THE FILE-DRAWER PROBLEM

September 6, 2012

Alexander L. Davis
Carnegie Mellon University
Department of Social and Decision Sciences
alexander.l.davis1@gmail.com

Contents

List of Tables

List of Figures

Abstract

This dissertation provides normative, descriptive, and prescriptive analyses of a scientist's decision to share data. The normative analysis (Chapter Two) concludes that, although there is no logical ground for determining whether data or theory is faulty when they conflict, data sharing policies that omit disconfirming data are unethical because they impose conventions on the reader, thus deceiving them. However, five experiments (Chapter Four) find that surprising disconfirmations are perceived to be caused by error, and future observations that are seen as diffuse are judged to be less worthy of publication. The second part of the normative analysis (Chapter Three) concludes that disconfirmations are more likely to be errors than affirmations only when the selection of true hypotheses is common. However, participants in the Wason rule discovery task (Chapter Five), who were asked to discover the rule that generated a set of three numbers (2,4,6), thought the opposite. With no penalty for incorrect error attributions, participants proposed triples that did not fit the rule (false hypotheses) more often than those that did fit the rule, but attributed error more often to disconfirmation than affirmation. Furthermore, they shared data based on their attributions of error, and these error attributions were affected by whether feedback was affirming or disconfirming, even after controlling for whether the data were actually error. The prescriptive analysis (Chapter Six) proposes methods of documenting data, methods, and statistical analyses so that penalties can be implemented when inferences are faulty or documentation is poor. The dissertation concludes with a recapitulation of the normative, descriptive, and prescriptive analyses and highlights directions for future work.

Part I

The Problem of Data-Sharing

Chapter 1

A Short History of Data Sharing

> The first principle is that you must not fool yourself—and you are the easiest person to fool. So you have to be very careful about that. After you've not fooled yourself, it's easy to not fool other scientists. You just have to be honest in a conventional way after that.
>
> —Richard Feynman, 1974 (1)

> In a desert prison, an older prisoner befriends a new arrival. The young prisoner talks constantly about escape, spinning plan after plan. After a few months, he makes a break. He's gone a week; then the guards drag him back. He's half dead, crazy with hunger and thirst. He wails how awful it was to the old prisoner: endless stretches of sand, no oasis, failure at every turn. The old prisoner listens for a while, then says, "Yep. I know. I tried those escape plans myself, 20 years ago." The young prisoner says, "You did? Why didn't you tell me?" The old prisoner shrugs: "So who publishes negative results?"
>
> —Janice Probst, 2006 (2)

> It can be proven that most claimed research findings are false.
>
> —John Ioannidis, 2005 (3)

In research as well as life, I make mistakes. I make a lot of them. I form hypotheses poorly, forget to measure age or gender, or create an instrument with very little construct validity. When an experiment doesn't come out the way I expect, these mistakes seem apparent. When the data come out the way I like, it's difficult to see flaw.

I don't want to bother others with my bad research, nor do they want to hear it. Pressures to feel both competent about myself and be evaluated positively by others push me toward hiding the data I see as flawed. Yet, when the results come out the way I want, I cannot convince myself that the 'bad research' that preceded the supposed discovery is irrelevant. This is the *file-drawer problem*, where each scientist must decide whether to share unwanted, disconfirming evidence with the scientific community.[1]

This shame, embarrassment, and disrepute associated with 'flawed results' has existed since the inception of institutional science. In the 18th century, measurement error and personal flaw were synonymous, as the "concealment of discrepancies in observation were not only common, they were considered a savant's prerogative. It was error that was seen as a moral failing" (10). Little has changed since then; in fact, the file-drawer problem seems to be getting worse. In 1990, 70% of results published in a sample of journals across a variety of scientific disciplines were statistically significant (11). This number increased to 86% in 2007, and is now over 90% in the social sciences. From what is published, in the last 300 years scientists have either become flawless researchers or adept at hiding their flaws.

Some have studied this process empirically. For example, Sterling (12) was one of the first to try to empirically verify the file-drawer problem in Psychology. He found that almost all (286/294 or 97%) reports that used significance testing from four prominent Psychology journals had 'statistically significant' results. While the published data may be accurate, the almost non-existence of published null results suggests not all data are shared.

Following this, Mahoney (13) constructed a fake paper on the efficacy of using reinforcement on children to modify their behavior, experimentally varying the results of the intervention. He sent the different versions of the paper to 75 reviewers of a journal (the *Journal of Applied Behavior Analysis*) that he knew would strongly favor efficacy of the treatment. When the results were what the reviewers (and their subfield) wanted to hear, they were more likely to recommend publication, and rated the paper as having higher methodological quality. When the data were not what they wanted to hear, they scrutinized it much more closely, and were three times as likely to find an unplanned typographical error in the manuscript.

Although omitting negative results from publication seems to be the *de facto* policy, there are those who have argued against it. For decades, Cohen (14; 15) and colleagues (16; 17) have lamented the low statistical power of psychological experiments, and argued that this implies that data are

[1] The dissertation is concerned with the inferences and behavior of the scientist who collects the data and must decide to submit it for publication, rather than an editor who decides to publish. Problems related to data sharing have been called a number of things. One is the file-drawer problem (4; 5). Another is selective reporting bias (6; 7). A third is the problem of disclosure (8). A fourth, is publication bias (6). A fifth is data availability (9). There are probably many more. The term selective reporting bias, where the researcher chooses to exclude data in a submitted article for publication, is most apt. The terms file-drawer problem and publication bias are ambiguously applied to both the journal and researcher decisions, although they may be narrowly defined to mean the former. Use of the term file-drawer problem should be interpreted as selective reporting bias in this dissertation.

suppressed from publication. Until recently psychologists have not been concerned. The watershed moment came when Daryl Bem published a paper in the prestigious *Journal of Personality and Social Psychology* providing an experimental demonstration of extra sensory perception (18). A flurry of criticism followed, focusing on statistical analysis, peer review, and publishing in Psychology (19). Following this, Simmons *et al.* (20) published a paper accusing psychologists of a culture of unethical data analysis and sharing practices.

Since then, these unethical practices have been exposed several times, including the cases of Marc Hauser, Diederick Stapel, and Dirk Smeesters. Failure to publish replications of Bem's original paper, most of which did not show the same effect, also spawned new 'file-drawer' websites where replications (especially failed ones) can be archived. Complementing this, several new projects have focused on independently replicating published psychology results (21; 22; 23). However, these file-drawer websites have so far had little success (24), and failed independent replications continue to "go unpublished, languishing in personal file drawers or circulating in conversations around the water cooler" (25).

The file-drawer problem is not limited to social science research. Medical research companies and medical schools have a strong financial incentive to make life-saving discoveries, while suppressing research that suggests their discoveries are false. Like Psychology, traditional publication bias approaches have found that most published medical research is confirmatory or statistically significant (26). For example, Hasenboehler *et al.* (27) found 74% of studies in orthopedic and general surgery reported positive findings, with another 9% reporting ambiguous or neutral findings.

What is published does not accurately reflect the research that is done. Some evidence supporting this comes from asking scientists what they do with their data. For example, Martinson, Anderson, and De Vries (28) surveyed 3,247 NIH funded scientists and found that overall, 0.3% admitted to falsifying data, 6% reported not presenting contradictory data, 10.8% withheld methodological details and results in published papers, and 15.3% reported dropping observations based on the 'gut' judgment that they were in error.

Other evidence comes from comparing published reports to other sources. For example, clinical trial registries indicate that one-third of trials still remain unpublished three years after conclusion (29). Many clinical studies submitted to IRBs produce negative results and are not published (30). AIDs trials with negative results take about twice as long to publish as those with statistically significant results (31). Anti-depressant trials submitted to the FDA do not match those published, inflating the effect size in the published literature by about one-third (32). Three nicotine treatment trials by Pharmacia went unpublished, but the successor and successful treatment was published in the *Journal of the American Medical Association* (33). The list goes on, including the anti-depressant paroxetine for children (34), lorcainide for myocardial infarction (35), reboxetine (36), Vioxx (37), and several anti-smoking therapies (naltrexone, mecalymine, and Habitral).

Even Physics, the paragon of science, has a history file-drawer problems. For example, 40% of results in one issue of *Review of Particle Physics* were omitted because of "strong sources of bias", "assumptions that the Particle Data Group does not wish to incorporate" or "inconsistency with other reported results" (38). In attempting to measure the charge of the electron, Millikan collected data from 140 oil drops but reported only 58, using his own judgment to determine which data were valid and which were invalid (39). Just like psychologists and biomedical researchers, physicists are "always doing experiments or making observations that disappoint them. They look for some phenomenon or relationship and they do not find it. Most of these negative experiments are forgotten and the results consigned to the file drawer" (40).

1.1 Proposed Causes

The two strongest explanations for the file-drawer problem are perverse incentives and error attributions. Perverse incentives are institutional rewards for reporting only successful findings to others. Error attributions are cognitive tendencies to see disconfirming evidence as flawed and affirming evidence as flawless.

1.1.1 Perverse Incentives

Simple, "eye-catching", and easy-to-comprehend stories yield publications (41). These publications, in turn, reward researchers with jobs and funding for further research (13). For example, the American Medical Association reports that greater than 70% of the funding for pharmaceutical research comes from industry (42). This funding is often dependent on the ability of the researcher to prove they can get positive results (41).

On the other hand, publishing results that contradict a flashy hypothesis can mean sacrificing one's career, or even intimidation from those who would rather not see the results published (43). Any result that suggests a therapy is ineffective puts that company's potential profits at odds with the public's well-being (44). These companies generally distort reports of adverse events, make them difficult to understand, or do not even measure or report them at all (45). The pressure to produce positive results comes not only from the medical research companies, but also from researchers who want to save lives, and patients who want to live. Unfortunately, it is difficult to produce uniformly positive experimental evidence for any theory, even if it is right.

When mixed or disconfirming results occur, it is “tempting for investigators to submit selected data sets for publication, or even to massage data to fit the underlying hypothesis” (46).

Physicists who study gravitational waves are acutely aware of the challenges of institutional incentives to produce flashy results. Researchers at the Large Interferometer Gravitational Wave Observatory (LIGO) try to detect the presence of gravitational waves, but these waves are so small that they still haven’t been detected in over 40 years of searching and with multi-billion dollar research budgets. Because these physicists never get to produce a discovery, they have difficulty convincing others of the value of their work. For example, one gravitational wave physicist consistently had his students criticized for not having made a discovery, making it difficult for them to graduate, get jobs, or tenure:

> And the reason for why it became big, at [my institution], and in my head, was fundamentally because of an incident that happened with two students who had done a beautiful job and they got this shit from my own colleagues. And I said I’m never gonna put students through this again. If we are going to continue with this, we are going to have to do it on a scale such that even if we don’t see anything, no goddam [expletives deleted] theorist, OK, can confront one of my students and say “What did you discover?” and give him a sneering, [expletive deleted], ride, OK? So it has to be something where the upper limit is good enough. And you say “Yeah we have made a scientific statement.” (pg. 668) (40)

1.1.2 Error Attributions

In any groundbreaking experiment, the difficulties of measurement are typically so extreme that any failed prediction could be attributed to a number of flaws, including bad design, an underpowered study, incorrect analyses, or chance (47). In these circumstances disconfirming results are both likely to occur and reasonably attributed to error. Thus, any data sharing policy that omits results that are attributed to error will lead to the file-drawer problem.

Take the famous Michelson-Morley experiment as an example. This experiment sought to measure the velocity of the earth through the aether, an invisible substance that all matter was hypothesized to be suspended within. Measurements of the aether worked much like measurements of the velocity of a car by putting one’s hand outside the window to measure the force of the wind. Just as one’s hand experiences resistance from the the air outside the car window, the earth was expected to experience resistance from the aether. However, the theoretical effect of this ‘aether wind’ was expected to be so small that measurement instruments needed to be extremely sensitive; so sensitive that “a mass of 30 grams placed on the end of one of the arms of an apparatus weighing tons was enough to upset the results dramatically” (pg. 34) (48). Sensitivity to vibration was only one among many possible measurement errors, including changes of temperature “as small as 1/100 of a degree” that would theoretically triple the effect of the aether wind. The measurement apparatus also could not be built out of metal, to reduce problems of vibration and increase the weight, because of magnetic fields. Nor could the device be made out of wood because of sensitivity to humidity.

In this ocean of possible measurement flaws, the eventual failure to detect the aether wind was disappointing but unsurprising. All of the experiments conducted by Michelson and Morley, and subsequently by Morley and Miller, were null results, regarded by their creators as failures, and were not followed up by Michelson, as he was “so disappointed at the result that instead of continuing he immediately set about working on a different problem: the use of the wavelength of light as an absolute measure of length” (48). In their time, the failure to measure the aether wind was an anomaly, and was explained as being caused by some combination of the many possible experimental flaws previously mentioned. However, as it was retrospectively consistent with Einstein’s General Theory of Relativity, rather than being flawed, the Michelson-Morley experiments are considered the greatest physics experiments ever conducted.

Another critical test of Einstein’s relativity theory had similar problems of measurement error. The solar eclipse of 1919 allowed Einstein’s theory to be compared against Newton’s theory of gravitation. In Newton’s theory, gravity should bend light during the eclipse. However, Einstein’s theory proposed additional bending due to the curvature of spacetime.

Sir Arthur Eddington led the main expedition to measure these light deflections during the eclipse (at Principe near Africa), while other research groups took measurements simultaneously (at Sobral in Brazil). The measurements of these different groups did not agree. As with the Michelson-Morley experiments, the measurement of the light deflections were extremely difficult, where “the difference in focal length between a hot and a cold telescope will disturb the apparent position of the stars to a degree which is comparable with the effect that is to be measured” (pg. 46-47) (48). Most of the measurements that were inconsistent with Einstein’s theory were considered ‘noisy’ (e.g., the Sobral astrographic plates), and removed from the data analysis. Interestingly, like Millikan in his oil drop experiments, Eddington attributed the results from the Sobral astrographic plates to systematic error, often without being able to explain why (pg. 51) (48). Like Millikan, Eddington was right.

Attributions of ‘noisy’ results to measurement error are not always right, however. Chemist and Nobel Laureate Irving Langmuir personally demonstrated this. Two chemists, excited about an apparent discovery, asked him to evaluate an effect (the Davis-Barnes effect) that relied on an observer counting flashes through a tube. After they demonstrated the effect to him, Langmuir concluded that they hadn't made a discovery, but were instead biased by their expectations. To prove this, Langmuir secretly changed the pattern of voltages used in the experiment, without notifying the experimenter (Barnes). Barnes, himself an esteemed professor at Columbia University, counted flashes in a pattern that was unrelated to the voltage changes, the supposed cause. When Langmuir confronted Barnes about this clear refutation, Barnes immediately generated explanations ad-hoc, that "the tube was gassy" and the "temperature has changed." Langmuir called this response, *pathological science*, where Barnes "immediately—without giving any thought to it had an excuse. He had a reason for not paying any attention to any wrong results. It just was built into him. He just had worked that way all along and always would. There is no question that he is honest; he believed these things, absolutely"(49). Langmuir proceeded to write a detailed letter to Barnes, arguing that he was "counting hallucinations." Like Feynman, Langmuir believed that "men, perfectly honest, enthusiastic over their work, can so completely fool themselves."

1.2 Possible Effects

The file-drawer problem is likely to have a slow but severe effect on a scientific field, eventually causing it to become completely hobbled or extinct, as was probably the case with Soviet Lysenkoism. This happens for a number of reasons. Flawed methods cannot be used to build better ones, instead making every researcher start from scratch (50). Flashy but wrong theories will be perpetually proposed and the refutations will be perpetually buried, leaving the field in a conceptual stasis (51). Negative results will continue to accumulate if not reported, as researchers "who are unaware of the contradictory experimental results repeatedly attempt to confirm or disprove the selected results in the literature" (52). Funding cuts will occur when discoveries are repeatedly overturned and replicable results do not surface, as happened with Title VII grants for physician education (2). Honest students will quit, as observed by Wagenmakers, "I've seen students spending their entire PhD period trying to replicate a phenomenon, failing, and quitting academia because they had nothing to show for their time" (25). Those who can get positive results, by ethical or unethical methods, will remain.

The file-drawer problem is particularly harmful in medical research. Treatments that seem initially useful must be abandoned, wasting time, money, and putting patients at risk, as in the case of zidovudine (31). Effect sizes are likely to shrink drastically upon replication or uncovering of the file-drawer, with the potential to make subsequent research based off of initially flawed results completely unusable (53). The file-drawer problem corrupts meta-analyses needed for evidence-based medical decision-making (27). This threatens patient safety, and potentially imposes high opportunity costs, as funding is diverted away from the search for real cures to ineffective treatments. Khan, Khan, and Brown (54) argue that unreported results can make ineffective drugs look effective, and if these are then used as active control groups (i.e., non-inferiority studies) in future studies, rather than using placebo controls, future drugs that are roughly equivalent to the active control (which actually has no effect) will seem to be effective. They strongly argue against eliminating the use of placebo controls because of the file-drawer problem.

The file-drawer problem is also likely one of the main causes of the poor replicability of published drug trial results.[2] Successful replications of research in haematology and oncology have been rare, as only 11% (6/53) of one systematic replication attempt were successful (46). Replications of antidepressant trials fared slightly better, with 48% (45/93) being significantly better than placebo (54). Part of this failure might be the fact that "there are no guidelines that require all data sets to be reported in a paper; often, original data are removed during the peer review and publication process" (46).

1.3 Overview of the Dissertation

Researchers and consumers of research are concerned by their experience with the publication system. This dissertation complements empirical research on the file-drawer problem that address its prevalence and causes (55; 11; 56; 12), extends previous philosophical (57; 58; 59) and mathematical analyses (3; 60), and builds on prescriptions based both on method and documentation (61; 62; 63; 64).

With this in mind, the dissertation explores the following two questions:

- What data sharing policies emerge in the face of unexpected and unwanted data?
- Is reasoning about the validity of data distorted by incentives?

To do this, I use the Normative-Descriptive-Prescriptive (NDP) framework of Behavioral Decision Research (65; 66; 67):

[2] This does not apply to drug approval, as FDA requires pre-approval of all trials that will eventually be submitted as evidence.

- *Normative Analysis*: In Part Two of the dissertation, the normative analysis examines how idealized, rational agents should reason and act when sharing data.
    - Chapter Two argues that data sharing is an ethical rather than epistemological problem.
    - Chapter Three provides a normative analysis of data sharing policies using probability calculus.
- *Descriptive Analysis*: In Part Three of the dissertation, the descriptive analysis builds on the normative analysis, trying to understand how humans actually reason and act compared to the ideal. The descriptive analysis uses two sets of experiments.
    - Chapter Four presents experiments where participants are asked to attribute the cause of unexpected results in a hypothetical psychology experiment.
    - Chapter Five presents experiments where participants discover a rule when there is the possibility of error and there are incentives to share or hide data from others.
- *Prescriptive Analysis*: In Part Four of the dissertation, the prescriptive analysis asks how real people can be brought closer to the ideal. The prescriptive analysis proposes methods of documenting and communicating uncertainty from scientific experiments.
    - Chapter Six presents a simple and open approach to documenting and sharing data.

Part II

Normative

Introduction to the Normative Analysis

Part Two of this dissertation discusses when selective reporting, where data are omitted from publication if they conflict with institutional incentives or are attributed to error, is normatively justified. Chapter Two evaluates data sharing from the viewpoint of philosophy of science and philosophy of statistics, and proposes an ethical standard for data sharing. Chapter Three evaluates two justifications for not sharing disconfirming data: 1) that disconfirmations are not as informative as affirmations, and 2) that disconfirmations are more likely to be error.

Chapter 2

Cleaning the Data or Cooking the Books

2.1 Introduction

Some scientists are distressed by the lack of rules for data sharing, worsened by the *de facto* action prescribed by their paradigms, where inconvenient data are routinely discarded (68). This distress is warranted. Fanelli (11) found that over the last 20 years negative results have begun to disappear from scientific journals, and that this is worse for the social sciences. Ioannidis (3) claims that most published research findings in medicine are provably false. Wasserman (69) even proposed completely eliminating peer review, opting for a more democratic collaborative review system based on pre-prints (http://arxiv.org/).

This alarming situation is not helped by poorly articulated rules about data sharing provided by scientific bodies like the National Science Foundation or the National Academy of Sciences. Rather than addressing the real problems scientists face when deciding whether or not to share messy data, these bodies limit their discussions of misconduct to "actions that are unambiguous, easily documented, and deserving of stern sanctions" (70); that is, Fabrication, Falsification, and Plagiarism.

Instead, researchers must decide whether removing data from a published report would constitute "cleaning the data or cooking the books" (70). These decisions are common, ambiguous, and have important consequences. Poor data sharing policies undermine the effective communication of scientific research. Suppressing unwanted data, for example that contradict a favored theory, waste the resources of those who try to build from it. However, sharing too much data can confuse our peers with an incomprehensible and unusable morass of 'mere facts' (71), while sharing faulty data can lead others to draw incorrect conclusions.

In this chapter I argue that sharing data is an ethical decision, where one chooses to minimize the chance of deceiving one's reader. Any policy for determining when data are faulty and should be excluded from communication is based on *convention*, and scientists may reasonably disagree on the conventions they find acceptable. Methods of cleaning data that irrevocably impose conventions on the reader make the 'cleaned' results deceptive and thus unethical.

2.2 Unwanted data

When shared with others, data that affirm a scientist's theory are likely to yield rewards, whereas data that suggest this theory is false will not. Although ideally the scientific community benefits the most from data that make the evaluation of theories clear, affirming or not, in practice refutations are not perceived as providing this clarity. As Lakatos argued, "there is no falsification before the emergence of a better theory" (59), which, if taken seriously, means that refutations are valueless unless accompanied by affirmation. Thus, scientists set perverse incentives for themselves and each other, where only convincing data are rewarded. In turn, they are provided only that, in the form of affirming results. Because refutations are valueless (or even of negative value), the interests of the scientific community and individual researcher are often at odds.

Take for example a researcher that proposes a specific hypothesis about behavior, adhering to the assumptions of the members of her community (e.g., her lab, her advisor). This may be that children perform Bayesian causal induction, that unconscious priming can be linked to sensory perception, or that prejudice can arise from arbitrary group distinctions. If the experimenter's specific hypothesis, derived from the paradigm of the community, is not supported by the data, it also casts doubt her abilities. Her lab members may suggest her experiment was poorly conducted or her calculations were incorrect, giving the researcher a bad reputation. As Kuhn and others have argued, once a paradigm is established and expectations are entrenched, failing to get an expected result is a "failure of the scientist", (pg. 35) (71), and "discredits only the scientist and not the theory" (pg. 80). Thus, a report of any result other than one that affirms the paradigm is merely a statement about the poor quality of the researcher. In contrast, if her hypothesis is supported, she is encouraged to publish the paper, and rewarded with future employment and funding.

In such situations, data sharing is a *signaling game*. In the signaling game, the researcher has private information (the data she's collected) about her hypothesis, and the community wants to reward a researcher if her hypothesis is interesting and the data support it, and not otherwise (72). The obvious solution is for the community to look at the data the researcher shares (the signal), hoping it will clarify the value of the researcher's hypothesis. However, if the cost to the researcher of omitting falsifying data is low enough, then she can always create a flawless picture of her theory with data, guaranteeing reward. In this case, the signal (data) is meaningless, and the community would be reasonable to ignore any data provided by the researcher. Since both the researcher and community do not know whether the researcher's hypothesis is correct (although the researcher has suggestive evidence), the community is forced to evaluate her hypothesis purely on a priori grounds, such as how interesting or surprising the theory is.

Even altruistic researchers, who are willing to be surprised and make discoveries, are affected by these perverse incentives. This is because they may only be able to get a hearing for their work if they achieve and maintain status as a respected scientist. In the archetypal communities described by Kuhn, that would mean avoiding results and discoveries that challenged the paradigm. In signaling game terms, if the scientific community does not value anomalous data, it would be in the interest of scientists who genuinely care about discovery to not share them. In sum, scientists are incentivized to not share disconfirming data. As long as these perverse incentives exist, ethical researchers are likely to suffer.

2.3 Incomprehensible data

If sharing data were purely about incentives, then changing incentives could possibly provide an easy solution to problems of publication bias. However, the incentive dilemma is hidden in an epistemic one. Data that do not affirm a theory are also more difficult to make sense of than those that demonstrate the expected.

Difficulty determining the meaning of data are severe and persistent problems that are often at the heart of scientific and philosophical debates. Philosophers of science and statistics, from Laplace (73) to Gelman and Shalizi (74), have grappled unsuccessfully with making sense of anomalous data. To see the ethical nature of data sharing, the epistemic veil surrounding data sharing decisions must be lifted.

2.3.1 Kuhn's Paradigms

The interpretability of data, or lack thereof, is determined by paradigm. The paradigm prescribes what to look at, how to take measurements, and resolves many other methodological and theoretical choices (sometimes called the "frame problem" (75)). Experiments that result in a known outcome within a paradigm are the only ones that scientists can unambiguously make sense of, as the paradigm serves to prepare the mind of the researcher to spot the type of phenomena the paradigm prescribes, while blinding the researcher to other types.

When there is no paradigm, useful data are hard to spot. Different scientists will observe “the same range of phenomena [and] describe and interpret them in different ways” (pg. 17) (71). Fact-gathering is a relatively random and undirected process.

When there is a paradigm, the meaning of data are still problematic. Observations that don’t fit into the paradigm are not necessarily reshaped to fit what was expected. Instead they may not even be seen, as data are partial records that are “immensely circumstantial” (pg. 16) (71). As a result the details needed to recognize anomaly may not even be recorded. That is, a paradigm can preclude even seeing potentially anomalous data.

The data may not make sense, even if there is a paradigm and the data are seen. Without an ability to generate causal explanations, data are “unrelated and unrelatable” mere facts, or even worse “not quite a scientific fact at all” (71). Instead, paradigmatic explanations that have previously succeeded will be invoked to explain anomalous data. For example, Joseph Priestley did not discover Oxygen, but instead another instance of his theory: “air with less than its usual quantity of phlogiston” (pg 53) (71).

Once data are recognized as anomalous they must be reconciled with theory. This can happen either because the theory is wrong, the data are flawed, or both. Any conflict between theory and data is really a conflict between an explanatory theory of the causal mechanisms of interest and an interpretive theory of the meaning of the data (59). At the most basic level, data are composed of observations that assume one’s visual and cognitive perceptions are accurate. At higher levels, entire theories of observation and instrumentation are used. For example, Galileo’s observations of “mountains on the moon and spots on the sun” were aided by a telescope, and thus the “optical theory of the telescope” (pg. 15) (59). As a result, when observation and theory do not align, there is not a conflict between fact and theory, but between two theories.

Data sharing policies are ambiguous because there is no basis for privileging statements based on an observational theory, such as the optical theory of the telescope, over other “non-observational” theories, such as Newton’s laws. No appeal to psychology could solve this problem, because the psychology of observation is itself dependent on an observational theory. All observations are theory-laden, and empiricism must assume a “psychology of observation” that is hoped to be accurate (59). There is no logical way to determine whether theory or data were wrong, meaning there is no purely logical guide to data sharing.

Instead, when theory and observation conflict, resolution must occur on extra-logical grounds; by ineffable tacit knowledge (76), severe tests of alternative explanations for the data (77), what has traditionally been done (71), the sagacity (78) or intuition (57) of the scientist, or what provides the most convenient and easy-to-work-with explanation (58). A decision must be made about whether to retain theory or observation (57).

2.3.2 Poincare's Conventionalism

The dominant approach to dealing with the conflict between observational and non-observational theories is *conventionalism*. In contrast to justificationism, which permits only logically valid arguments, conventionalism allows some knowledge to progress by decision, without having to provide reason. Conventionalism chooses the meaning of data based on convenience; what is easiest to work with and understand.

Take for example Henri Poincare's (58) discussion of choosing between Euclidean or non-Euclidean geometries. According to Poincare, if one discovered "negative parallaxes" or proved that "all parallaxes are higher than a certain limit", this would not disprove Euclidian geometry, because we could also "modify the laws of optics, and suppose that light is not rigorously propagated in a straight line." This conclusion is reasonable, and possibly true, but not provable without also using geometric assumptions. Instead, the decision to throw out the theory or the data depends on which is "more advantageous", or convenient, and in this way "Euclidean geometry has nothing to fear from fresh experiments" (pg. 73).

Poincare points out that convention is not logically justifiable. Decisions are not "synthetic a priori intuitions," and are also not "experimental facts". However, conventions are not arbitrary since they have an "experimental origin" (pg. 110). Poincare only allows decisions to be made about the fundamental theories, where "experiment may serve as a basis for the principles of mechanics, and yet will never invalidate them" (pg. 106).

To Poincare, experiments are used as conventions to form the basis of theories, but not to refute or invalidate them. Conventionalism allows some data to be meaningless or useless to the person collecting the data, but quite useful to someone who holds different conventions. Failing to share data means imposing one's conventions on the reader, conventions that, if possible, require explicit articulation by the author.

2.3.3 Popper's Falsificationism

Karl Popper's viewpoint was that only refutations, or falsifications, were valuable data, as only they could conclusively disprove theories. This focus on falsification rather than affirmation was a radical change in philosophy of science (57), as he changed the status of affirmative data from the most valuable (as held by the positivists) to useless.

To understand Popper's approach, consider the following (universal) statement: if the weight placed on a string is greater than its tensile strength, the string will break. Now, consider a singular statement: we placed an object on a string that has greater weight than the tensile strength of the string. Putting these two statements together yields the prediction that the string will break. In this example, if the string does not break, then the universal statement is disproved. This type of simple syllogistic reasoning is the foundation of Popper's approach. It is deductive because the universal statement cannot be verified (inductively) by finding every string and every object that could be put on the string and measuring whether the string breaks when the object is placed on it. This is because the universal statement applies to all strings and all objects; we cannot search all space and time for each object and string to verify the universal statement.

In contrast, singular statements (e.g., the string broke) refer to a specific space and time. It is simple enough to verify a singular statement: we only need one instance of a string and one instance of an object at any space and at any time. Popper calls the asymmetry between being able to verify singular but not universal statements, *unilateral decidability*. It is unilateral decidability, where one can verify a singular but not a universal statement, that allows a theory to be falsified, but not verified, by observation.

To make his approach empirical, Popper talks of observations as occurrences. For example, "I observe a glass containing water at noon on Saturday in Pittsburgh" is an occurrence because it is a spatio-temporally limited singular statement. An event is a class of occurrences that have the same form but differ only in individual names (e.g., observing a glass of water at any time). Popper thus calls potentially falsifying observations *basic statements*, which are singular statements "asserting that an observable event is occurring in a certain individual region of space and time" (pg. 85).

Popper refers to "inter-subjective" agreement to determine whether basic statements are true; that is, multiple people see an occurrence and agree that it has occurred. If there is disagreement, the basic statements can be tested against other basic statements, ad infinitum, in the hope of reaching inter-subjective agreement. Eventually, testing will resolve to observations where everyone agrees, that are "easy to test." Once an easy-to-test statement is verified, the entire chain of tests that led to it can conclusively resolve in falsification. If one can't come to this point, then Popper believes either the phenomenon was not inter-subjectively testable, not observable, or that language (communication) has trapped us. Popper calls inter-subjectively agreeable basic statements that are repeatedly demonstrated and are inconsistent with a theory *reproducible effects*. Only these reproducible effects can falsify a hypothesis.

In sum, Popper's Falsificationism proposes that a theory, once sufficiently axiomatized and checked for consistency (i.e., it is not self-contradictory), to be falsifiable and thus scientific, it needs to rule out not just occurrences, but at least one event. If this event occurs the theory (universal statement) is falsified. The only important data are falsifying events, and theories need only be falsifiable to be scientific.

Although it appears to provide an unambiguous policy for data sharing (i.e., share only falsifying reproducible effects), his approach eventually failed. This is because at any point there is no logical way of disproving what Popper called "conventionalist strategems" that could be used to save a theory against falsification. The four "main" stratagems were: 1) introducing ad hoc auxiliary hypotheses that make the theory consistent with falsifying data, 2) changing the ostensive definitions of the data or the theory to make a falsification into verification, 3) challenging the data directly, as being "insufficiently supported, unscientific, or not objective, or even on the ground that the experimenter was a liar" (pg. 60-61), or 4) a claim that the theory was misapplied or misinterpreted by a fallible theoretician. Popper warned social scientists about these stratagems (pg. 62). Indeed, they appear to be common reasons for not sharing data as each conventionalist stratagem can be invoked to argue that the data should not be shared with the scientific community.

Popper's way of avoiding the infinite regress of trying to prove an observation with other observations was to accept a limited form of conventionalism. According to Popper, one could decide to accept a basic statement as true without further justification. These are called *accepted* basic statements. All other statements must be proved and subjected to additional "inter-subjective tests" or provisional agreement. Popper's conventionalism differs from Poincare's in that Popper allows these decisions to only apply to observables (basic statements) rather than theories (universal statements).

2.3.4 Lakatos' Research Programmes

Popper's student, Imre Lakatos, recognized the shortcomings of both Poincare and Popper's perspectives. For Lakatos, Poincare is too conservative: a theory can always be preserved as long as it is convenient to do so. Popper's is, in contrast, too risky: A falsifying statement could, if accepted based on a fallible convention, rule out a true theory. Instead, for Lakatos the only tenable position is that "all propositions in science are fallible" (pg. 19) (59); that is "scientific theories are not only equally unprovable, and equally improbable, but they are also equally undisprovable."

Lakatos, like Popper, saw conventionalist decisions as a necessary part of a philosophy of science. He solves Popper's problem by requiring that any conventionalist decision take into account whether the decision leads to the prediction of new facts. These new facts must be “improbable or impossible” without using the newly modified theory. If the modification does lead to the prediction of such new facts, Lakatos considers the decision *theoretically progressive* and thus admissible. This is Lakatos’ approach: a new theory must explain at least as much as the old one and also predict new facts at a rate that outpaces the adjustment of the theory to fit anomalies. For Lakatos, to allow refutation without creatively (progressively) proposing an ad-hoc defense “shows nothing but the poverty of our imagination” (pg. 35) (59). For Lakatos, stunning predictions are the valuable data that should be shared.

2.3.5 Summary

Kuhn, Poincare, Popper, and Lakatos provide many arguments why data may be invalidated. Kuhn tells us that we may not see them, they may not make sense outside our paradigm, or we may be convicted as incompetent if we do share them. Poincare licenses us to discard any data if it is convenient to do so, so as to maintain a tractable theory. Popper tells us that only falsifiers matter, based on reproducible effects. Lakatos only cares about the prediction of new facts as important data. All of these approaches appeal to conventions, rather than logical justifications, to reject data.

2.4 Statistical Decisions

Lakatos and Popper both wanted a formal mathematical approach that could assist scientists in resolving conflicts between data and theory. In parallel, but mostly independent of these philosophical discussions, statisticians provided such rules. The *Frequentist* approach makes decisions about data that are not statements about the truth value of a hypothesis directly, but instead are justifications for provisional rejection of a statistical hypothesis with known error rates. The *Bayesian* approach, on the other hand, allows truth values to be assigned to hypotheses directly in terms of subjective probabilities.

2.4.1 Frequentist

Fisher’s null hypothesis significance testing

In the early 20th century, around the time when Popper elaborated Falsificationism, R.A. Fisher developed an approach for statistical hypothesis testing to resolve conventionalist decisions about conflicting theory and data. [1]

[1] While founding modern statistics (79; 80) by inventing and refining concepts of sufficiency, consistency, efficiency, maximum likelihood estimation, deriving sampling distributions, and

Fisher's approach, *null hypothesis significance testing*, was to only accept statements that indicate a well-defined sample of data are inconsistent with a *null hypothesis* (e.g., that two sample proportions are equal). In this approach, the theoretical frequency distribution of the null hypothesis is constructed and the probability of observing the sample data given this distribution is calculated. If this probability is sufficiently small (usually less than 5%), the data are judged to be inconsistent with the null hypothesis, or *statistically significant*, and the null hypothesis is rejected.

Fisher believed that null hypothesis significance testing provided exactly what scientists needed to resolve the conflict between theory and data, "simple rejection of a hypothesis, at an assigned level of significance" (pg. 40) (83). To him, this is "all that is needed, and all that is proper, for the consideration of a hypothesis in relation to the body of experimental data available" (pg. 40). Like the conventionalist approaches of Popper and Lakatos, this rejection is tentative, where "no irreversible decision has been taken" (pg. 38) (83).

Data that fail to reject the null hypothesis demonstrate a lack of experimental understanding, as experimental knowledge is gained from knowing "how to conduct an experiment which will rarely fail to give us a statistically significant result" (pg. 14) (84). Aside from rejecting the null hypothesis, there is no other purpose of an experiment which "may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (pg. 16) (84).[2] Based on this reasoning, Fisher's prescription is to only to share results that are statistically significant, and to ignore the rest (pg. 1244 as cited in (81)):

promoting randomization in experimental design, he also was engaged in bitter conflicts with other scientists, including physicists Arthur Eddington and Harold Jeffreys, and statisticians Karl Pearson, and especially Jerzy Neyman (81). His personal and derisive attacks on the Neyman-Pearson Frequentists, Subjective Bayesians, and Objective Bayesians, still casts a shadow over debates between these factions on hypothesis testing and statistical induction. In light of this, it is not surprising to find that Fisher's views, often intentionally extreme to avoid concessions to others (80; 82), can be easily misinterpreted as justifications for not sharing data.

[2] As to uncontrolled causes, or auxiliary hypotheses, Fisher argued for randomization rather than coming up with "an exhaustive list of such possible differences appropriate to any one kind of experiment, because the uncontrolled causes which may influence the result are always strictly innumerable" (pg. 18) (84). Fisher also attempted to derive an objective method of inverse probability, where statistical hypotheses could be given probability distributions. He called this fiducial probability. Suppose one has a pivotal quantity (a function of a statistic and parameter whose distribution does not depend on the parameter). This pivotal quantity can be inverted and a distribution can be solved for the parameter, giving it a fiducial, rather than posterior, distribution. From this fiducial distribution, the probability of the parameter lying in any region of the distribution can be calculated. Fisher's fiducial argument to confidence interval estimation is now used, and is equivalent to Neyman's unconditional confidence interval (85; 82) although their interpretations were different.

> "it is usual and convenient for experimenters to take 5 percent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard...If P is between 0.1 and 0.9, there is certainly no reason to suspect the hypothesis tested."

This pervasive idea may be the single strongest convention that determines data sharing since Fisher popularized his approach. Not sharing data that are not statistically significant may be *the* data sharing problem, one that Fisher invented and advocated.

Neyman-Pearson Powerful Tests

Jerzy Neyman and Egon Pearson (son of Karl Pearson) modified the Frequentist foundations Fisher set down (86). They agreed with Fisher that it is possible to use mathematics to guide decision about statistical hypotheses, but instead appealed to a decision-theoretic account of hypothesis testing. Their approach was concerned with creating a pre-planned analysis that could minimize errors from two types of decisions: 1) concluding that the null hypothesis is false when it is true (Type 1 Error), and 2) concluding that an alternative hypothesis is false when it is true (Type 2 Error). To do this, the experimenter judges which error is more important, and chooses an error level that this test must not exceed (Type 1 Error or α). Once this is determined, a test statistic is created so as to minimize Type 2 Error (β). This test is called the *most powerful test of level alpha.*

To calculate a most powerful test, at least one alternative hypothesis ($\neg H$) must be proposed, otherwise the "problem of an optimal test of H is indeterminate" (pg. 104) (87). In contrast, Fisher was content to only specify a null hypothesis H and the complement $\neg H$ without any specific alternatives. By calculating the Type 1 and Type 2 error levels, one can get an idea of how often false-positive and false-negative decisions will be made. The calculation of alpha and beta levels also provide guides for fixing one's experimental design by "(i) alter[ing] the design of the experiment, (ii) try[ing] to find a more powerful test, (iii) increas[ing] the level of significance and (iv) increas[ing] the number of observations." (pg. 107) (87).

Thus, Neyman and Pearson developed a method for maximizing the chance that a result will be statistically significant given a specified set of alternative hypotheses. In this sense, they constrain the data sharing problem to sharing statistically significant results among studies that have high power. If this rule were followed, more results would be shared if studies could be planned with high power, or less data would be shared if higher power could not be achieved.

Mayo's Error Statistics and Severe Tests

Deborah Mayo (77) proposed a philosophy of statistics that generalizes the decision-theoretic approach of Neyman and Pearson. Her proposal is that if a *severe test* is conducted, which has a low risk of making Type 1 and Type 2 errors, and a theory is not refuted by this severe test, then there is good evidence that the theory is true. That is, Mayo's severe test is one where a hypothesis would have a high probability of being rejected if it were false in the light of inconsistent data and a low probability of being rejected if it were true. A hypothesis severely tested in this way is not the same as Popper's "all theories yet to be refuted" because the severe tests are custom designed to disprove the hypothesis.

In this approach, what separates science from pseudoscience is the ability to learn from error or failed predictions. If a failed prediction gives "rise to a fairly well defined problem; specifically, the problem of how to explain it" (pg. 33) and can be "pinned to a specific hypothesis" (pg. 34), then the approach is scientific. Possibly failed auxiliary hypotheses, or potential sources of error, are separated and given thorough methodological and statistical examination with the goal of producing reliable experimental knowledge that remains even after the theories that inspired the experiments are proven false.

Mayo explicitly rejects any communication of data that "prevents the determination of valid error probabilities" (pg 297), for example by "treating pre-designated and post-designated tests alike" (pg. 296). Thus, for Mayo, any procedure that invalidates error probabilities, which failing to share data arguably does, is inadmissible.

Summary

The Frequentists argue that if data are very unlikely given a hypothesis (i.e., have a low p-value) then this is evidence, but not proof, that the hypothesis is false. While the decision about the truth of the hypothesis itself depends on broader (non-statistical) scientific judgment, significance tests can provide the best formal guide. Statistically significant data are seen as meaningful and worth sharing with the scientific community, whereas non-significant data are not. From Fisher's point of view, failing to get statistically significant results means one does not understand the experiment that was conducted, and one should try again rather than inform others about this failure. Fisher's mathematical viewpoint provides easy justifications, documented and analyzed by Greenwald (88)[3], for sweeping confusing and unexpected results

[3] He argues that psychologists see data as synonymous with the hypothesis the researcher wants to test. Data that fail to reject the null hypothesis are seen as equivalent to data that support the null hypothesis, which the researcher literally believes to be false. Thus, non-significant data are useless. He also summarizes three other arguments made by psychologists

under the rug.

2.4.2 Bayesian

The alternative approach to making statistical decisions is the Bayesian one (also called inverse probability). It is more flexible, and more audacious, allowing probabilities to be assigned to explanatory hypotheses directly. In this paradigm, probabilities are not long-run relative frequencies of events, but instead, "relative, in part to ignorance, in part to our knowledge" (pg. 6) (73). That is, probabilities are epistemic states of a person.

For example, suppose a person flips a fair coin which has known Frequentist properties (e.g., in an infinite sequence of identical flips of this coin the relative frequency of heads to tails is 0.5). However, the coin flipper views the outcome of the flip but does not tell us. Clearly the probability he assigns to heads or tails is either 1 or 0, but our probability has not changed. His probability is not the same as ours. Thus, "probability is a function not only of the coin, but also of the information to the person whose probability it is. Thus subjectivity occurs, even in the single flip of a fair coin, because each person can have different information and beliefs" (pg. 5) (89).

There are two main schools of Bayesian statistics: *Subjectivist* and *Objectivist*. Both schools agree on two fundamental elements of probabilistic beliefs: 1) that they must be coherent, and 2) that they must operate according to Bayes' Rule. Coherence assures that one cannot be given a series of bets that guarantee a loss of money. Bayes' Rule guarantees that coherence is maintained in the light of new data (conditioning). What the two schools of Bayesian statistics disagree on is where prior probabilities (or prior total evidence; (90)) should come from; that is, from 'objective' or 'subjective' sources.

Subjectivist

Subjectivist Bayesians argue that hypotheses have truth value based on subjective belief. For them, scientific judgment depends on beliefs that are "yours alone, and need not be the same as what someone else would say, even someone with the same information as you have, and facing the same

against reporting non-significant results: 1) that finding non-significant results is not a discovery and does not advance science, therefore it is not worth reporting; 2) that statistical significance is evidence of correct experimental design and hypothesis; and 3) that "there are too many ways (including incompetence of the researcher), other than the null hypothesis being true, for obtaining a null result." He provides refutations for each of these arguments, the strongest being refutation of the third: while it is true that the incompetence of a novice is likely to lead to noisy results and failure to reject the null hypothesis, incompetence of the novice and expert both can lead to systematic bias that can support one's hypothesis. That is, errors due to incompetence cut both ways.

decisions" (pg. 1) (89). When deciding whether to believe an experimental result, the Subjectivist Bayesian implicitly or explicitly assigns a subjective probability to hypotheses and data directly. This subjective probability can be elicited, with varying degrees of success, using proper scoring rules (e.g., (91)). By assigning prior probabilities and likelihood functions to each hypothesis under consideration, the Subjectivist can deductively derive the posterior probabilities of each hypothesis given new data. This posterior probability is directly applicable to any hypothesis; that is, the Subjectivist Bayesian can hold a probability of an explanatory theory, which Frequentists and Falsificationists will not do. Because of this, the Subjectivist Bayesian approach has great advantages in terms of representing sparse or unobservable data, while additionally taking into account beliefs and other factors that play into a decision that a Frequentist would not admit.

This approach also appears[4] to have a built-in way of dealing with ad-hoc auxiliary hypotheses, Popper's first conventionalist stratagem. Each auxiliary hypothesis introduced ad-hoc must have a prior probability assigned to it, which is likely to be small, since it was not considered ex ante (by definition, it wasn't considered ex ante because it is ad-hoc).

To a Subjectivist Bayesian, data are valuable if they change one's posterior beliefs. If the data produce no change, then they produce no value, as one would not be willing to pay money to reveal the outcome of an experiment that could produce these data. When sharing data, the Subjectivist Bayesian cares about her beliefs about what other people believe. If one believes that data will change the posterior beliefs of others, then the data are expected to be valuable to them. The adoption of personal beliefs about the value of data to oneself and others can be seen as a special type of personal convention that may be justifiable to oneself, but not to others.

Objectivist

Since its inception, the Subjectivist approach has been surrounded by controversy. It was seen as adding unwarranted "subjectivity" into science, which was meant to be objective in principle. As a result it was strongly rejected by many philosophers and scientists, including most of those previously mentioned. For example, Fisher's problem with using Bayes' Rule for inference was due to "lack of experimental knowledge" about the prior (83). If there was some way of establishing the long-run frequency of information in

[4] The ability to accommodate ad-hoc auxiliary hypotheses is severely limited, either requiring logically omniscient priors or uncomputable functions (92), or if the true hypothesis is not in the support for a prior, a Bayesian has little hope of discovering the cause, regardless of the amount of data obtained (74). So, "thinking Bayesian" is only provisionally helpful, affording us formal inductive rules only as long as we are willing to pretend that we are logically omniscient.

the prior, Fisher would undoubtedly accept that. Without such proof, Fisher found no use for priors.

To deal with this, the Objectivist Bayesians argue that in many circumstances we should have a *unique* posterior belief that is determined by an objective prior, called the *informationless prior*. The intent was to provide a method where different observers arrive at the same conclusions (posterior distributions) after observing some data because they all agree that there is one correct prior. That is, the Objectivist Bayesians tried to come up with a method where prior belief is not purely determined by the opinions of the decision-maker.

Although the use of informationless priors can be useful, it is contradictory in many circumstances (90). This includes the three major approaches to informationless priors: 1) Laplace's (73) *principle of insufficient reason*, 2) the *invariance principle* of Jeffreys (93), and 3) the *principle of maximum entropy* of Jaynes (94). While the "informationless" priors they derive end up being quite simple (e.g., a uniform distribution), the justifications for these priors are quite complex and technical. Each has its advantages and disadvantages, and their derivations are out of the scope of this paper.

Just as the Subjectivists hold personal prior beliefs, Objectivist Bayesians make the choice of prior beliefs by appealing to reasonable principles that select informationless priors. Like all those mentioned before, the choice of 'informationless prior' has no unique stance among priors, making it a conventionalist decision.

2.5 Ethical Conventionalism

Each of the philosophical and statistical approaches discussed can be used to argue for data sharing policies that invoke some form of *conventionalism*: Poincare's chooses based on what is easiest to work with, Popper's chooses falsifiers, Lakatos' chooses successful predictions, Fisher's chooses statistically significant results, Neyman and Pearson's choose statistically significant results with high power, Mayo's chooses severe tests, Bayesians choose priors and value data that affect the posterior beliefs of others. None of these approaches resolve the conflict between theory and data without a conventionalist choice. As a result, any decision made by a researcher to report or omit data involves a convention that is implicitly or explicitly imposed on the reader. These conventions can be very effective, making the interpretation of the data easy, but can be very misleading if the reader either does not know about the reporting conventions or does not hold the same conventions. For example, someone holding Fisher's convention of only sharing statistically significant results could mislead a Bayesian, who may find very few data points (even as little as one data point that cannot even be given a p-value in Fisher's convention) very useful.

Thus, any approach that imposes conventions on the reader without making them explicit or allowing the reader to examine the data according to her own conventions is deceptive, and as a result, unethical. To avoid this possible deception, any ethical data sharing policy must not force conventions on the reader. Ethical data sharing allows the reader, if he or she desires, to reconstruct the original data without convention, or with minimal convention imposed by the data sharer. For this reason, even if the meaning of data are unclear, they still need to be documented. Even though the documentation is highly theory-laden, driven by the paradigm, the trace provided in the scientific record can be reconstructed by those with differing opinions about appropriate convention.

This perspective resolves important questions about data sharing, such as whether we should report 'warm-up' experiments that preceded our 'main' experiments that demonstrate a discovery. This is because, among the available conventions held in the community, data must be documented and made retrievable (although not necessarily included in the main analysis) if another member of the community would consider that data as evidence. The ethical data sharing policy depends on the members of the community and their conventions. If reasonable members of the community hold that conversations with strangers or spouses count as evidence, then these must be properly documented. Most often, documentation will begin when data are collected after instruments are developed, or during the process of developing measurement instruments. Thus, as most people in a community would find each experiment in a series of experiments informative, it would violate their conventions to omit these data. Given this policy, the reader can make a realistic judgment about the validity of the researcher's hypothesis without becoming confused or misdirected by too much, highly uncertain, or weak evidence. This also makes clear the community's duty to clearly articulate and compile the conventions of their members, as is done in the CONSORT statement (95).

2.6 Conclusion

I conclude by proposing three simple rules to guide ethical data sharing. These rules follow from the principle that any data sharing policy, to be ethical, must not impose conventions on the reader:

- *Rule 1*: Communicate research by your own conventions, making them as explicit as possible.
- *Rule 2*: Provide justification for and documentation of these

conventions. That is, for all data omitted that another person in the community could want, specify why this data was not shared.

- *Rule 3*: Provide a traceable account of other data not extensively detailed so that others can examine them according to their own conventions.

When the veil of epistemic and game-theoretic concerns is removed, data sharing is a question of ethics. It is about honesty. It is about not fooling ourselves and others.

Chapter 3

Rational Analyses of Data Sharing

3.1 Introduction

The work in this chapter extends the work of Overall (96), Ioannidis (3) and Shafto (60) by evaluating four normative questions of data sharing:

1. Are disconfirmations less informative than affirmations, and thus less worthy of sharing?

2. Are disconfirmations more likely to be error than affirmations, and thus less worthy of sharing?

Throughout this Chapter, I use Wason's 2-4-6 rule discovery task[1], the Neyman-Pearson decision-theoretic approach to hypothesis testing, and Bayesian analysis.

3.2 The Differential Diagnosticity Conjecture

One argument against sharing disconfirming data is the *Differential Diagnosticity Conjecture* (DDC): affirmation of a hypothesis is more informative than disconfirmation. As a result, disconfirming data need not be shared with the scientific community. The most general version of this conjecture can be formulated in Bayesian terms, to make the notion of 'informativeness' precise, in the following way: data that are high probability under our hypothesis change our posterior beliefs more than data that are low probability ($DDC1$).

[1] In the Wason task, a person is given the numbers 2-4-6 and told that they were generated by a hidden rule. The person can then propose new sets of three numbers and get feedback on whether those numbers also fit the rule.

3.2.1 A Simple Case

In the simplest case, suppose, as Popper does, a universal statement is our hypothesis: All swans are white (H). The complement to this hypothesis is a singular existential statement: There exists a non-white swan ($\neg H$). In this case, H assigns probability 1 to all white swans and probability 0 to all non-white swans. The informativeness of data will depend solely on the prior probability of H, $P(H)$. If $P(H) > 0.5$, then disconfirmation is more informative than affirmation because $P(H) - 0 > 1 - P(H)$. The opposite also holds. Thus, if we are willing to admit a prior probability of our hypothesis, $P(H)$, then $DDC1$ holds or does not hold in an arbitrary manner depending only on our prior beliefs.

3.2.2 Differential Diagnosticity Conjecture 1

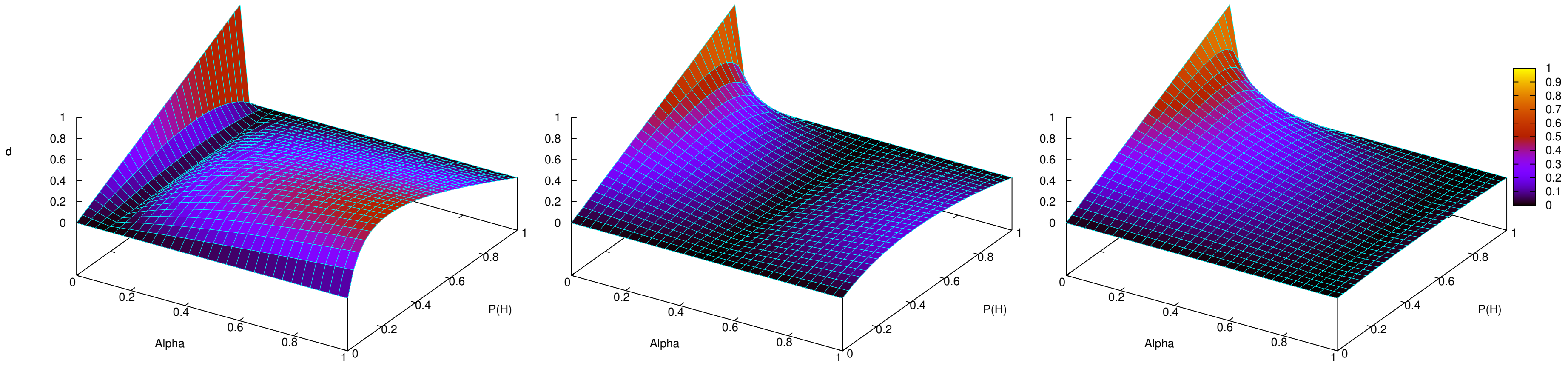
In a more general case one can allow H to assign high probability (but not necessarily 1) to white swans, and low (but not necessarily zero) probability to non-white swans. Consider the absolute difference between prior, $P(H)$, and posterior, $P(H|D)$, probabilities of some hypothesis H given some new data D as a measure of informativeness. This difference measure is called d and it has some nice properties which make it preferable to alternatives such as the log-ratio, log-likelihood ratio, and Carnap's r (97).

Let the informativeness of new data D, using the d measure mentioned above, called $d(H|D)$ normalized with the L_1 norm for simplicity ($\|X\|_1$ is a fancy way of saying the absolute value of X), be as follows:

$$d(H|D) = \|P(H|D) - P(H)\|_1 = \|\frac{P(D|H)P(H)}{P(D|H)P(H) + P(D|\neg H)P(\neg H)} - P(H)\|_1 \tag{3.1}$$

To simplify, substitute the symbols $\alpha = P(D|H)$, $\beta = P(D|\neg H)$ and $x = P(H)$ giving:

$$d(H|D) = \|\frac{\alpha x}{\alpha x + \beta(1 - x)} - x\|_1 \tag{3.2}$$



(a) $\beta = 0.1$

(b) $\beta = 0.5$

(c) $\beta = 0.9$

Figure 3.1: Figures showing $d(H|D)$ for varying values of β.

Figures 3.1a-c show three graphs of $d(H|D)$ with different values of $\beta = \{0.1, 0.5, 0.9\}$. It can be seen that one learns more when $P(D|H) = \alpha$ and $P(D|\neg H) = \beta$ are farther away from each other. That is, one learns more when the probability of the data under our hypothesis is very different from the probability of the data under other hypotheses, regardless of whether the hypotheses themselves are likely or unlikely a priori. The greater the difference, the more the data suggest one hypothesis over another, and the more we learn. This pattern does not substantially change depending on the prior probabilities of the hypotheses. This is intuitive.

What is not intuitive is that the rate of change is greater when $\alpha < \beta$ than $\alpha > \beta$. When $\alpha < \beta$, the graph is always downward sloping, meaning less probable data lead to more change in belief than more probable data. When $\alpha > \beta$, the graph is always upward sloping, meaning more probable data lead to more information than less probable data. However, the rates are asymmetric, when $\alpha < \beta$ the slopes are much sleeper than when $\alpha > \beta$.

What is the meaning of this asymmetry? This is merely due to the effect of the numerator or denominator on the value of a ratio. Suppose I have the function $f = \frac{y}{x}$. If I take $y = x = 1$, then decrement x by 0.9 then $f = 10$, however, if I increment y by 0.9 then $f = 1.9$. This asymmetry is thus merely due to the choice of numerator or denominator, or, more relevant here, whether I choose H or $\neg H$ for the numerator.

The log-likelihood ratio does not have this sensitivity. Thus $DDC1$ is true only in an arbitrary sense that I've chosen to put one hypothesis in the numerator over another, or that I've chosen not to use the log-likelihood ratio.

If $DDC1$ were properly translated in terms of the general Bayesian analysis described above, it would read as follows, and be correct: "I learn more when data are very likely under my hypothesis, and very unlikely under alternative hypotheses, than when the data are equally likely under both hypotheses. Alternatively, I learn more when data are very unlikely under my hypothesis and very likely under alternative hypotheses, than when the data are equally likely under both hypotheses. This relationship is perfectly symmetric."

3.2.3 Differential Diagnosticity Conjecture 2

Now, let's examine a second form of the DDC: data that are high probability under a hypothesis, that allows rejection of the complement of this hypothesis, are more informative than data that are low probability, that do not allow rejection of the complement. Or, more simply, data that reach statistical significance are more informative than data that do not reach statistical significance ($DDC2$).

One can formalize $DDC2$ with a hybrid Bayesian-Neyman-Pearson approach, which assigns probability to data given a hypothesis based on a rejection region, but also allows assigning probability to hypotheses directly. Suppose that some person proposes some hypothesis H. Consider two cases. When receiving affirmation, the data are in the rejection region for $\neg H$ (usually the null hypothesis), meaning the data have low probability under $\neg H$. Denote this $Data \in RR$ or "data in the rejection region." When receiving disconfirmation, the data are not in the rejection region for $\neg H$, indicating they have high probability for $\neg H$. Denote this $Data \in \neg RR$. Also assume both hypotheses have positive probability less than one and have non-zero Lebesgue measure.

If one substitutes conventional rejection rules, then $P(Data \in RR|\neg H) = \alpha$, where α is the usual significance level (0.05) for Type 1 Error, and $P(Data \in RR|H) = \beta$, where β is the Type 2 Error. The parallel conditional probabilities and Neyman-Pearson interpretation are summarized in the table below:

Variables	Conditional Probability	Neyman-Pearson
$1-\beta$	$P(Data \in RR\|H)$	Power
α	$P(Data \in RR\|\neg H)$	Type 1 Error
β	$P(Data \in \neg RR\|H)$	Type 2 Error
$1-\alpha$	$P(Data \in \neg RR\|\neg H)$	Negative Predictive Value
x	$P(H)$	Probability of Hypothesis

To tell whether affirmation is more informative than disconfirmation, we look at the ratio of the $d(H|D)$ measures given affirmation or disconfirming evidence.

$$\frac{d(H|conf)}{d(H|disc)} = \frac{\|P(H|Data \in RR) - P(H)\|_1}{\|P(H|Data \in \neg RR) - P(H)\|_1} \tag{3.3}$$

With this setup, if the ratio of the differences in equation 3.3 is greater than 1, then affirmation is more informative than disconfirmation, and if the ratio is less than 1, then the opposite holds. Thus, substituting the variables for conditional probabilities, the ratio of the differences is as follows:

$$\frac{d(H|Data \in RR)}{d(H|Data \in \neg RR)} = \frac{\|\frac{1-\beta}{(1-\beta)x+\alpha(1-x)} - 1\|_1}{\|\frac{\beta}{\beta x+(1-\alpha)(1-x)} - 1\|_1} = \|\frac{x(\alpha+\beta-1)+1-\alpha}{x(\alpha+\beta-1)-\alpha}\|_1 \tag{3.4}$$

From this, affirmation is more informative than disconfirmation whenever:

$$\|\frac{x(\alpha+\beta-1)+1-\alpha}{x(\alpha+\beta-1)-\alpha}\|_1 > 1 \tag{3.5}$$

I eliminate corner solutions where $x \in 0, 1$. The result is that affirmation and disconfirmation are equally informative whenever the following equation holds:

$$\beta = \frac{2\alpha - 2\alpha x + 2x - 1}{2x} \tag{3.6}$$

Affirmation is more informative than disconfirmation whenever the following inequality holds:

$$\beta > \frac{2\alpha - 2\alpha x + 2x - 1}{2x} \tag{3.7}$$

The graph of this inequality is shown in Figure 3.2. From both the equation and the graph, it is clear that as $P(H) = x$ increases, β must be higher (power must be lower) for $DDC2$ to hold. As α decreases, β must be lower (power must be higher) for $DDC2$ to hold.

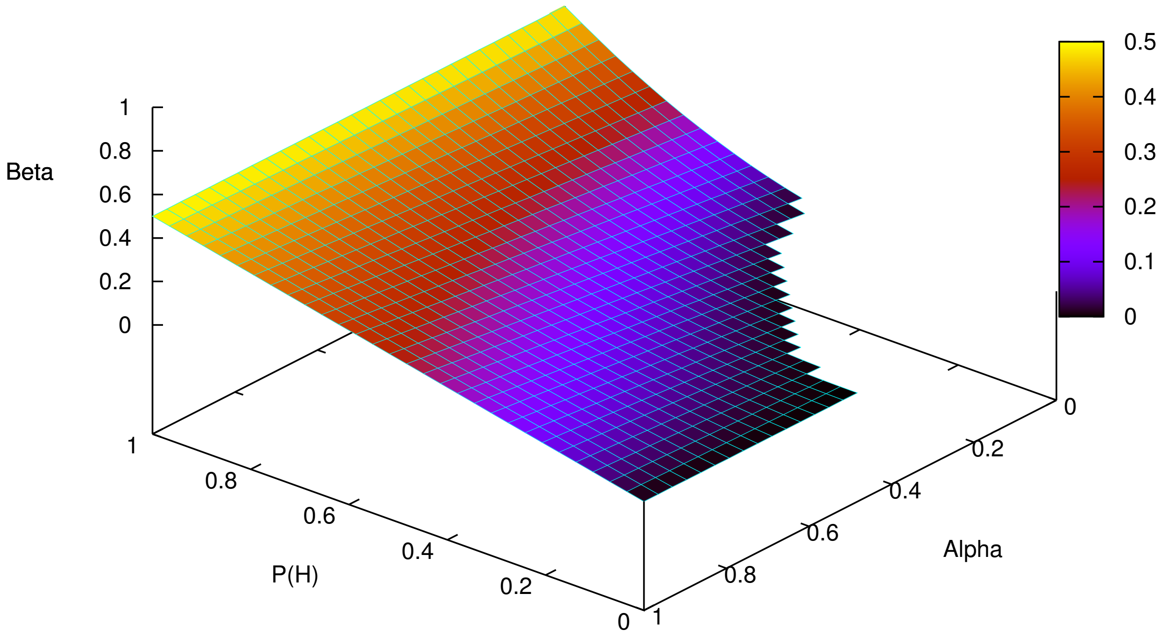


Figure 3.2: Graph showing when affirmation is more informative than disconfirmation.

For most social scientists, α is fixed at 0.05, but power varies. This special case can be worked out. The following equation holds:

$$\beta > 0.95 - \frac{0.45}{x} \tag{3.8}$$

The graph of this function is shown in Figure 3.3. As can be seen, $DDC2$ holds under the following conditions: when $P(H) = 0.47$, β must be greater than zero. When $P(H) = 1$, β must be greater than 0.5. If $P(H) < 0.47$, β can be any value. So, if β is between 0 and 0.5, affirmation can be more informative than disconfirmation. For any fixed alpha level, *the higher the value of or prior beliefs, the less likely one is to get more information from affirmation than disconfirmation.* It generally does not hold when β is very low and $P(H)$ is very high.

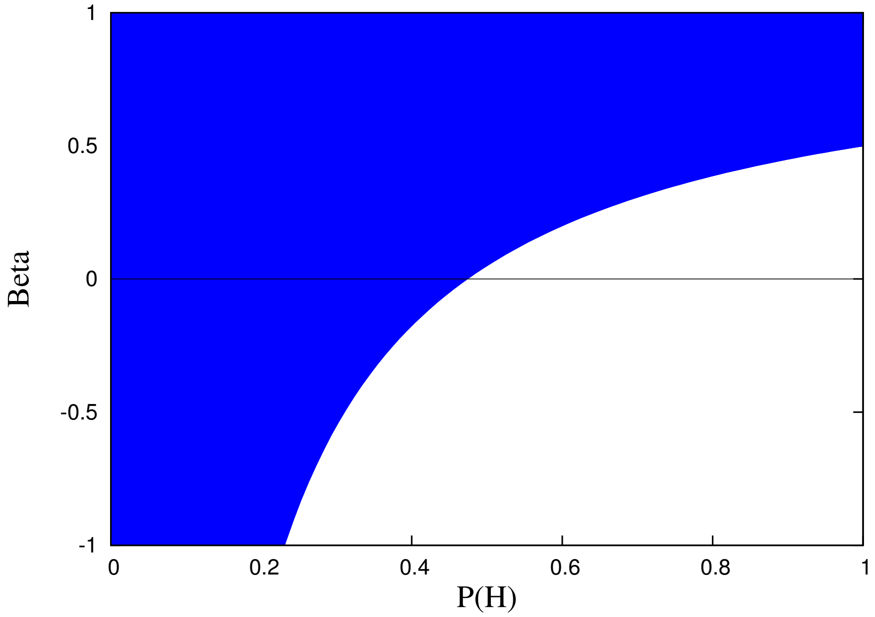


Figure 3.3: Graph of $\beta > 0.95 - \frac{0.45}{x}$ with shaded region where affirmation is more informative than disconfirmation for $\alpha = 0.05$.

Let's pick a few values for power to examine this further. Substituting Type 2 Error for power:

$$Power < 0.05 + \frac{0.45}{x} \tag{3.9}$$

Rearranging:

$$d(H|conf) > d(H|disc) \leftrightarrow \left\{ \begin{array}{ll} x < \frac{0.45}{Power - 0.05} & \text{if } Power > 0.05 \\ x > \frac{0.45}{Power - 0.05} & \text{if } Power < 0.05 \end{array} \right\}$$

If Power is close to 1, then $P(H)$ must be less than 0.5. If Power is close to 0.5, then $P(H)$ must be less than 1. If Power is less than 0.5, then $DDC2$ always holds. Since Power and $P(H)$ are usually low, $DDC2$ is likely to hold. The $DDC2$ has more validity than $DDC1$ in the Neyman-Pearson world we live in.

Why is this different than $DDC1$? The Neyman-Pearson approach assigns the lower error rate, α to the higher value hypothesis. In this sense, one cannot just reverse H and $\neg H$ to show that the converse also holds; one cannot switch H and $\neg H$ because we've assigned them different error rates based on their value.

In sum, if we consider H and $\neg H$ symmetric, then the differential diagnosticity conjecture ($DDC1$) has no meaning; *in general, data are more informative if they are assigned very different probabilities by different hypotheses, and this relationship is symmetric.* On the other hand, if one uses the Neyman-Pearson approach, and requires errors to be smaller for H than $\neg H$, then the differential diagnosticity conjecture ($DDC2$) holds for the cases social scientists usually face. As a result, $DDC2$ is a logical (although not particularly good) defense of discarding disconfirming data.

3.3 Blaming the Method

An alternative reason for throwing out disconfirming results is that a disconfirmation is more likely to be an error than an affirmation. I call this the *Blaming the Method Conjecture* (BMC). I again use the Neyman-Pearson approach allowing for prior probabilities or base rates.

The following table shows the behavioral commitment to make the judgment that $\neg H$ is false when $Data \in RR$ and $\neg H$ is true when $Data \in \neg RR$:

	H	$\neg H$
$Data \in RR$	A	B
$Data \in \neg RR$	C	D
	$P(H)$	$P(\neg H)$

The Type 1 Error is the probability of the data being in the rejection region given that the statistical hypothesis H is false:

$$\text{Type 1 Error} = P(Data \in RR|\neg H) = \alpha = \frac{B}{B+D} \qquad (3.10)$$

In contrast, this is not the same as the posterior probability of an error given the data are in the rejection region:[2]

$$P(error|Data \in RR) = \frac{B}{A+B} = \frac{\alpha P(\neg H)}{\alpha P(\neg H) + (1-\beta)P(H)} \tag{3.11}$$

Similarly, Type 2 Error is the probability of the data not being in the rejection region given that the statistical hypothesis H is true:

$$\text{Type 2 Error} = P(Data \in \neg RR|H) = \beta = \frac{C}{A+C} \tag{3.12}$$

This is, again, not equal to the posterior probability of an error given the data are not in the rejection region, which is equal to:

$$P(error|Data \in \neg RR) = \frac{C}{C+D} = \frac{\beta P(H)}{\beta P(H) + (1-\alpha)P(\neg H)} \tag{3.13}$$

Thus, this is a confusion of questions and confusion of inverses. When one says, “I throw out the data because errors are more likely to occur or disconfirmation than affirmation” one is saying:

$$P(error|Data \in \neg RR) > P(error|Data \in RR) \leftrightarrow \frac{B}{A+B} < \frac{C}{C+D} \tag{3.14}$$

This is not equivalent to the Type 1 Error rate being smaller than the Type 2 Error rate:

$$P(Data \in RR|\neg H) < P(Data \in \neg RR|H) \leftrightarrow \frac{B}{B+D} < \frac{C}{A+C} \tag{3.15}$$

To evaluate the correct question, one must find out when the posterior probability of error is more likely when failing to reject the null hypothesis than when rejecting it:

$$\frac{P(error|Data \in \neg RR)}{P(error|Data \in RR)} > 1 \tag{3.16}$$

[2] A quick note on why Tversky and Kahneman said that low power studies increase type 1 error (98). It can be seen from the equation above that, once one rejects the null hypothesis $Data \in RR$, the expected Type 1 Error is only equal to α if $\alpha P(\neg H)+(1-\beta)P(H) = P(\neg H)$. The lower the power $(1-\beta)$, the higher the Type 1 Errors among studies that reject the null hypothesis $Data \in RR$. Overall (96) called this *conditional Type 1 Error*. It is also interesting to note that conditional Type 1 Error also increases when one is less likely to pick correct hypotheses (i.e., increasing in $P(\neg H)$).

Rearranging:

$$\frac{\frac{\beta P(H)}{\beta P(H)+(1-\alpha)P(\neg H)}}{\frac{\alpha P(\neg H)}{\alpha P(\neg H)+(1-\beta)P(H)}} > 1 \leftrightarrow \beta(1-\beta) > \frac{\alpha(1-\alpha)(1-P(H))^2}{P(H)^2} \tag{3.17}$$

Using our usual $= 0.05$, and $P(H) = x$, the conjecture is true if:

$$0 > \frac{0.0475(1-x)^2}{x^2} - \beta(1-\beta) \tag{3.18}$$

The plot of this graph is shown in Figure 3.4:

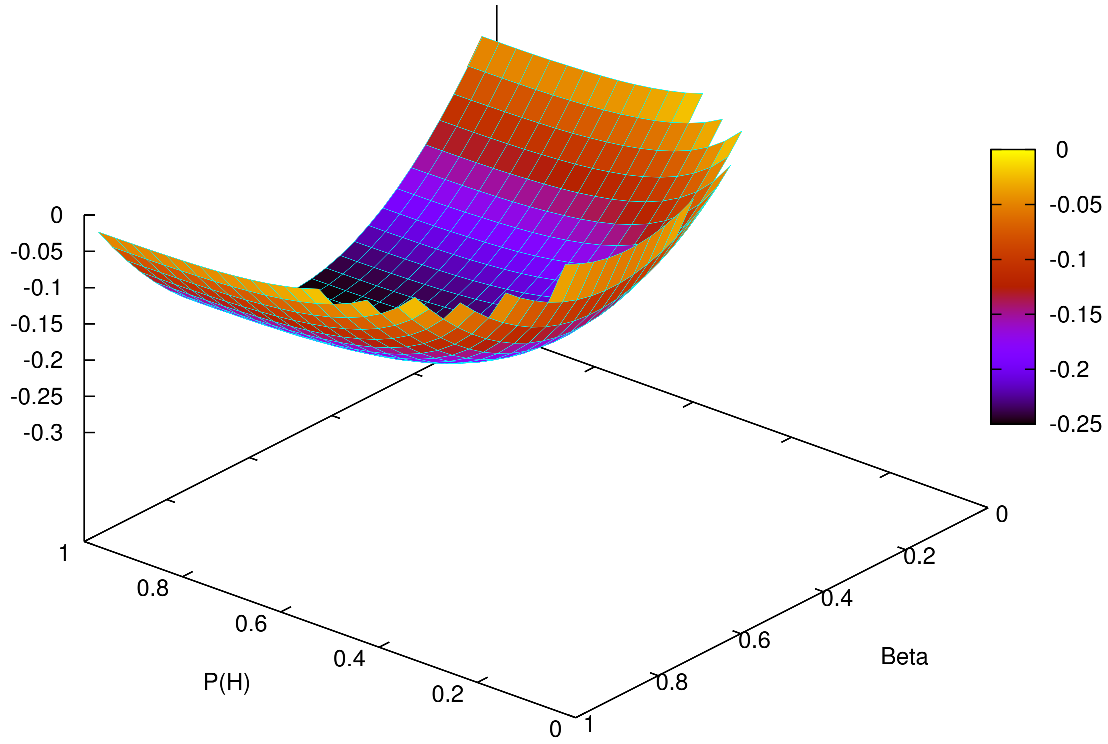


Figure 3.4: Graph showing region where disconfirmation is more likely to be error than affirmation.

Two facts can be gleaned from the equation and this graph:

1. The higher the $P(H)$, the more likely the BMC is to hold.
2. The larger $\|\beta - 0.5\|_1$ is, less likely BMC to hold.

Thus, the general conclusion is that: 1) if our hypothesis is rarely true, which is usually the case, then the BMC does not hold, 2) if the power is extremely high or low, then BMC is unlikely to hold. The BMC depends mostly on the prior probability that the hypothesis is true, and less so on the power of the test. Since we usually deal with circumstances where the prior probability that the hypothesis is true is low, BMC is usually false. In general, we should expect more errors when we get affirmation than disconfirmation when we are generally poor at choosing true hypotheses.

3.4 Conclusion

In this chapter I've analyzed two justifications for not sharing disconfirming data.

Section 1 evaluates the conjecture that disconfirming data are not informative as affirming data, and thus they need not be shared. This is called the differential diagnosticity conjecture (DDC). In the general case, the DDC is false: data are more informative if they are assigned very different probabilities by different hypotheses, and this relationship is symmetric. However, if one is unwilling to treat hypotheses symmetrically, then affirmation is more informative than disconfirmation.

Section 2 evaluates a different reason for not sharing disconfirming data: that disconfirming data are more likely to be error than affirming data. Using the Neyman-Pearson hypothesis testing framework, this is shown to be false in the cases social scientist usually face, where the probability of picking a true hypothesis is low. That is, in general we should expect more errors when we get affirmation than disconfirmation when we are generally poor at choosing true hypotheses.

Part III

Descriptive

Introduction to the Descriptive Analysis

In Part Three of the dissertation, the descriptive analysis examines whether people behave according to standards set forth in the normative analysis. Chapter Two concluded that, although there is no logical ground for determining whether data or theory is faulty when they conflict, data sharing policies that omit disconfirming data are unethical because they impose conventions on the reader, thus deceiving them. In the descriptive analysis, Chapter Four complements Chapter Two by examining whether lay participants judge that surprising disconfirmations are not worthy of being published because they are attributed to error, thus privileging theory over data and imposing conventions on readers.

Chapter Three tells us that when the value of hypotheses is symmetric and the error probabilities of false positives and false negatives are equal, affirmation and disconfirmation provide the same information. In addition, one should expect more errors from affirmation than disconfirmation when one is generally poor at choosing true hypotheses. Chapter Five complements Chapter Three by evaluating whether lay participants adhere to these normative principles. Using the Wason 2-4-6 rule discovery task, participants are given a known rate of error in feedback for hypotheses they test, and are given the chance of sharing the data they collect with another participant also trying to solve the rule. If they are rational, then they should attribute error to feedback whenever they strongly believed their hypothesis a priori and the feedback was disconfirming, or if they strongly disbelieved their hypothesis a priori and the feedback was affirming.

The descriptive research also has a reflexive or 'meta' purpose. The approach and experiments described in Part Three reflect my problems and training along with their natural evolution. As a result, the research itself is an example of the phenomenon to be described, as the experiments frequently propose hypotheses, fail in their predictions, and then invoke error to explain the failure. By making this process transparent, not only in the behavior of the subjects of the experiments but also in the experimenter, others can learn from both the method and results. The research exposes the not-frequently-discussed but very important elements that lead to the file-drawer problem, where 'pre-tests' and 'pilot-tests' are flexibly defined and reported with the benefit of hindsight and potentially distorted by pressures to publish. If this process is hidden, it cannot be addressed, discussed, and improved.

Chapter 4

Surprises, Error, and Data Sharing

4.1 Introduction

Every experiment has the potential for unexpected results—otherwise it would not be worth conducting [1]. When surprises arise, scientists need to account for them. Those results may suggest new theories. Or, they may just raise questions about the soundness of the experimental design—and the auxiliary hypotheses needed to interpret the data that it produces (59). In Psychology, those questions might include when research participants understand the instructions and stimuli as intended, whether the set-up conveyed unintended clues or incentives, and whether mistakes were made in data entry or statistical analysis. The weaker the empirical or theoretical support for these assumptions, the more the interpretation of unexpected results must rely on scientific judgment (99).

Researchers' confidence in that judgment should be shaken by knowing that they already answered these questions as best they could when designing a study. The need to make such inferences acknowledges that every study requires an assessment of construct validity, as researchers simultaneously evaluate their substantive theories and their methodological assumptions (100). Unexpected results require particularly judicious assessments. If researchers accept those results uncritically, then they may allow flawed methods to undermine valuable theories. If researchers challenge the methods hypercritically, then they may unreasonably defend flawed theories.

The history of Physics provides a famous example of making progress by discounting surprising experimental results. While attempting to measure the charge of an electron, Nobel laureate R.A. Millikan discarded multiple unexpected data points, confidently attributing them to error in his

[1] We thank the late Robyn Dawes for reminding us of this principle.

experimental apparatus. Most of those instances occurred during an ambiguously defined "warm-up period" where he "gradually refined his apparatus and technique in order to make the best measurements." (101, p. 13) However, Millikan also rejected later (post-warm-up) observations where "there were no obvious experimental difficulties that could explain the anomaly." He attributed these anomalies to nothing more explicit than "something wrong with the thermometer." (39, p. 13) Later work found that Millikan's intuitions were generally right, even though he did not articulate reasons for them—and, indeed, could not have known the source of the anomalies given scientific knowledge at the time. (His experimental apparatus was unreliable with charges greater than 30e.) Had Millikan pursued the anomalies, he would have delayed studies that made important contributions to Physics, despite their flaws.

As in Millikan's case, it may be necessary to "explain away…odd results" to avoid having research "instantly degenerate into a wild-goose chase after imaginary fundamental novelties." (Michael Polanyi quoted by (102, p. 63)) Psychological research has identified processes that can support and undermine such judgments. On the one hand, surprising results can induce a greater subjective need for better explanations, prompting deeper probing and reflection (103). On the other hand, such results can prompt "explaining away" results that disconfirm favored theories by unfairly attacking auxiliary hypotheses (104).

In a less happy example from Physics, Rene Blondlot's purported discovery of a new type of electromagnetic radiation, called n-rays, "touched off a wave of self-deception that took years to subside." (105, p. 170) His supporters included respected physicists who uncritically reported expected effects when they placed n-ray sources (e.g., gas burners used for lighting, heated silver or sheet iron) in front of electric spark generators, while accusing scientists who failed to observe those effects (106) of poor training.

Both Millikan and Blondlot attributed unexpected results to measurement error. Such *error model*[2] explanations include attributing unexpected data to uncontrolled, unintended, or unknown experimental artifacts. Error models can capture valid intuitions and keep science moving until deeper understanding arrives, as with Millikan. However, as in the case of Blondlot and his supporters, error models can also immunize hypotheses against valid challenges from disconfirming data (102; 107; 108; 109). Psychological processes that could enable error-model thinking include blaming the method (110), biased assimilation (111), confirmation bias (112), and belief perseverance (113).

[2] More technically, an error model is a causal explanation that renders the substantive theory conditionally independent of the data when invoked, thus making the data not informative for the theory.

Error models can be created in foresight (for potential surprises) or hindsight (for actual ones). The voluminous research on hindsight bias (114; 115) suggests that the two perspectives will produce rather different error models. In hindsight, explanations will naturally focus on the observed outcome, whereas foresight will consider possible outcomes (116). Accounts of hindsight bias can be derived from many theories of human memory, judgment, and formal reasoning, including mental model "rejudgments" (117), q-morphisms (118), sense-making (119), causal judgment (120), Bayes nets (121), and causal models (122; 123). In general, the psychological evidence on hindsight bias echoes the creeping determinism account originally proposed by Fischhoff (124), in which learning about an outcome modifies one's prior beliefs to make it seem more likely. Generally speaking, the more unexpected the outcome, the stronger this sense-making process and the greater the resultant bias will be —unless the surprise is so extreme or obviously random that one cannot generate an acceptable causal explanation (119).

Treating expected and unexpected results differently creates the risk of accepting weak, but welcome results uncritically, while learning too little from potentially informative surprises—leading to "well-intentioned scientists making well-intentioned (although biased) decisions. . . leading to incorrect results" (68, p. 58). The Large Interferometer Gravitational-wave Observatory (LIGO) represents one ambitious attempt to reduce that risk. Members of this "big science" project specify a priori rules for removing spurious data prior to statistical analysis, so that these decisions are not unduly affected by the data themselves (125; 126). Those rules seek to balance those scientists' desire to include as many of their (very expensive) observations as possible, while excluding spurious ones that could undo their work. However, even in that mature science, it is hard to anticipate all possible problems (e.g., a private plane flying into the restricted air space over an interferometer, perturbing the observations), making some post hoc interpretation inevitable.

Excluding data is straightforward when outright fabrication is discovered (127). It is much harder in the situations usually faced by scientists, where, "data are not published in good journals, or even in bad journals" but instead are "sitting in my file drawer" (68, p. 58), after being rejected because they "didn't work [or were] pilot studies" (p. 58).

Open-access data advocates argue that all data must be shared, so that the community of scientists can evaluate their relevance directly and discern the "story of the failures that make the successes possible" (128, p. 15). Some claim that unshared data are "experimental failures" (129, p. 24). Yet, for working researchers to adopt these norms, they need to feel that they are more like Blondlot (potentially mistaken) than Millikan (potential Nobel laureates). They also need peers who value learning from failures as well as successes.

A priori rules are needed most when the differences are greatest between the error models produced in foresight and hindsight, namely, when the evidence disconfirms researchers' hypotheses, prompting them to generate flexible alternative hypotheses that may overfit the data (130; 20). The present studies examine the role of error models in interpreting and sharing experimental results, focusing on foresight-hindsight differences.

As a platform for these studies, we use a design introduced by Slovic and Fischhoff (116). It has participants assess the probability of replicating the initial observation of a hypothetical experiment. Foresight participants assess that probability for two possible outcomes. Hindsight participants are told that one of those outcomes was, in fact, observed. The conditional probability of replication should be the same in both conditions. However, Slovic and Fischhoff found that hindsight participants see replication as more likely, consistent with being less able to see how the initial study could have turned out otherwise. We begin by repeating Slovic and Fischhoff's original study, in order to establish a baseline for the following studies, examining how people account for more and less expected results. Incidentally, we assess the robustness of a widely cited study, thirty-plus years later.

4.2 Experiment One

4.2.1 Method

Participants evaluated the four hypothetical studies presented in Experiment One of Slovic and Fischhoff (116), using their stimulus materials. These studies tested whether: 1) a virgin rat would exhibit maternal behavior following a blood transfusion from a mother rat, 2) seeding a hurricane with silver-iodide crystals would diminish its wind velocity, 3) goslings would be imprinted on a duck if exposed to its quacking before hatching, and 4) children could take another person's perspective when judging the position of a dot on a large Y. Foresight participants first assessed the probability of each outcome occurring, then its probability of replication on all, some, or none of 10 additional observations — should it be observed on a single initial observation. Hindsight participants were told that one of the two outcomes had occurred, and then assessed its probability of replication. The design was 4 (study: rat, hurricane, duck, Y-test) by 2 (time: foresight vs. hindsight) by 2 (outcome: A or B) with repeated measures on the first factor and repeated measures on the last factor in the foresight condition, whose participants gave probabilities of replication for both outcomes.

Participants

All 268 participants were paid volunteers who responded to an Amazon Mechanical Turk (MTurk) ad offering them 1 dollar for participation in a 7-minute study. Mason *et al.* (131) found that, when paid more, MTurk participants work longer but do not perform better (in terms of accuracy). Horton *et al.* (132) found that MTurk participants replicated results from several classic judgments studies originally conducted with traditional (e.g., student) samples. A two-part attention filter (133; 134; 135) at the beginning of the experiment assessed whether participants were paying attention. Only the 173 participants who passed both its parts (one easier, one harder) were included in the analysis. According to participants' reports, their average age was 32 years old (range = 18 – 81) and 56.6% were women.

4.2.2 Results

Table 4.1 reveals a clear hindsight bias in responses to the first hypothetical study. Foresight participants said that, if the first virgin rat demonstrated maternal behavior (after receiving a blood transfusion from a mother rat), there was a 27.8% chance of that happening on all 10 subsequent cases. Hindsight participants told that the initial case had turned out that way gave a 49.4% probability to consistent replication. The corresponding means in (116) were 30% and 44%, respectively. The mean probability of no replications was 32.8% in foresight and 16.2% in hindsight (in the previous study, 29% and 7%). The other outcome (B) showed complementary results, also fairly similar to those before.

The other three hypothetical studies revealed similar patterns (Tables 4.2-4.4): An initial observation was seen as significantly more likely to be replicated consistently when it was reported to have happened (hindsight) that when it was considered as a possibility (foresight). It was also judged significantly less likely never to be repeated. In each case, the means were similar to those in (116) — although that is not a necessary condition for replicating the pattern of responses.

4.2.3 Discussion

Slovic and Fischhoff (116) found that people see the results of the first observation of a study as more likely to be replicated in hindsight than in foresight. In this exact replication of their Experiment One, that result held true. In their Experiment Two, Slovic and Fischhoff (116) found similar results when foresight participants considered only one of the two possible outcomes, rather than both (as in Experiment One), indicating that their lower confidence in replication was not due to focusing less on each outcome.

Table 4.1: Judged probability that the initial observation will replicate in all, some, or none of 10 replication trials for virgin rat study.

Outcome	Response	Foresight M (SD, N)	Hindsight M (SD, N)
Outcome A (maternal behavior)	All	28 (29, 61)	49 (32, 62)
	Some	39 (29, 61)	34 (28, 62)
	None	33 (32, 61)	16 (20, 62)
Outcome B (no maternal behavior)	All	47 (37, 61)	67 (34, 50)
	Some	34 (29, 61)	23 (28, 50)
	None	20 (24, 61)	10 (20, 50)

Table 4.2: Judged probability that the initial observation will replicate in all, some, or none of 10 replication trials for hurricane study.

Outcome	Response	Foresight M (SD, N)	Hindsight M (SD, N)
Outcome A (intensity increases)	All	40 (33, 61)	52 (30, 62)
	Some	34 (27, 61)	34 (26, 62)
	None	27 (28, 61)	15 (16, 62)
Outcome B (intensity decreases)	All	37 (33, 61)	51 (34, 50)
	Some	36 (28, 61)	36 (31, 50)
	None	27 (28, 61)	14 (18, 50)

These participants had no natural reason to prefer observing either outcome, unlike actual investigators, who may care deeply about how studies turn out. However, these participants did have natural expectations, expressed in the probabilities that foresight participants gave for the possible outcomes of the first observation. As seen in Table 4.5, one outcome was significantly more likely for three of the four hypothetical studies (virgin rat, hurricane, Y-test), whether measured by the mean probability or the percentage of participants assigning a probability greater than 50%.

From both perspectives, the most likely of the eight outcomes was Outcome A in the Y-test study. As seen in Tables 4.1-4.4, that outcome also produced the weakest hindsight bias, as though it was so strongly expected that reporting its occurrence had relatively little impact (although it was not so likely as to encounter a ceiling effect). Conversely, reporting Outcome B in the Y-test study, the least expected of the eight, had a particularly large hindsight effect, indicating willingness to abandon outcome A, given a single contrary observation. These results are consistent with participants generating causal explanations for explaining whatever they observe, with those more often being error models when they observe the unexpected.

Experiments Two through Five examine these processes, as revealed in

Table 4.3: Judged probability that the initial observation will replicate in all, some, or none of 10 replication trials for gosling study.

Outcome	Response	Foresight M (SD, N)	Hindsight M (SD, N)
	All	36 (33, 61)	56 (34, 62)
Outcome A (approaches goose)	Some	34 (29, 61)	34 (19, 62)
	None	30 (32, 61)	10 (12, 62)
	All	49 (34, 61)	73 (32, 50)
Outcome B (approaches duck)	Some	34 (30, 61)	19 (24, 50)
	None	17 (24, 61)	8 (16, 50)

Table 4.4: Judged probability that the initial observation will replicate in all, some, or none of 10 replication trials for the Y-test study.

Outcome	Response	Foresight M (SD, N)	Hindsight M (SD, N)
	All	45 (34, 61)	56 (31, 62)
Outcome A (places dot in area A)	Some	37 (31, 61)	33 (28, 62)
	None	18 (23, 61)	10 (12, 62)
	All	18 (24, 61)	31 (28, 50)
Outcome B (places dot in area B)	Some	34 (32, 61)	46 (31, 50)
	None	48 (36, 61)	23 (20, 50)

attributions for results of the Y-test study, the one with the most and least expected initial observations. If participants use error models to accommodate unexpected results, then they should invoke explanations such as "experimental error" or "methodological problems" more often with Outcome B than with Outcome A.

4.3 Experiment Two

4.3.1 Method

Experiment Two replicated Experiment One, with two differences: (a) Participants considered just one hypothetical study, the Y-test, in order to elicit a fuller, more focused set of beliefs.[3] (b) Participants assessed the probability that each of four causes accounted for the results. Thus, the design was 2 (foresight vs. hindsight) by 2 (expected outcome [area A] vs.

[3] Experiment Two originally included all four studies from Experiment One. However, for the sake of simplicity, we decided to focus on the Y-test results, which used the expected and unexpected outcomes, hence best fit our research interests.

Table 4.5: Mean foresight probability and proportion of probabilities greater than 50 for each outcome of the initial observation. One-sample t-test compares P(A) to 50%.

	P(A)	P(B)			
Study	M (SD)	M (SD)	One-Sample t-test	P(A)>50	P(B)>50
Virgin Rat	40 (25)	59 (26)	t (60) = 2.66, p = 0.01	14/61	32/61
Hurricane	52 (27)	40 (26)	t (60) = 2.90, p = 0.005	29/61	12/61
Gosling	51 (28)	52 (27)	t (60) = 0.65, p = 0.52	28/61	25/61
Y-test	65 (28)	24 (22)	t (60) = 4.14, p < 0.001	39/61	5/61

unexpected outcomes [area B]), between-subjects.

Participants

For Experiment Two, all 664 participants were paid volunteers who responded to an Amazon MTurk ad offering them $1 for participation in a 7-minute experiment. Experiment Two used the same attention filter as Experiment One, with 468 individuals (70%) passing both tests. Their average age was 31 years old (range: 18 – 63); 50% were women.

Materials

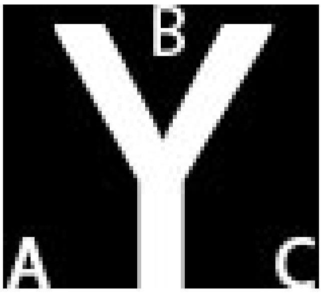
In Experiment Two, all participants received the same introductory instructions as used in Slovic and Fischhoff (116), followed by their description of the Y-Test study.

> In the pretest of an experiment that she intends to run in the future, an experimenter will place a 4-year-old child in front of an easel with a large Y on it, with a dot in the lower left-hand third of the letter. The child will then be taken around to the back of the easel where he will see another Y. He will be asked to draw a dot in the "same position" on that Y as the one he had just seen.
>
> The possible outcomes are (a) the child places a dot in Area A (the lower left-hand third), (b) the child places a dot in Area B (the upper third), or (c) the child places a dot in Area C (the lower-right hand third).

Participants were then asked for predictions and attributions (using the Area condition as an example below). The brackets provide our interpretation of each response.

Foresight.

Figure 4.1: Image of Y shown to participants.



> If the child places a dot in Area A, what is the probability that:(Note: These four probabilities should total 100%.)
>
> 1. The child's understanding of the experimenter's instructions caused the child to place the dot in Area A. [Valid Method]
> 2. Some error in the experiment caused the child to place the dot in Area A. [Invalid Method]
> 3. Random chance caused the child to place the dot in Area A. [Chance]
> 4. There was some other cause not already mentioned above. [Other]

Hindsight. The instructions for hindsight participants differed in reporting the first observation:

> Result: The child placed a dot in Area A (the lower left-hand third).

4.3.2 Results

Table 4.6 shows median probabilities assigned to the four causal explanations.[4] Based on the results of Experiment One (and (116)), we treat area A as expected and area B as unexpected. We use medians rather than means because of skewed distributions and outliers. We conducted median regressions (136; 137) of the two experimental factors on the probabilities assigned to the four potential causes, using non-parametric bootstrap to estimate standard errors (138).

For each cause, results were generally in the predicted direction, although not always significantly so. The experimental method was judged more valid when the initial observation was expected rather than unexpected (i.e., the

[4] We also asked exploratory questions not reported here, regarding participants' overall judgments of the strength of the experimental design and how the results should be treated.

Table 4.6: Median probability, standard error, and sample size for four causal attributions for Experiment Two. Md = median. Sample sizes are (H = hindsight; U = unexpected): $HU = 115$; $FU = 122$; $HE = 118$; $FE = 114$.

Possible Cause	Condition	Foresight Md (SE)	Hindsight Md (SE)
Valid Method	Expected (A)	60 (7.7)	60 (7.5)
	Unexpected (B)	30 (5.4)	50 (3.9)
Invalid Method	Expected (A)	5 (2.5)	4.5 (1.5)
	Unexpected (B)	10 (1.8)	10 (1.2)
Chance	Expected (A)	20 (2.1)	10 (1.6)
	Unexpected (B)	20 (2.3)	20 (2.2)
Other	Expected (A)	8.5 (2.3)	9.5 (2.3)
	Unexpected (B)	20 (3.6)	12.5 (3.9)

child placed the dot in area A), in both foresight (60% vs. 30%) and hindsight (60% vs. 50%). The corresponding main effect for the difference between the probability assigned to Valid Cause in the expected and unexpected conditions (difference = -30; 95% CI: [-50, -10]) was statistically significant, $t(464) = 3.00$, $p < 0.05$, $d = 0.14$. There was also a non-significant interaction, with unexpected evidence reducing the probability assigned to Valid Method more in foresight than hindsight (difference = 20; 95% CI: [-6, 46]), $t(464) = 1.53$, $p > 0.05$, $d = 0.07$.

Conversely, participants assigned similar probabilities to the method being Invalid after an unexpected observation than after an expected one, in both foresight (5% vs. 10%) and hindsight (4.5% vs. 10%), a non-significant main effect (difference = 5.0; 95% CI: [-1.3, 11.3]), $t(464) = 1.59$, $p > 0.05$, $d = 0.07$; nor was there the expected interaction (with larger effects in hindsight), $t(464) = 0.00$, $p > 0.05$, $d = 0.00$.

Chance was assigned a greater role with unexpected results in hindsight (20% vs. 10%), but not in foresight (20% vs. 20%), reflected in both a main effect of hindsight (difference = -10; 95% CI: [-14.9, -5.1]), $t(464) = 4.08$, $p < 0.05$, $d = 0.19$, and an interaction between the two factors, $t(464) = 2.11$, $p < 0.05$, $d = 0.10$. Finally, Other Causes received higher probabilities with unexpected results, in both foresight (20% vs. 8.5%) and hindsight (12.5% vs. 9.5%), with a significant main effect (difference = 10; 95% CI: [-0.9, 19.1]) $t(464) = 2.20$, $p < 0.05$, $d = 0.10$; but no interaction.

4.3.3 Discussion

As predicted by the assumption that people rely on error models to explain unexpected outcomes, participants who considered the less expected result (the child placing the dot in area B) assigned significantly lower probabilities to the method being valid and significantly higher probabilities to chance and to other causes. They did not assign significantly higher probabilities to an invalid method, although there was a trend in this direction. These patterns were observed in both foresight and hindsight, except for one significant interaction: chance was assigned a greater role for unexpected results in hindsight, but not foresight. Thus, it appears that people can invoke error model thinking in foresight as well as they can in hindsight—if asked to do so.

One possible interpretation of these results is that error models are equally available in foresight and hindsight—if people explicitly consider how they will account for an unexpected result. However, that assessment may not happen as naturally in foresight. That could be true for actual researchers, if they do not press as hard as they might in foresight to think about possible confounds, as well as for participants in Experiment One, asked to consider, but not explain, an unexpected first observation. In contrast, the attribution tasks of Experiment Two embody a kind of debiasing procedure, making potentially useful alternative explanations more available in foresight—and the outcomes less likely. Without having elicited probabilities of replication in Experiment Two, we cannot know. Experiment Three does that, adding the probability question from Experiment One to the attribution task of Experiment Two.

We predicted that these trends would be stronger with the unexpected observation, insofar as it creates a greater need to explain the result. The lack of significant interactions with any of the non-chance causes (Valid Method, Invalid Method, Other Causes) suggests that participants found ways to deal with the unexpected result more thoroughly in hindsight.

Experiment Three looks more closely at participants' use of error models by having them generate a cause for the initial observation on their own, using an open-ended format. We code this cause into the four categories of Experiment Two. After providing that cause, participants perform the attribution task of Experiment Two, with a refinement of the Valid Method category, phrasing more clearly it in terms of the hypothesis guiding the investigator, that the child could rotate the image mentally, assuming that a valid method was needed for it to emerge. Experiment Three also offers participants an open-ended opportunity to explain their reasoning.

Finally, we extend the experimental task by asking participants to imagine themselves as the investigators, then say how they would treat the results of a study with responses from an additional 10 children, in terms of whether the data should be published, replicated, or discarded. If participants believe that an unexpected result is due to error, then they should see it as not worth publishing because it does not properly test the hypothesis—just as Millikan discarded measurements that he thought were contaminated by uncontrolled variation, such as "something wrong with thermometer."

4.4 Experiment Three

Experiment Three changes the methodology of Experiment Two in four ways: (a) Participants generated their own causal explanations before assigning probabilities to pre-defined categories. (b) We clarified those categories by explicitly offering the non-error explanation (the child rotated the image). (c) We asked participants how they would treat research results in terms of publication, imagining themselves as scientists. (d) We added a manipulation check.

4.4.1 Method

Experiment Three followed Experiment Two, with a 2 (foresight vs. hindsight) by 2 (Area: A, B) design. We added an open-ended attribution task and a question about data sharing and revised the structured attribution question.

Participants

All 448 participants were paid volunteers who responded to an Amazon MTurk ad offering $1 for participation in a 7-minute experiment. Experiment Three used the same attention filter as before, with 359 (80%) individuals passing. Fifteen (3%) failed the manipulation check, indicating area C. Among the remaining 344 participants, the average age was 32 years old (range: 18 – 68); 151 were women (44%).

Materials

The instructions followed Experiment Two, with these modifications.

For Foresight: After reading the study's design, participants were asked to explain one potential outcome of the first observation.

> Please explain why you think the child could place the dot in Area A [or B]. (open-ended) [OpenCause]

Participants then answered the modified version of the attribution question:

> What is the probability that? (Note: These four probabilities should total 100%.)
>
> 1. The child's ability to mentally rotate the image caused the child to place the dot in Area A. [Rotate]
> 2. Some error in the experiment caused the child to place the dot in Area A. [Invalid Method]

3. Random chance caused the child to place the dot in Area A. [Chance]
4. There was some other cause not otherwise mentioned. [Other]

Participants then answered the probability-of-replication question from Slovic and Fischhoff (1977) and Experiment One. Participants next answered a new question asking how they would treat those observations:

> If the replication of this experiment with 10 additional children comes out the way you expect, which of the following actions would you recommend that the scientist take:
>
> 1. Collect more data before publishing [MoreData]
> 2. Publish without collecting more data [Publish]
> 3. Do not publish any of the data [NoPublish]

An open-ended question asked them to explain this recommendation. Finally, participants completed the following manipulation check:

> Where did the child put the dot?
>
> 1. Area A
> 2. Area B
> 3. Area C

For Hindsight participants, the tasks were the same, except that they were told, “The child placed the dot in Area A [or B].”

4.4.2 Results

Table 4.7 shows judgments of the three replication possibilities. As in Experiment Two, we used median regression with non-parametric bootstrapped standard errors for statistical tests.

Experiment Three replicates the previously observed hindsight effect, but only for the expected outcome. Participants told that the first child had placed the dot in the expected place (A) gave higher probabilities to that happening on the next 10 observations than did participants who considered that outcome as a possibility (50% vs. 30%). Consistent replication of the unexpected result (B) was, however, equally likely in hindsight and foresight (10% vs. 10%). The corresponding interaction was marginally significant (difference = -20; 95% CI: [-43, 2.6]), t (339) = 1.77, p = 0.08, d = 0.10. Conversely, the expected result was judged less likely never to replicate (on

Table 4.7: Median probability (Md) and standard error (SE) for expected replication for Experiment Three. Sample sizes are (H = hindsight; U = unexpected): $FU = 71$; $HU = 87$; $HE = 101$; $FE = 84$.

		Foresight	Hindsight
Expected Replication	Condition	Md (SE)	Md (SE)
All	Expected (A)	30 (8.2)	50 (7.5)
	Unexpected (B)	10 (2.4)	10 (2.9)
Some	Expected (A)	45 (9)	30 (5.5)
	Unexpected (B)	50 (5.8)	50 (5.8)
None	Expected (A)	6.6 (2.2)	1 (1.6)
	Unexpected (B)	20 (4.1)	25 (4.8)

Table 4.8: Mean number of participants choosing each category. Publish judgments could be in one of three categories, which can be modeled using a Dirichlet distribution. Standard errors generated from 10000 simulations from a posterior Dirichlet distribution (139) with improper Dirichlet(0,0,0) priors.

		Foresight	Hindsight
Publishing Recommendation	Condition	% (95% CI)	% (95% CI)
More Data	Expected (A)	0.79 [0.69, 0.87]	0.82 [0.74, 0.89]
	Unexpected (B)	0.85 [0.75, 0.92]	0.83 [0.75, 0.90]
Publish	Expected (A)	0.18 [0.11, 0.27]	0.17 [0.10, 0.25]
	Unexpected (B)	0.09 [0.03, 0.16]	0.09 [0.04, 0.16]
No Publish	Expected (A)	0.04 [0.01, 0.09]	0.01 [0.00, 0.04]
	Unexpected (B)	0.07 [0.02, 0.14]	0.08 [0.03, 0.14]

the next 10 observations) in hindsight that in foresight (1% vs. 7%), whereas the unexpected result was judged more likely never to replicate once it had been observed than when it was just a possibility (25% vs. 20%). Here, too, though, the interaction was not statistically significant (difference = 10; 95% CI: [-3, 23]), t (339) = 1.57, p = 0.12, d = 0.08.

Data Sharing Judgments

As seen in the top section of Table 4.8, participants overwhelmingly recommended collecting more data before publishing, for both expected and unexpected results. Among the minority who recommended publishing, the rate was twice as high with expected results than with unexpected ones, although the difference was not significant. Conversely, not publishing was more common with unexpected results; again, not significantly so. These patterns were the same in hindsight and foresight (i.e., with no significant interactions).

Causal Attributions

Table 4.9 shows the probabilities assigned to the four explanations of the initial observation. Rotate attributes the first observation to the child's having (or lacking) the ability to rotate the display mentally (as revealed by a valid method). As expected, the probabilities assigned to that explanation were higher when results were consistent with that ability (A vs. B), in both foresight (58% vs. 20%) and hindsight (67% vs. 25%). The main effect (difference = -40; 95% CI: [-55, -25]) was statistically significant, t (339) = 5.30, p < 0.05, $d = 0.29$. However, that difference was not significantly greater in hindsight (difference = -7; 95% CI: [-24, 10]), t (339) = 0.80, $p >$ 0.05, $d = 0.04$.

Table 4.9: Median probability (Md), standard error (SE) for four causal attributions for Experiment Three. Sample sizes are (H = hindsight; U = unexpected): $FU = 71$; $HU = 87$; $HE = 101$; $FE = 84$.

Possible Cause	Condition	Foresight Md (SE)	Hindsight Md (SE)
Rotate	Expected (A)	58 (6.8)	67 (5.2)
	Unexpected (B)	20 (3.5)	25 (2.8)
Invalid Method	Expected (A)	5 (1.6)	1 (1.4)
	Unexpected (B)	10 (3.1)	10 (2.6)
Chance	Expected (A)	20 (2.6)	10 (2.1)
	Unexpected (B)	25 (3.2)	20 (2.0)
Other	Expected (A)	10 (2.3)	5 (2.2)
	Unexpected (B)	20 (3.4)	20 (2.8)

Although there was a trend for participants to see the method as Invalid after an unexpected observation than after an expected one, in both foresight (5% vs. 10%) and hindsight (1% vs. 10%), this main effect (difference = 5.0; 95% CI: [-1.6, 11.6]) was not significant t (339) = 1.54, $p > 0.05$, $d = 0.08$; nor was the interaction corresponding to the weakly greater effect in hindsight, t (339) = 0.88, $p > 0.05$, $d = 0.05$. Chance was evoked less for unexpected results in hindsight (15%) than in foresight (20%), a main effect (difference = -10; 95% CI: [-16, -4]), t (339) = 3.30, $p < 0.05$, $d = 0.18$. Other Causes were invoked more with unexpected results, in both foresight (20% vs. 10%) and hindsight (20% vs. 5%), with a significant main effect (difference = 10; 95% CI: [1.6, 18.4]) t (339) = 2.40, $p < 0.05$, $d = 0.13$, but no interaction.

We coded participants' open-ended explanations into the four categories of the structured attribution questions, adding a category for uninformative responses (e.g., "the child placed the dot," "I don't know why"). Table 4.10 shows typical examples. Other Cause explanations implied a valid method that revealed a different process than Rotate.

Table 4.10: Causal categories coded from open-ended responses.

Category	Subcategory	Examples
Rotate	Spatial Rotation	The child was unable to mentally rotate the image. The child has bad spatial rotation. The child is able to mentally rotate the image. The child has good spatial rotation.
	Perspective-Taking	The child placed the dot based on his point of view. The child responded to the relative or absolute position.
Invalid Method	Faulty Task	The instructions were ambiguous. The task was confusing.
	Faulty Child	The child was not paying attention. The child was too young to understand instructions. The child's brain is not developed.
Chance		The child placed the dot randomly. The child guessed. The child placed the dot based on luck.
Other Causes	Ambiguous	That's where the dot was on the other side. The child placed the dot in the same place.
	Task-Child Interaction	The child places the dot based on the shape of the Y. The child is left-handed. The child looks at this area first. The experimenter coached the child on the response.
	Memory	The child remembered where the dot was. The child forgot where the dot was.
Miscellaneous		The child placed the dot. A vacuous response. An uninterpretable response.

Table 4.11 shows the proportion of participants providing explanations in each category (e.g., 8 of the 84 (10%) who considered the expected observation in foresight attributed it to the child's mental rotation ability). For the few participants (11) who gave more than one explanation, we only included the first. As predicted, Invalid Method explanations were much more likely with unexpected results than with expected ones, in both foresight (28% vs. 1%) and hindsight (34% vs. 0%), a significant main effect (difference = 31%; 95% CI: [24%, 38%]) t (339) = 9.09, $p < 0.05$, $d = 0.49$.

4.4.3 Discussion

As in Experiment One and Slovic and Fischhoff (116), participants expected the initial result to replicate consistently in 10 additional observations more often in hindsight than in foresight — although that difference emerged here only with the expected observation (A). Conversely, the probability of never replicating the initial observation was less likely in hindsight than foresight,

Table 4.11: Proportion of participants making each attribution in open-ended response. Standard errors generated from 10000 simulations from a posterior Dirichlet distribution with Jeffreys' Prior (140; 141) Dirichlet(1/2,1/2,1/2,1/2,1/2,1/2) since there was a category with zero observations. $FU = 71$; $HU = 87$; $HE = 101$; $FE = 84$.

		Foresight	Hindsight
Causal Category	Condition	% (95% CI)	% (95% CI)
Rotate	Expected (A)	0.10 [0.04, 0.17]	0.13 [0.07, 0.20]
	Unexpected (B)	0.07 [0.03, 0.14]	0.13 [0.07, 0.20]
Invalid Method	Expected (A)	0.01 [0.00, 0.05]	0.00 [0.00, 0.02]
	Unexpected (B)	0.28 [0.19, 0.39]	0.34 [0.25, 0.44]
Chance	Expected (A)	0.04 [0.01, 0.09]	0.01 [0.00, 0.04]
	Unexpected (B)	0.06 [0.01, 0.11]	0.00 [0.00, 0.03]
Other	Expected (A)	0.81 [0.70, 0.87]	0.85 [0.76, 0.90]
	Unexpected (B)	0.46 [0.35, 0.58]	0.49 [0.38, 0.59]
Miscellaneous	Expected (A)	0.05 [0.02, 0.11]	0.01 [0.00, 0.05]
	Unexpected (B)	0.13 [0.06, 0.22]	0.03 [0.01, 0.09]

with a non-significant trend for a larger difference when it was the expected one.

Few participants recommended publishing the results, even when 10 children responded in the same way as the first, although somewhat more supported publication if the result was expected. These recommendations were similar in foresight and hindsight.

As in Experiment Two, participants invoked Invalid Method more with unexpected results than with expected ones, both when choosing among fixed options (Table 6) and when offering their own (Table 8). Moreover, these attributions were similar in hindsight and foresight, again suggesting that they are available if they are explicitly sought, as required by our attribution and data sharing tasks.

4.5 Experiment Four

Experiment Four replicates Experiment Three with several refinements designed to provide more sensitive measures. Based on the open-ended explanations in Experiment Three, Experiment Four divides the Invalid Method category in the structured attribution question into "something wrong with the child" and "something wrong with the task." Next, because participants in Experiment Three so uniformly wanted a much larger sample before publication, Experiment Four adds a task asking them to predict the outcomes for 100 additional trials, then indicate whether they would publish that result.

We also extend our study of error models in two ways. First, we examine whether people who attribute results to a flawed method also feel that all outcomes are equally likely, by asking participants to predict how many of 100 additional children will place their dot in each of the three areas. Finally, we ask how the researcher should respond, should that pattern actually be observed.

We predicted that an unexpected result (B) will encourage participants to believe that “anything can happen,” leading them to predict more uniform distribution across the three areas and to urge more cautious researcher responses. As before, unexpected results should be more strongly attributed to methodological problems.

4.5.1 Method

Design. Experiment Four was a 2 (foresight vs. hindsight) by 2 (Area: A, B) design.

Participants

For Experiment Four, participants were paid volunteers who responded to an Amazon MTurk ad offering them 1 dollar for participation in a 7-minute experiment. 312 of 465 individuals (67%) passed the attention filter. Their average age was 31 years old (range: 18 – 67); 135 were women (43%).

Materials

The instructions were the same as Experiment Three, with these modifications:

(a) Before learning (hindsight) or anticipating (foresight) the outcome of the initial observation, participants guessed the researcher’s hypothesis:

> What do you think the researcher’s hypothesis is (give your best guess)? [Hypothesis]

(b) After considering that initial result, participants answered a modified version of the structure attribution question from Experiment Three:

> What is the probability that? (Note: These five probabilities should total 100%.)
>
> 1. The child’s ability to mentally rotate the image caused the child to place the dot in Area A. [Rotate]

2. The child was not paying attention, and this caused the child to place the dot in Area A. [Faulty Child]
3. The task was confusing, and this caused the child to place the dot in Area A. [Faulty Task]
4. Random chance caused the child to place the dot in Area A. [Chance]
5. There was some other cause not otherwise mentioned. [Other]

(c) Participants then predicted the next 100 observations and assessed their implications:

> In a replication of this experiment with 100 additional children, how many children will place the dot in the following areas:
>
> 1. Area A
> 2. Area B
> 3. Area C
>
> If the replication of this experiment with 100 additional children comes out the way you expect, how should the researcher evaluate the hypothesis you guessed?
>
> 1. Have less confidence in the hypothesis
> 2. No change
> 3. Have more confidence in the hypothesis
>
> If the replication of this experiment with 100 additional children comes out the way you expect, which of the following actions would you recommend that the researcher take?
>
> 1. Collect more data before publishing [MoreData]
> 2. Publish without collecting more data [Publish]
> 3. Do not publish any of the data [NoPublish]

4.5.2 Results

Causal Attributions

Table 4.12 shows the probabilities assigned to the five causal explanations of the initial observation. Participants gave a higher probability to the child's mental rotation ability (as revealed by a valid method) when the dot was placed in area A rather than area B, in both foresight (40% vs. 10%) and hindsight (35% vs. 10%). The main effect (difference = -30, 95% CI: [-45, -15]) was statistically significant, t (308) = 4.11, p < 0.05, d = 0.23, with no interaction.

Conversely, participants assigned higher probabilities to the two Invalid Method explanations after an unexpected observation (area B) than after an expected one (area A). For Faulty Child, that was true in both foresight (25% vs. 10%) and hindsight (20% vs. 10%), with a significant main effect (difference = 15; 95% CI: [8, 22]), t (308) = 4.12, p < 0.05, d = 0.23, and no interaction. For Faulty Task, this was also true in both foresight (20% vs. 10%) and hindsight (25% vs. 10%), again with a significant main effect (difference = 10; 95% CI: [4, 16]) t (308) = 3.14, p < 0.05, d = 0.18, and no interaction. Attributions to Chance and Other Causes were unrelated to the reported outcome.

Table 4.12: Median probability (Md), standard errors (SE) for five causal attributions for Experiment Four. Sample sizes are (H = hindsight; U = unexpected): FU = 79; HU = 72; HC = 83; FC = 78.

Possible Cause	Condition	Foresight Md (SE)	Hindsight Md (SE)
Rotate	Expected (A)	40 (6.3)	35 (5.5)
	Unexpected (B)	10 (3.0)	10 (3.5)
Faulty Child	Expected (A)	10 (1.8)	10 (1.9)
	Unexpected (B)	25 (3.1)	20 (2.3)
Faulty Task	Expected (A)	10 (2.9)	10 (2.0)
	Unexpected (B)	20 (2.6)	25 (4.5)
Chance	Expected (A)	10 (2.8)	10 (1.3)
	Unexpected (B)	10 (2.7)	10 (0.9)
Other	Expected (A)	6 (2.5)	5 (2.1)
	Unexpected (B)	10 (0.8)	10 (1.0)

Thus, an expected result was more often attributed to a theory, whereas an unexpected result was more often attributed to methodological problems, with the child or the task. There were no significant hindsight-foresight interactions, indicating similar responses to actual results and potential ones.

Posterior Predictions

Table 4.13 shows participants' predictions for the 100 additional children. Both A and B were more likely when observed with the first child, with the main effect being significant for A (difference = -20; 95% CI: [-8, -32]), t (308) = 3.30, p = < 0.05, d = 0.19 and for B (difference = 10; 95% CI: [0.6, 19]), t (308) = 2.13, p < 0.05, d = 0.12. These differences were the same in foresight and hindsight.

Table 4.13: Median number of children expected to place the dot in each area (Md) and standard error (SE) for Experiment Four. Sample sizes are (H = hindsight; U = unexpected): $FU = 79$; $HU = 72$; $HE = 83$; $FE = 78$

		Foresight	Hindsight
Predicted Placement	Condition	Md (SE)	Md (SE)
Area A	Expected (A)	60 (4.7)	60 (3.6)
	Unexpected (B)	40 (3.2)	40 (3.6)
Area B	Expected (A)	10 (3.3)	10 (2.4)
	Unexpected (B)	20 (3.1)	25 (2.9)
Area C	Expected (A)	23 (2.2)	20 (2.4)
	Unexpected (B)	30 (1.7)	25 (2.8)

In order to assess participants' tendency to treat the three areas (A,B,C) as equiprobable, we calculate Shannon Entropy ($ShEn$), as a measure of the diffuseness or "flatness" of their distribution of predicted dot placements:

$$ShEn(A, B, C) =$$

$$\text{-P(A)}\times \log_2(P(A)) - P(B) \times \log_2(P(B)) - P(C) \times \log_2(P(C))(4.1)$$

Here, P(A) is the proportion of children (out of 100) predicted to place the dot in area A, and so on. With three response categories, the measure ranges from 0 (all 100 in one category) to 1.585 (a uniform distribution).

Overall, median Shannon Entropy was higher with an unexpected initial observation (B) than with an expected one (A), in both hindsight (1.44 vs. 1.10) and foresight (1.35 vs. 1.16), a significant main effect (difference = 0.20; 95% CI: [0.02, 0.38]), t (308) = 2.23, $p < 0.05$, $d = 0.13$, and no interaction. Thus, an unexpected initial result produced a stronger tendency to believe that "anything can happen" in the next 100 observations.

Belief Change

When they predicted the researcher's hypothesis, many participants explicitly indicated an area: A (84), B (12), or C (45). Among the others, 45 gave answers implying A or C (e.g., "the same position"; "the mirror position").

Most participants (not shown) thought that if their prediction for the 100 observations came true, then the researcher should be more confident in her original hypothesis, regardless of the first observation that they considered—even though considering that observation significantly affected their predictions for those 100 observations. This tendency was equally strong in hindsight and foresight.

Participants who predicted flatter distributions (as indicated by higher $ShEn$) were less likely to believe that the researcher should increase her confidence should those results be obtained (r = -0.13; 95% CI: [-0.24, -0.02]), t (310) = 2.35, $p < 0.05$, and more likely to believe that she should decrease it, (r = 0.12; 95% CI: [0.01, 0.23]), t (310) = 2.20, $p < 0.05$, should she observe such ambiguous results. Thus, participants with less confident predictions (flatter distributions) saw those data as less diagnostic, hence meriting less change in confidence.

Data Sharing Judgments

Most participants (not shown) recommended collecting more data before publishing even if the 100 observations turned out as they had predicted — although fewer did so than with the 10 additional observations in Experiment Three (82% (49/358) in Experiment Three versus 65% (92/312) in Experiment Four). These judgments were unrelated to the initial result (A or B) or whether it was reported as observed (hindsight) or possible (foresight).

Those who expected flatter distributions were less likely to recommend publishing without collecting more data (r = -0.19; 95% CI: [-0.30, -0.08]), t (310) = 3.47, $p < 0.05$, more likely to recommend collecting more data before publishing, (r = 0.16; 95% CI: [0.05, 0.26]), t (310) = 2.77, $p < 0.05$, and more likely to recommend not publishing any of the data (r = 0.11; 95% CI: [0.00, 0.22]), t (310) = 2.00, $p = 0.05$. Thus, participants were less inclined to recommend sharing the data with the scientific community when they had more diffuse expectations.

4.5.3 Discussion

When participants considered the child placing the dot in the expected area, they were more likely to attribute that result to a substantive theory, that the child could mentally rotate the image, and less likely to attribute it to methodological problems, such as that the child was not paying attention or found the task confusing. Participants in Experiments Two and Three attributed the unexpected result to chance or an ‘other cause’, such as placing the dot where the child looks first, more than an expected result. However, in this experiment, where the the response format captured their error models, participants attributed unexpected results to error, and not chance or other substantive theories.

The initial observation also made that result seem more likely in replications of the same experiment. Although that was true for both the expected and the unexpected initial result, observing the latter made future results seem less predictable, in the sense of being more uniformly spread out across the possible outcomes. That difference was not just a reflection of making the unexpected outcome seem more likely, thereby leveling the distributions. Rather, participants who considered B as the initial observation also saw C as significantly more likely, in both foresight (30 vs. 23) and hindsight (25 vs. 20) (diff in medians = 10 (95% CI [4,16]), t (310) = 3.57, $p < 0.05$, $d = 0.20$. Apparently, the unexpected result led to thinking about alternative causal models. As evidence that the entropy measure captured participants' uncertainty, it was correlated with how confident participants expected the researcher to be and how strongly they would recommend publishing the results.

We once again observed no foresight-hindsight differences with any of the present measures. As in Experiments Two and Three, the results suggest that the less certain perspectives of foresight are available in hindsight, if inferential processes are structured to evoke them. In a debiasing study, (116) provided such structure by requiring hindsight participants to give reasons why the unreported outcome might have occurred. Here, we asked them to reflect on the causes and disposition of the results. However, because these inferences were made after participants assessed the probability of replication, we have no direct evidence regarding their effectiveness as a debiasing procedure.

4.6 Experiment Five

In Experiment Four, foresight and hindsight participants were equally likely to invoke error models, both when they generated their own explanations and when they chose among explanations that we offered. The similarity of those attributions, with and without outcome knowledge, suggests that people could generate error models at any time. However, the intuition motivating our studies is that they typically do not do so until, in hindsight, unexpected outcomes motivate them to think even harder about what might have gone wrong.

If those additional reasons came from their own minds, rather than being generated in response to unexpected evidence, then they should have been incorporated in their prior knowledge. Considering these error models before data collection would thus have allowed orderly, even Bayesian, updating. However, having those considerations arise from unexpected observations means that such updating may be biased by the very results that prompted it. For example, hindsight bias should make reasons consistent with the results disproportionately accessible. Confirmation bias should give those reasons disproportionate credibility. If so, then researchers' inferences will be biased toward supporting their initial hypotheses by virtue of undermining the credibility of unexpected (and perhaps unwanted) results. Conversely, expected results will not prompt such a search for additional error model reasons.

Experiment Five creates conditions closer to actual foresight. In the *complete prior* condition, participants assess the potential relevance of three possible error models before they observe the data. These three explanations vary in how strongly they favor areas A, B, and C. In each of three incomplete prior conditions, one of these three explanations is omitted, so that it could be generated after observing the outcome of the experiment. We expect participants to overweight the credibility of explanations that are initially omitted, then discovered when needed.

4.6.1 Method

Experiment Five was a 2 by 4 design, crossing two possible outcomes (A, B) with four sets of explanations given to participants before they considered an outcome.

Participants

For Experiment Five, participants were paid volunteers who responded to an Amazon MTurk ad offering them 1 dollar for participation in a 7-minute experiment. Using the same attention filter left 969 of 1628 individuals (60%) who passed. Their average age was 30 years old (range: 18–70); 408 were women (42%).

Materials

The instructions were the same as in Experiment Four, except that before considering the initial result, participants answered a modified version of the structured attribution question from Experiment Four. In the complete prior condition, the question was:

> Which of the following do you think could possibly affect the experimental results (check all that apply)?
>
> 1. The task is confusing. [Uniform]
> 2. The children selected for the study are left-handed. [Non-Uniform area A]
> 3. Children like putting things in the middle, to maintain symmetry. [Non-Uniform area B]
> 4. Some other cause. [Other]

In the three other incomplete prior conditions, one of the three alternative explanations (Non-Uniform area A; Non-Uniform area B; Uniform) was omitted. They were meant to be available to explain outcome A, outcome B, or all three outcomes, respectively.

Participants were then told the first child placed the dot in either area A or area B and were asked to attribute the cause:

What is the probability that? (Note: These five probabilities should total 100%.)

1. The child's ability to mentally rotate the image caused the child to place the dot in Area A. [Rotate]
2. The task was confusing, and this caused the child to place the dot in Area A. [Uniform]
3. The child was left-handed, and this caused the child to place the dot in Area A. [Non-Uniform area A]
4. The child likes putting things in the middle to maintain symmetry, and this caused the child to place the dot in Area A. [Non-Uniform area B]
5. Random chance caused the child to place the dot in Area A. [Chance]
6. There was some other cause not already mentioned. [Other]

Participants then predicted the next 100 observations and made data sharing judgments, as in Experiment Four.

4.6.2 Results

Causal Attributions

Before considering any observations, 51% of participants thought that the task being confusing could affect the results, 37% that children being left-handed could do so, and 41% that children's preference for symmetry could. Thus all three of these explanations had plausible effects.

The probability assigned to the child having the mental ability to rotate the image (Rotate) was unaffected by which of the possible causes were mentioned before the result was observed. That probability was significantly higher when the child placed the dot in area A (25; 95% CI [22, 28]) rather than B (10; 95%CI [8, 12]), t (967) = 9.45, $p < 0.05$, $d = 0.30$, consistent with that ability explaining the former result, but not the latter.

Participants assigned the same probability to the child's confusion (Uniform) affecting the results regardless of whether that explanation was mentioned before the initial observation was reported. That probability was higher with the unexpected observation (B) than with the expected one (A), (20; 95% CI [18, 22]) vs. (10; 95% CI [8, 12]) , t (967) = 7.36, $p < 0.05$, d = 0.24.

The probability assigned to the child being left-handed (Non-Uniform area A) was affected by which of the possible causes were mentioned before the result was observed. Using means rather than medians (there was little re-sampling variance of the median in the bootstrap), participants assigned higher probabilities to the child's left-handedness affecting the initial observation when that explanation was mentioned before the observation was reported. This occurred both when they were told the child placed the dot in area A (13; 95% CI [10.7, 15.8]) vs. (9; 95% CI [7.1, 10.8]), t (277) = 2.65, p = 0.009, d = 0.16, and when they were told the child placed the dot in area B (7.9; 95% CI [5.7, 10.1]) vs. (5.6; 95% CI [4.3, 7.0]) t (207) = 1.78, p = 0.077, d = 0.12. Thus, an explanation of an expected result was judged less plausible when it was not mentioned before observing the results, regardless of whether the results confirmed or disconfirmed those expectations. There was also a significant main effect of area, such that the median probability assigned to the left-handed explanation was higher for participants told that the child placed the dot in area A (10; 95% CI [6, 10]) rather than area B (5; 95% CI [5, 5]), t (967) = 5.69, $p < 0.05$, d = 0.18.

Next, participants told that the child placed the dot in area B assigned higher mean probabilities to the symmetry explanation (Non-uniform area B) when it was mentioned before the result was observed (24.4; 95% CI [20.2, 28.7]) compared to when it was not (17.9; 95% CI [14.3, 21.5]), t (201) = 2.34, p = 0.02, d = 0.16. This relationship did not hold for participants told that the child placed the dot in area A (8.3; 95% CI [6.1, 10.5]) vs. (7.2; 95% CI [4.8, 9.5]), t (280) = 0.74, p = 0.46, d = 0.04. There was also a significant main effect, such that the median probability assigned to symmetry as a cause was much lower for participants told that the child placed the dot in area A (1.5; 95% CI [0, 5]) compared to participants told area B (20; 95% CI [15, 20]), t (967) = 9.79, $p < 0.05$, d = 0.31. Thus, as with the expected result, explanations for the unexpected result seemed less likely when not mentioned before the result was observed. Unlike the explanation for the expected result, this only happened when the result was consistent with the explanation. No differences emerged for chance and other causes for outcome or prior condition.

Thus, both non-uniform explanations were both judged less likely in hindsight when they were not mentioned in foresight. Attributions to the uniform (anything goes) explanation were unaffected by mentioning it in foresight.

Posterior Predictions

Participants predicted more of the next 100 observations in area A when told that the initial observation was there (med=60; 95% CI [60, 65]) than when told area B (35; 95% CI [33, 40]), t (967) = 10.8, $p < 0.05$, d = 0.35. The same was true when area B was reported (med=30; 95% CI [30, 33]), compared to when it was not (10; 95% CI [10, 15]), t (967) = 9.91, $p < 0.05$, d = 0.32.

As in Experiment Four, area C was predicted more often by participants told that the initial observation was B (med=25; 95% CI [25, 27]) rather than A (20; 95% CI [20, 20]) t (967) = 6.55, $p < 0.05$, $d = 0.21$, even though area C was not mentioned.

As measured by $ShEn$, the distributions of these predictions were flatter after the unexpected initial observation than after the expected one (A:1.18; 95% CI [1.16, 1.22]) vs. (B: 1.44; 95% CI [1.37, 1.49]), t (967) = 7.48, $p <$ 0.05, $d = 0.24$. There were no main effects or interactions of mentioning explanations.

Data Sharing Judgments

Most participants again recommended collecting more data before publishing, even if the 100 observations turned out as they had predicted. These patterns were unrelated to the initial observation and to which error models were mentioned beforehand. As in Experiment Four, those who expected flatter distributions (with higher $ShEn$) were less likely to recommend publishing (r = -0.12; 95% CI: [-0.19, -0.06]), t (903) = 3.7, $p < 0.05$, but not more likely to recommend not publishing any of the data (r = -0.003; 95% CI: [-0.08, 0.07]), t (733) = 0.09, $p > 0.05$. There were no other main effects or interactions between outcome reported, explanation mentioned, and publish judgments. Thus, participants were more likely to recommend collecting more data when they had diffuse expectations for the outcome of exact replications of the same experiments.

4.6.3 Discussion

In Experiment Five, the two non-uniform explanations (the child was left-handed, the child prefers symmetry) were both assigned higher probabilities of causing the result when mentioned before participants learned the outcome of the initial observations. That effect was greatest when the reported observation was consistent with the explanation, suggesting that causal models can be overlooked unless prompted — by asking or observation. Attributions to a diffuse explanation, producing uniform expectations (the child was confused), did not increase when it was mentioned sooner, rather than later, suggesting that such error models are always available. Publication judgments and predictions were unrelated to which explanations were mentioned (and omitted).

As in Experiment Four, participants found area C more likely when the initial observation was unexpected (B). This again suggests that the surprise made that seemingly unrelated outcome more plausible, consistent with "surprise" participants making more diffuse predictions overall, and being less likely to recommend publishing, even when the data confirmed their predictions.

4.7 General Discussion

We present five experiments examining how the evaluation of scientific evidence differs when the results are expected or unexpected and when considered in foresight or hindsight. Experiment One repeats Experiment One of Slovic and Fischhoff (116), thirty-five years later with an online (MTurk) sample, and finds similar results: an initial observation seems more likely to be replicated when considered in hindsight compared to foresight. Subsequent experiments examined responses to the most expected and unexpected results, among the four studies evaluated in Experiment One.

In Experiment Two, contrary to our expectations, participants were equally likely to attribute expected and unexpected results to methodological error, in both foresight and hindsight. Experiments Three and Four addressed a possible methodological problem with Experiment Two. Appropriate to our topic, it was a measurement error in how we elicited attributions to such problems. Experiment Three elicited explanations with an open-ended format, allowing participants to use their own concepts and terms. Experiments Four and Five offered fixed alternatives based on these responses. In all three experiments, unexpected results were more often attributed to flawed methods (or "error models"), but not to chance or other causes, compared to expected ones. These effects were similar in foresight and hindsight, suggesting that explicitly asking about alternative explanations equates these perspectives. Experiment Five affirmed this observation by systematically varying which explanations were mentioned before any results were reported. Mentioning explanations consistent with non-uniform predictions increased attributions to them, especially when consistent with the outcome. Invocation of the uniform error model explanation, making no specific predictions, was unaffected by whether it was mentioned before or after the initial observation.

Experiments Four and Five also found that reporting an unexpected outcome led to flatter predictions for 10 or 100 additional observations, also consistent with surprises evoking error models. In these predictions, observing the unexpected outcome (B) also increased the probability of the unrelated outcome (C), as though anything was now possible. We found that participants who gave flatter predictions were also less willing to recommend publishing the results rather than collecting more data.

Many previous studies have found that unexpected results are more likely to be attributed to error (142; 111; 13; 143; 144; 145). However, in these studies, the expected outcomes were typically also desired ones, for example, that capital punishment is an effective (or ineffective) crime deterrent (111).

Thus, participants attributed outcomes that were unwanted as well as unexpected to error. One exception is the finding that people often invoke error when confronted with disconfirming feedback in the Wason rule discovery task (146; 109), although even there, participants may become invested in a favored hypothesis. Here, participants considered studies run by others on a neutral topic, hence had expectations without desires. Masnick *et al.* (147) found a similar result using short vignettes about the efficacy of pedagogical techniques. In that study, participants who had natural expectations about the efficacy of the techniques, but no investment in their success, attributed unexpected results to methodological flaw.

The experiments can provide guidance to practicing researchers. Researchers wring their hands worrying about 'overfitting' unexpected data that weren't considered in foresight (130). There is a real risk of becoming committed to a weak and unwarranted theory that just happens to fit the data well. In Experiments Two through Four, explanations of data were seen as equally probable in foresight and hindsight. This suggests that one needn't worry so much about whether these explanations will seem disproportionately likely. Instead, our results indicate that, as long as the set of explanations remains fixed, judgments of them will not be affected.

However, completely failing to consider an explanation entirely is another matter. Experiment Five found that explanations that predicted specific results (e.g., area A or area B), as opposed to uniform explanations, are judged more probable when considered before observing the results, especially when they are consistent with the results. Thus, researchers may be overly skeptical of theories that were generated after the data are observed, or conversely, not skeptical enough of explanations set forth ahead of time. More "natural" explanations, that come to mind easily when designing an experiment, are also seen as relatively more likely after observing the results compared to explanations that may have required more thought (and even empirical observation) to generate. Deeper probing in foresight may help, by making sure all explanations that are serious possibilities are considered before observing the results. Unfortunately, there is no limit to this time consuming and often frustrating process, so the termination of this process ultimately depends on a judgment that the so-far-considered explanations are 'good enough'.

For both Experiments Four and Five, diffuse data were seen as due to a flawed experiment. As a consequence, researchers should reflect on the temptation to lock diffuse data away into their file-drawer (4). Researchers should be careful to examine data that may initially seem diffuse and uninformative, as it may be possible to discover systematic sources of error in the noise. As long as there are clues to the noise in the data, they can be very helpful for planning new experiments. They should lay the foundation to make future experiments sensible, as Experiments Two and Three did in our case, where an initial failure to find that surprises were attributed to error in Experiment Two laid the foundation for discovering the categories of error that needed to be included in Experiment Three.

One limit to the present research is its reliance on structured attribution options. The open-ended responses in Experiment Three revealed some of the diversity in how people intuitively formulate error model explanations. The structured options based on these responses revealed patterns missed by the ones that we produced intuitively in Experiment Two. Nonetheless, there is more to be learned by eliciting participants normal ways of thinking. As second limit is reliance on a single experimental stimulus, the Y-test study. As revealed in the manipulation check of Experiment Three, some participants interpreted it differently than we had expected. Any confusion on their part might have limited their ability to generate alternative explanations (and our ability to understand what they produced).

The results point to several directions for future research. In our experiments we do not look at the number and type of mentioned and omitted explanations. We often consider only one explanation before observing our results. When an explanation is the only one considered beforehand, alternative explanations generated after observing the data may gain little credibility, compared to if we had considered multiple explanations beforehand. Experiment Five did not test this, because in all conditions participants considered at least three explanations beforehand. Similarly, all of our experiments include all explanations at the time of measurement, possibly affecting the probability assigned to them compared to if they were left in an 'all else' category. Explanations neither mentioned beforehand nor included at the time of measurement may be disproportionately ignored, similar to omission of possible failure modes in fault trees (148).

In a dynamic context, one could look at how participants create error models and data veto judgments in a flexibly defined “warm-up” period, similar to that attributed to Millikan. For example, examining how people make decisions about whether to continue pursuing research goals in the face of apparent anomalies. This pits Millikan's warm-up period, where one throws away anomalous data to maintain the research goal, against Polanyi's (and Mayo's (77)) wild goose chase, where one pursues anomalies as they arise before continuing the research project. These are two very different and important research strategies. Their relative merits may be more easily decidable on psychological rather than normative (philosophical) grounds.

Other interesting future directions involve possible field experiments of the generation of causal explanations and data veto policies for unexpected results before or after scientific experiments are conducted. Disciplines amenable to this would include physics, such as work done at the Large Hadron Collider or

LIGO, pharmaceutical drug discovery, and psychological research. These contexts are likely to elicit much richer causal reasoning. The meaning of publishing the data in this context is well-understood by those involved, so a manipulation such as specifying ex ante versus an ex post data veto policy, simulating what was done by LIGO, would also be very informative.

4.8 Conclusions

Kuhn (71) asked, “How do scientists proceed when aware only that something has gone fundamentally wrong at a level with which their training has not equipped them to deal?” (p. 86) He answered, in effect, that they naturally attribute unexpected results to flawed experimental method and expected ones to the theory that guides them. It takes an accumulation of unexpected results, along with a deep insight, to prompt a scientific revolution. Here, we found similar treatment of expected and unexpected results with lay participants evaluating a single study. The practical implications of these results are seen in the data sharing policies that participants revealed. Although participants were generally cautious about publishing any results, they had much more confidence (less diffuse predictions) in ones that confirmed their expectations. Thus, by implication, unless scientists follow pre-specified data veto rules, they risk disproportionately discarding unexpected results.

The task used here was taken from a study of hindsight bias, wherein people struggle to retrieve the uncertainty of foresight, increasing their confidence that an observed result with be replicated. Producing reasons why another result was possible reduces that bias, by enriching hindsight. Conversely, considering a fuller set of possible causal principles prior to observing any results can reduce the tendency to invoke error models to explain unexpected results, by enriching foresight.

Chapter 5

Incentives, Error, and Data Sharing

Hypothesis testing has been found to follow a *positive test strategy*. Researchers collect data that they expect to conform to their prior beliefs and then exaggerate its information value (149), while discounting any inconsistent evidence that comes their way (150; 110; 107; 111). This result has been found with both simple experimental tasks and in dynamic artificial environments, such as simulated molecular biology (151), programmed robots (152), and multiple-cue probability learning (153). Similar patterns have been found in scientific laboratories. For example, in an observational study of a biological sciences laboratory, Dunbar (110) found that scientists did not immediately reject their hypotheses after they were contradicted by data. Rather, their first reaction was to invoke experimental error (150; 110; 107). In thirty-seven experimental treatments conducted by one biologist, twenty-one had unexpected results, most of which were treated as errors (110).

After inspecting their data for errors, researchers must decide whether to communicate any data that they consider flawed. If the researchers' error attributions are accurate, then omitting these errors from published reports may avoid distracting readers. If they are inaccurate, then failing to publish those data will allow false theories to emerge and persist. Justified or not, data attributed to error are unlikely to be shared. For example, statistical significance is often (incorrectly) interpreted as the probability of error in data, and is usually a necessary condition for publication (11; 12).

Data sharing decisions are not only affected by whether the data are perceived to be faulty, but also by professional rewards for publishing positive (usually statistically significant) results. These rewards can produce a healthy motivation to make a discovery, such as finding a successful anti-cancer drug. However, rewards may also undermine accurate data inspection by increasing scrutiny of results that indicate the discovery is false, while simultaneously making affirming results a wanted relief from the pressure to produce (104). Supporting this account, there is evidence that higher rewards for publishing are associated with publication bias (154; 56).

Here we present three experiments on decisions to share possibly faulty data. We use Wason's 2-4-6 rule-discovery task (155). It asks participants to try to discover the rule that generated a set of numbers (2, 4, 6), by proposing a new set of three numbers (a proposed triple), then getting feedback as to whether the numbers that they proposed fit the rule. We use Penner and Klahr's version (109), in which participants are told that some percentage of the time, the feedback will be false—a feature that adds something like the uncertainty that is inevitable with scientific inferences.

We add several new features to the task. (a) Before receiving feedback, participants assess the probability that it will affirm their expectations. (b) After receiving it, they indicate whether they would share each trial, including the feedback, with a second person trying to discover the same rule. In Experiment One, the sharing decision is done at the end of the task; in Experiments Two and Three it is done immediately after they make their error judgments. (c) We also use two types of incentives intended to simulate the rewards that may lead to motivated reasoning. Experiment Two provides participants with a large incentive ($100) for correctly guessing the rule, and a small incentive ($1) for concluding that they do not know the answer. Experiment Three provides participants with an incentive to convince a matched participant that they discovered the rule, whether or not they actually did.

Using this task, we first replicate the finding that error is more likely to be invoked with disconfirming than with affirming feedback (109; 146; 108). We then examine whether these error attributions are justified using two evaluative criteria: (a) *accuracy*, defined as whether the judgments are correct; and (b) *Bayesian consistency*, defined as attributing feedback to error if and only if either: (i) participants strongly expected the triple they proposed to fit the rule but it did not, or (ii) participants strongly expected the triple they proposed to not fit the rule but it did.[1] *Selective reporting* is the degree to which trials attributed to error are not shared with the matched participant, compared to those attributed to other sources.

[1] More precisely, using Bayes' Rule it can be shown that feedback should be attributed to error whenever one believes that there is greater than an 80% prior probability that the triple fit the rule, but the feedback indicates it does not fit, or conversely one believes there is less than a 20% prior probability that the triple fit the rule, but the feedback indicates that it does fit.

5.1 Experiment One

Experiment One looks at whether error attributions are consistent with prior beliefs, whether error attributions correspond to actual error, and whether trials are less likely to be shared with another person when the feedback is attributed to error. We used the Wason 2-4-6 rule discovery task with feedback error (109). On each trial there was a 20% chance that the feedback was in error.

5.1.1 Method

Participants

Eighteen Carnegie Mellon University undergraduates completed the task for course credit.[2] They were on average 21 years old (range: 18 – 38). There were 7 women. One participant gave no valid responses.

Procedure

Participants were seated at a computer, asked to sign an informed consent document, and then instructed that they had 30 minutes to complete the task. Participants completed the task online as a Qualtrics questionnaire with embedded Javascript used for feedback. Each page (trial) of the questionnaire had the same format. In order, participants proposed a rule, proposed a new triple, assessed the probability that the triple they proposed fit the Actual Rule, received feedback, judged whether the feedback reflected error, and then decided if they wanted to give their Final Answer. They were reminded to record all responses both on the computer and on the spreadsheet they were given. After participants decided to stop new trials and give their Final Answer, or thirty minutes had passed, they were asked to review their spreadsheet and mark the trials that they thought should be shared, in order to help a new participant solve the problem.

Materials

The materials were a modification of Penner and Klahr's (109) version of the Wason 2-4-6 rule discovery task, with the study introduction rewritten to increase readability and comprehension, based on pretesting using cognitive

[2] A random-effects meta-analysis of the effect of disconfirming feedback on error attributions from Penner and Klahr's Study One (numerical broad and narrow conditions), Penner and Klahr's Study Two (numerical/narrow) (109) and Gorman (146), indicated an overall effect size of Hedges' $G = 0.43$, 95% CI [0.33, 0.53]. The sample sizes of the control groups in these studies were 15, 15, 25, and 24, respectively, indicating eighteen participants should provide sufficient power to detect the effect of feedback on error attribution.

interviews. The study also used computerized, rather than hand-written, feedback.

Introduction

Participants were shown the following introduction on the computer along with a separate paper copy as a reminder:

> "You will be given three numbers that are related somehow. For example: 3, 5, and 15. This is called a triple. There are many possible rules that could relate these three numbers. We have selected only one of them. The rule that we selected is called the Actual Rule. You will not be given the Actual Rule. Your task is to discover it. The initial triple on the next page is an example drawn from the Actual Rule."
>
> "Our study is using several versions of this task. Yours is a particularly difficult one. Sometimes, even if your Proposed Triple FITs the Actual Rule, the computer may output that it DOES NOT FIT. Conversely, sometimes, when your Proposed Triple DOES NOT FIT the rule, the computer may output that it FITs. On any trial there is a 20% chance that you will get false feedback. For each trial if you think false feedback occurred mark "F" in the "Feedback" column on your spreadsheet. If you think true feedback occurred, mark "T" in the "Feedback" column."
>
> "At any time you may try to guess the Actual Rule that we selected. This is called the Final Answer. You only get one Final Answer and it may be wrong. Once you make your Final Answer you can no longer get feedback from the computer and the experiment will end."

Initial Triple

At the top of each page, the initial triple (2,4,6) was shown. Participants were told:

> "The initial triple above is an example drawn from the Actual Rule."

Proposed Triple

After writing their best explanation of the initial triple, they were instructed to propose a new triple:

> “You may propose additional triples to help you discover the Actual Rule. The computer will tell you whether the triple you proposed fits the Actual Rule. Record all information on the spreadsheet you were given. Write one number of your triple in each box below.”

Prior Probability

On each trial, before they received feedback, participants assigned a probability that the triple they proposed fit the actual rule, by answering the following question:

> “What is the probability that the triple you proposed fits the Actual Rule? (must be a number between 0 and 100)”

We denote this $P(TFTR)$ for ‘[P]robability that the [T]riple [F]its [T]he actual [R]ule’.

Error judgment

Immediately after receiving feedback that the triple fit (FIT) or did not fit (DNF) the rule, participants judged whether they thought the feedback was due to error:

> “Do you think this feedback was true or false? (True/False)”

Final Answer

After participants felt they had completed enough trials, or the 30–minute window expired, they were asked to make their Final Answer:

> “Write your Final Answer for the Actual Rule in the box below (it can be mathematical or in words).”

Data sharing

Participants then decided which trials they wanted to share with a new participant:

> “In this experiment, a trial is a page where you proposed a rule, a triple, a probability estimate, received feedback, and judged whether you thought the feedback was false or true.”
>
> “In a future experiment we will have a new participant try to discover the same rule you tried to discover.

> You can choose trials that you think will help him or her solve the rule. For each trial you indicate, all of the information would be shared, including:
>
> 1. your rule
> 2. the proposed triple
> 3. your probability estimate
> 4. the feedback
> 5. whether you thought the feedback was false or true
>
> In the space below, please indicate the trials you conducted that you think would help this person."

5.1.2 Results

Unless otherwise noted, all estimation was done using hierarchical logistic models with subject-level varying intercepts (156; 157). The model assumes that multiple observations from the same person are conditionally independent given the subject-specific intercept. Tests, standard errors, and p-values based on these models were calculated using non-parametric bootstrap with 200 simulations per statistic (138).

Performance

Participants completed a median of eight trials. Each participant's task performance score was determined by their final answer, scored on a 5–point scale awarding one point for each element of the rule that they had discovered. The five elements were: 1) even numbers, 2) consecutive numbers, 3) ascending numbers, 4) the lower bound is 2, and 5) the upper bound is 100. All six participants who scored zero used a mathematical formula that was either unspecific (e.g., $x + 2$) or not a rule (e.g., $(2 + 6)/2 = 4$). Among the seven participants with a score of 1, six included even numbers in their answer, and one mentioned ascending numbers. Of the four participants who scored 2 on the task, two mentioned consecutive even numbers, and two mentioned ascending evens. The one participant who scored a 3 on the task guessed sequential even numbers less than 100.

Error Attributions

Replicating Penner and Klahr (109), participants judged disconfirming feedback to be error more often ($38\%; SE = 5.9\%$) than affirming feedback ($8.6\%; SE = 4.4\%$), $t(162) = 3.77$ $p < 0.05$, $d = 0.30$. In multiple regression, there was only a main effect of feedback type on attributions of error ($t(159) = 3.1$, $p = 0.0075$), with no significant main effect of actual error ($t(159) = 1$, $p = 0.46$) or interaction between the two factors ($t(159) = 0.72$, $p = 0.62$).

Bayesian Consistency

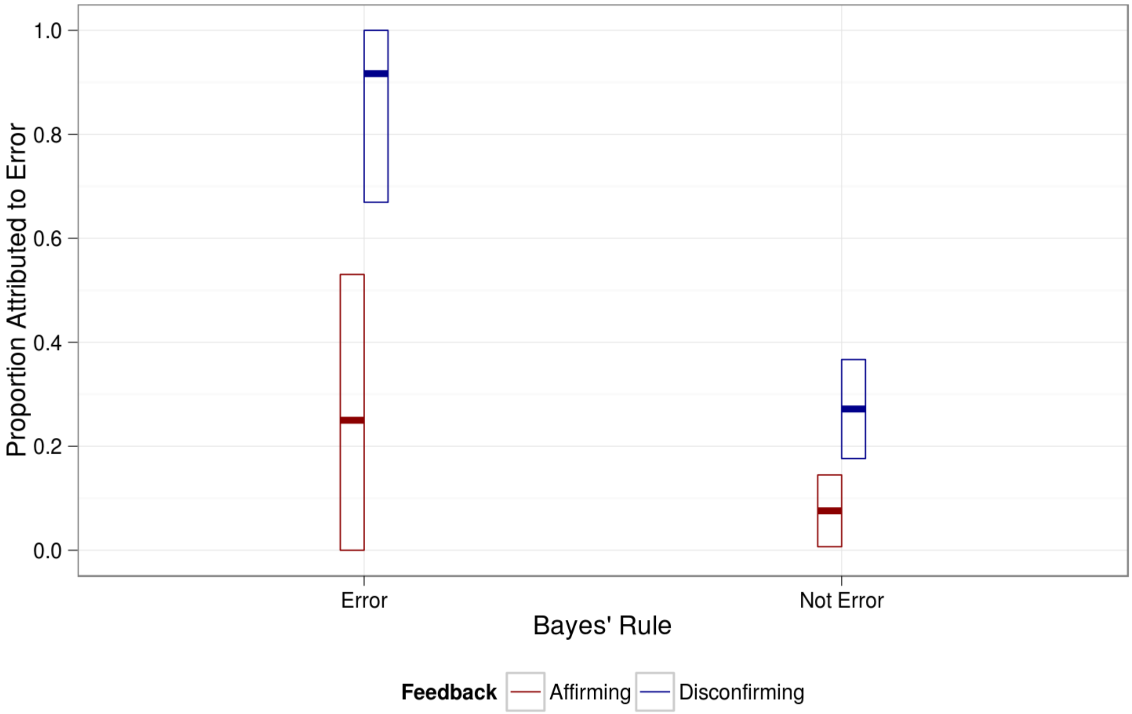


Figure 5.1: Proportion of trials attributed to error depending on whether Bayes' Rule predicted error attribution and whether the feedback was affirming or disconfirming.

These error attributions were consistent with prior beliefs. When participants received disconfirming feedback, they correctly attributed 11 of 12 trials to error when they strongly expected the triple to fit the rule beforehand ($P(TFTR) > 0.8$), and incorrectly attributed 15 of 65 trials to error when the strength of their prior beliefs did not justify attributing the feedback to error, ($P(TFTR < 0.8)$), $\chi^2(1) = 23$, $p < 0.05$, $\phi = 0.5$. When receiving affirming feedback, they correctly attributed 1 of 4 trials to error when they strongly expected the triple to not fit the rule beforehand ($P(TFTR) < 0.2$), and 4 of 63 incorrectly when the strength of their prior beliefs did not justify attributing the feedback to error ($P(TFTR > 0.2)$), $\chi^2(1) = 1$, $p = 0.31$, $\phi = 0.12$. However, in multiple regression, there was a main effect of feedback type on attributions of error ($t(159) = 3$, $p = 0.0095$), no significant main effect of Bayes' Rule requiring error attribution ($t(159) = 1.2$, $p = 0.37$), and no interaction between the two factors ($t(159) = 1.4$, $p = 0.3$). The overall correlation between their judgments and the consistency criterion was $\phi = 0.38$, $\chi^2(1) = 23$, $p < 0.05$.

Accuracy

Although error attributions were consistent with prior beliefs, they did not match actual error. When participants believed that feedback was false, it was as likely to be accurate as inaccurate (23% vs. 29%), $\chi^2(1) = 2.6$, $p = 0.35$, $\phi = 0.11$.

Data Sharing

Participants were as likely to share data when feedback affirmed their hypothesis as when it did not, (40%; $SE = 11\%$ vs. 34%; $SE = 12\%$), $t(122) = 0.48$, $p > 0.05$. They were also equally likely to share feedback when they saw it accurate or inaccurate (40%, $SE = 10\%$ vs. 32%, $SE = 10\%$), $t(122) = 0.53$, $p > 0.05$. When including both main effects and the interaction between actual error and attribution of error to predict whether each trial would be shared, there was neither a significant main effect of error attribution ($t(119) = 0.55$, $p = 0.68$), actual error ($t(119) = 0.045$, $p = 0.8$), or an interaction between the two factors ($t(119) = 0.19$, $p = 0.78$).

5.1.3 Discussion

The results replicate the findings of Gorman (108) and Penner and Klahr (109), who found that people are more likely to question feedback when it disconfirms their hypothesis. For error attributions, most of these judgments were normatively justified, matching the consistency criterion on 92 of 144 trials (64%). In spite of this consistency, participants were unable to identify actual error. Finally, on a task new to this study, participants shared information at equal rates regardless of whether the feedback was affirming or disconfirming and regardless of whether it was attributed to error.

The positive test strategy (149) entails seeking affirming evidence and discounting disconfirming evidence. Experiment One found that this strategy is both internally consistent and inaccurate. Participants were, however, no less likely to share disconfirming or seemingly flawed data. Although this pattern of data sharing contradicts the positive test strategy, we observed that some participants had difficulty interpreting the open-ended data-sharing question. Namely, when asked which trials they wanted to share, some responded with a triple (e.g., "2, 4, 6"), rather than a trial (e.g., "trial 3").

Experiment Two addresses this problem by using a fixed response format after each trial rather than an open-ended one at the end.

5.2 Experiment Two

Experiment One replicated the positive test strategy found in previous studies, with participants invoking error more often for disconfirming feedback. These error attributions were justified in terms of the consistency criterion, but not the accuracy criterion. The second experiment examines the effects of incentives on these judgments, using a monetary payoff that encourages participants to convince themselves that they know the rule. Specifically, participants were offered $100 for guessing the Actual Rule correctly, and $1 for concluding that they do not know it. This incentive scheme sought to encourage motivated reasoning, so that hopeful participants believe that they've reached a correct conclusion rather than assess their knowledge candidly.

Experiment Two also improves the data-sharing decision. Immediately after receiving feedback, participants make a binary (Yes–No) decision about whether each trial should be shared. Having data-sharing decisions at the end of each trial rather than at the end of the experiment sought to make it clearer that the sharing decision applies to the current trial, resolving any ambiguity about whether triples or trials should be shared. It also elicits sharing judgments earlier in the task, before participants might become tired or frustrated.

5.2.1 Method

Participants

Fifty-eight Carnegie Mellon University undergraduates participated in the experiment for course credit. There were thirty-four women, with average age of 20 years (range: 18 – 24).

Design

Participants were randomly assigned to either the control or the incentive condition. This was a one-way between-subjects design with two levels.

Procedure

The entire experiment lasted 30 minutes. Participants were given informed consent, instructions, and the response sheet. At the end of the experiment, they were asked to leave their email address, with the promise that they would be contacted later if they had solved the rule to receive their bonus payment. This delay of bonus payment was done to prevent participants from telling their friends the correct answer.

Materials

All materials were the same as those in the first study except for the following three changes. First, participants were given the spreadsheet, but were not required to use it.

Second, in the financial incentive condition, participants were told:

> "At the end of the experiment you will be given a chance to win money by guessing the rule. If you decide to guess the rule you will receive 100 dollars if the guess is exactly correct, but 0 dollars if the guess is incorrect. On the other hand, you can decide that you do not know and receive 1 dollar for sure."

Third, decisions to share a trial were made immediately after participants made their error attributions:

> "We are also interested in how people share information. In a future experiment, a new participant will try to discover the same Actual Rule that you are trying to discover. You can share information with this new participant to help him or her solve the Actual Rule. If the new participant solves the rule, you will receive an additional 50 dollars."

The trials were described in the same way as Experiment One, but the sharing judgment was now binary:

> "Do you think this trial should be shared with a new participant? (Yes/No)"

5.2.2 Results

Incentives and Performance

Incentives doubled the median number of trials from 4.5 to 9.[3] Using the same scoring method as Experiment One, those in the incentive condition scored

[3] Although the median number of trials increased, a non-parametric Kolmogorov-Smirnov (KS) test for differences in empirical cumulative distributions indicates no differences in distribution. Between Experiment One and the control condition of Experiment Two, the KS test was $D = 0.31$, $p = 0.23$. Between Experiment Two control and incentive conditions, the KS test was $D = 0.32$, $p = 0.11$. Thus, although the medians were different, the distributions of trials between the studies and conditions were similar.

about the same on average ($M = 1.58$, $SD = 1.21$) as those in the control condition ($M = 1.66$, $SD = 1.21$), t (56) = 0.80, $p > 0.05$. One participant solved the rule exactly, and was compensated with a $99 Amazon gift card.

Error Judgments

As in Experiment One, those in the control condition were significantly more likely to see feedback as error when it was disconfirming (29%, $SE = 5.8\%$), than when it was affirming (10%, $SE = 4.1\%$), $t(171) = 2.89$, $p < 0.05$, $d = 0.22$. In contrast, participants in the incentive condition were equally likely to attribute error to disconfirming feedback (16%, $SE = 3.7\%$) and to affirming feedback (20%, $SE = 6.7\%$), $t(309) = 0.62$, $p > 0.05$, $d = 0.04$. Thus, although we expected the incentives to increase motivated reasoning, they appeared to reduce the tendency for participants to attribute disconfirming results to error. In multiple regression, there was a significant main effect of feedback type ($t(476) = 2.5$, $p = 0.038$), incentive ($t(476) = 2.4$, $p = 0.05$), and a significant interaction between the two factors, where disconfirming feedback only increased error attributions for those in the control condition ($t(476) =$ -3.1, $p = 0.0069$). There were no other main effects, two-way, or three-way interactions between feedback type, incentive condition, and actual error.

Bayesian Consistency

As in Experiment One, error attributions for participants in the control condition were consistent with their prior beliefs. For affirming feedback, they correctly attributed 3 of 10 trials to error and incorrectly attributed 0 of 47 trials to error, $\chi^2(1) = 3.9$, $p = 0.057$, $\phi = 0.24$.[4] For disconfirming feedback they correctly attributed 9 of 13 trials to error and incorrectly attributed 16 of 82 trials to error, $\chi^2(1) = 10$, $p < 0.05$, $\phi = 0.32$. The overall correlation between their error attributions and the consistency criterion was $\phi = 0.24$, $\chi^2(1) = 10$, $p = 0.0015$.

Participants in the incentive condition exhibited similar consistency. For affirming feedback, they correctly attributed 12 of 38 trials to error and incorrectly attributed 15 of 76 trials to error, $\chi^2(1) = 2.9$, $p = 0.088$, $\phi = 0.16$. For disconfirming feedback they attributed 11 of 31 trials to error correctly and incorrectly attributed 22 of 155 trials to error, $\chi^2(1) = 16$, $p < 0.05$, $\phi = 0.29$. The overall correlation between their error attributions and the consistency criterion was $\phi = 0.19$, $\chi^2(1) = 12$, $p = 0.001$.

[4] A hierarchical model could not be used for the control condition. Only one participant both made an error attribution and should have not made an error attribution. Thus, only one subject-level intercept could be fit, as all other participants had zero probability of judging error. To deal with this we pool all of the data together to get an approximate answer.

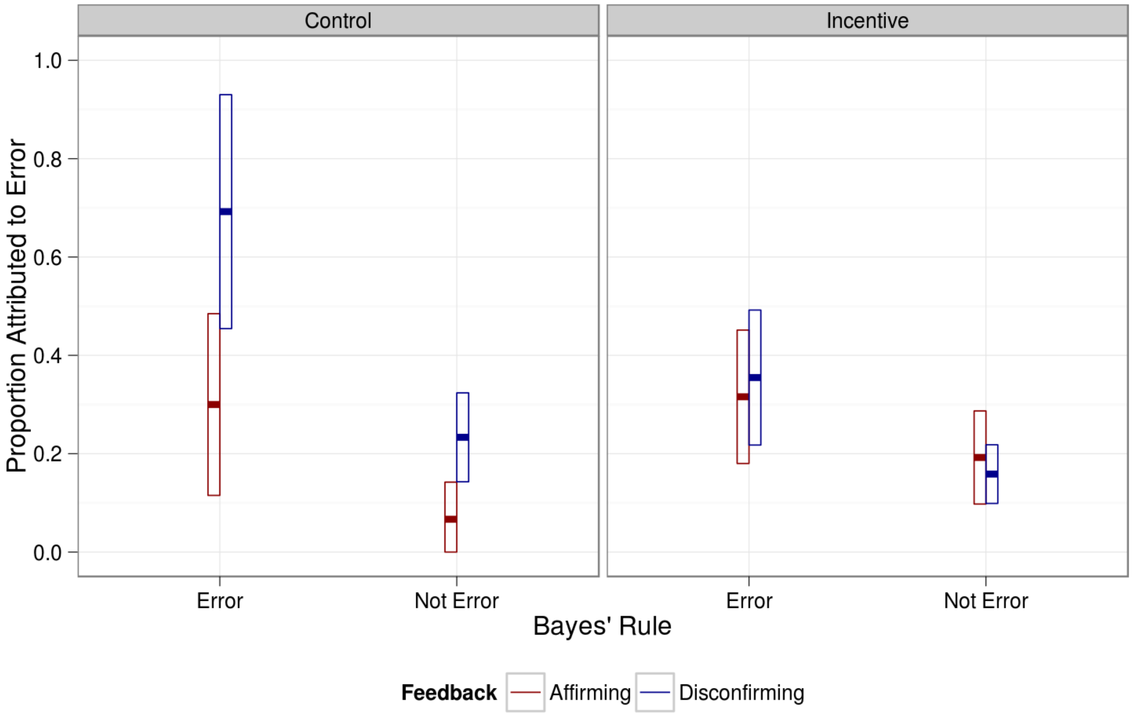


Figure 5.2: Proportion of trials attributed to error depending on whether Bayes' Rule predicted error attribution and whether the feedback was affirming or disconfirming.

Accuracy

As in Experiment One, participants in the control condition were unable to identify when actual errors occurred. They correctly identified 26% of actual errors and incorrectly identified 20% of non-errors as error, $\chi^2(1) = 2.2$, $p = 0.37$, $\phi = 0.093$. For the incentive condition, participants were also unable to identify actual error. They correctly identified 21% of actual errors and incorrectly identified 21% of non-errors as error, $\chi^2(1) = 1.3$, $p = 0.44$, $\phi = 0.056$.

Data Sharing

In contrast to Experiment One, participants in the control condition shared a smaller proportion of trials when the feedback was disconfirming (84%, $SE = 8.5\%$) than when it was affirming (93%, $SE = 5.5\%$), $t(171) = 1.96$, $p = 0.05$, $d = 0.15$. Similarly, they shared a smaller proportion of trials when they judged the feedback to be an error (79%, $SE = 12\%$) than when they judged it to be accurate (91%, $SE = 6.1\%$), $t(171) = 1.98$, $p < 0.05$, $d = 0.15$.

Participants in the incentive condition also shared a smaller proportion of trials when the feedback was disconfirming (84%, $SE = 6.2\%$), than when it was affirming (94%, $SE = 3.7\%$), $t(309) = 2.95$, $p < 0.05$, $d = 0.17$. They also shared a smaller proportion of trials when they judged the feedback to be an error (71%, $SE = 11\%$) than when they judged the feedback to be accurate (91%, $SE = 3.8\%$), $t(309) = 3.94$, $p < 0.05$, $d = 0.22$.

In multiple regression, there was only a significant main effect of error attribution ($t(476) = 2$, $p = 0.053$), and a marginally significant interaction between actual error and incentive condition, such that those in the incentive condition were more likely to share actual errors than those in the control condition ($t(476) = 1.7$, $p = 0.086$). There were no other main effects, two-way, or three-way interactions between feedback type, actual error, and incentive condition.

5.2.3 Discussion

Experiment Two again found that participants more often attribute error to disconfirming feedback when given no incentive beyond their intrinsic motivation to solve the problem. However, participants who were offered a large incentive for getting the rule attributed error to affirming and disconfirming feedback at equal rates. Although we had expected the incentive for getting the rule to increase motivated reasoning, it actually reduced the tendency for participants to attribute disconfirming feedback to error. It did not, however, lead to error attributions that were either more accurate or more consistent with prior expectations. Participants in the control condition met the consistency criterion on 125 of 152 trials (82%), which was a higher rate than those in the incentive condition (217 of 300 trials, 72%). One possible explanation is that the incentive helped participants maintain a more balanced perspective on the likelihood of error after receiving feedback; however, in spite of their motivation, they lacked the understanding (e.g., of Bayes' Rule) needed to respond consistently. An alternative explanation is that participants in the incentive condition rushed through the prior probability and error attribution questions in order to complete more trials, thereby creating more chances to propose triples and get feedback. This strategy would reduce consistency and make attributions of error more equal across feedback types, and is consistent with the finding that participants in the incentive condition completed twice as many trials in the same time period as those in the control condition.

For both the control and the incentive groups, participants shared disconfirming feedback less frequently than affirming feedback. They also shared feedback that they attributed to error less frequently than feedback that they saw as accurate. Those error attributions were loosely justified by internal consistency, but not by accuracy. Extrapolating to scientific contexts, researchers may have defensible reasons to omit data from publication based on their expectations, but that this consistency may not prevent harm to those who must use the data. Before reaching that conclusion, we address one possible artifact in Experiment Two's procedure: placing the sharing decision immediately after the error attribution task, perhaps suggesting that the two should be related. Experiment Three remedies this possible confound by eliciting data sharing decisions and error attributions both during each trial and at the end of the task, also allowing participants to reflect on all the data before making their final error attributions and data-sharing decisions.

Finally, Experiment Two's incentive scheme sought to motivate participants to believe they knew the rule. However, the value of data are usually determined not by the person who collects the data themselves, but by others, such as reviewers (for journals) or regulatory bodies (for drug approval). These people, who are external to the data collection process, determine the reward to the researcher based on their prior beliefs and their evaluation of the data shared with them. To simulate this incentive system more closely, Experiment Three uses the natural expectations that participants have about how to convince another person. We expect that an incentive to convince another person should increase the preference for discounting disconfirming feedback.

5.3 Experiment Three

Experiment Three replicates Experiment Two with several modifications. Most importantly, a new condition provides an incentive for participants to convince another person that their proposed Final Answer is correct, with data-sharing as the sole mode of communication between them. To do this, we embed the Wason task in a teacher–learner game, a type of principal–agent game (158; 60). In this task, the participant collecting the data (the teacher) shares data with another person (the learner) who has to guess the rule based on the data that the teacher decides to share.

The teacher is in one of two incentive conditions. The *compatible* incentive condition rewards both the teacher and learner if the learner guesses the rule. In the *perverse* incentive condition, the learner's rewards remain the same, but the teacher receives money if the learner accepts the teacher's Final Answer. Thus, the perverse incentive allows the teacher to distort the data supplied to the learner, potentially increasing her own payoff while reducing the learner's reward. In this scenario, the teacher knows the entire game structure, but the learner does not. Specifically, the learner is not told that the teacher does not have to share all the trials that were conducted, and the teacher is told that the learner only knows about the shared trials.

Experiment Three also deals with two methodological issues brought up in Experiment Two. One is that participants in the incentive condition attributed error to affirmation and disconfirmation equally, but were slightly less consistent in their error attributions than participants in the control condition. This may have reflected their rushing through the task to complete more trials. To reduce this threat, we use a penalty for making incorrect prior probability and error attributions. Any payoff to the participant is reduced in proportion to their inaccuracy on these two measures. This penalty prevents them from performing one element of the task well (collecting many trials) at the cost of the other elements (rushing through error attributions). The second was the possibility that participants assumed that the data sharing and error attribution judgments should be related because they occurred sequentially on each trial. This could create a false correlation between the two measures based on the participant's belief that the experimenter put the two questions close to each other for a reason. To deal with this, we also elicit data sharing decisions and error attributions at the end of the task, using a fixed-response format rather than the open-ended format used in Experiment One.

5.3.1 Method

Participants

One hundred Amazon Mturk volunteers completed the task for $5. There were 46 women, with average age of 32 years (range: 18–65).

Design

The design was a 2 level (perverse or compatible incentive) between-subjects design.

Materials

The procedure and materials were the same as in Experiment Two except for the following modifications. First, participants completed three 'practice trials' to help them understand the task. They were then told the following:

> "We are also interested in how people share information. The information comes in trials. A trial is a page where you proposed a triple and received feedback. The practice trials you conducted are shown below. For each trial you share, another person will get the triple you proposed and the feedback you received. The person will also receive the Final Answer you propose at the end of the task, regardless of the trials you share."

Participants were then told about possible bonus money:

> "Both you and the person you share trials with can earn up to a $5 bonus in addition to the $5 you receive for participating in the experiment."

The perverse incentive condition was followed with this text:

> "How you earn bonus money:
>
> - If the other person thinks your Final Answer matches the Actual Rule exactly, then you get $5.
> - If the other person thinks your Final Answer does not match the Actual Rule at all, then you get $0.
> - If the other person thinks your Final Answer somewhat matches the Actual Rule, then you get somewhere between $0 and $5."
>
> "How the person you are sharing trials with earns bonus money: The person you are sharing trials with can also earn money.
>
> - This person gets the most money ($5) by correctly judging how well your Final Answer matches the Actual Rule.
> - If this person thinks your Final Answer matches the Actual Rule, but it does not, the other person gets less money.
> - If this person thinks your Final Answer does not match the Actual Rule, but it is does, the other person gets less money."

Those in the compatible incentive condition were told:

> - "If the other person's guess matches the Actual Rule exactly, then you both get $5.
> - If the other person's guess does not match the Actual Rule at all, then you both get $0.
> - If the other person's guess somewhat matches the Actual Rule, then you both get somewhere between $0 and $5."

Finally, participants were told the penalty for making incorrect attributions:

> "Penalty for wrong answers
>
> Any bonus you get will be reduced if your false feedback and probability judgments are wrong. Thus, to earn the most money you should make your false feedback and probability judgments as accurate as possible."

5.3.2 Results

Incentives and Performance

As in Experiment Two, participants in the compatible and perverse incentive conditions completed a median of about 8 trials (9 and 7, respectively), $t(97)$ = -0.12, $p = 0.91$, d = -0.012.

Error Judgments

Both during (38% vs. 4.8%) and at the end of the task (41% vs. 12%), those in the compatible incentive condition were more likely to see feedback as in error when it was disconfirming than when it was affirming, ($t(510) = 7.7$, $p < 0.001$, $d = 0.34$; $t(525) = 6.5$, $p < 0.001$, $d = 0.28$, respectively). Similarly, both during (39% vs. 9.3%) and at the end of the task (47% vs. 7.9%), those in the perverse incentive condition were significantly more likely to see feedback as in error when it was disconfirming than when it was affirming ($t(537) = 7.3$, $p < 0.001$, $d = 0.32$; $t(504) = 8.5$, $p < 0.001$, $d = 0.38$, respectively).

Bayesian Consistency

For both incentive groups, adding the penalty for incorrect error attributions and probability judgments greatly improved accuracy and consistency, as compared to Experiments One and Two. For the compatible condition, the overall correlation between their error attributions and the consistency criterion was $\phi = 0.37$, $\chi^2(1) = 76$, $p < 0.001$. Participants in the perverse incentive condition exhibited even greater consistency, $\phi = 0.55$, $\chi^2(1) = 179$, $p < 0.001$.[5]

Accuracy

Participants both in the compatible and perverse incentive conditions were able to accurately identify error during the task ($\chi^2(1) = 79$, $p < 0.001$, $\phi = 0.37$; $\chi^2(1) = 139$, $p < 0.001$, $\phi = 0.5$, respectively). Participants in the compatible incentive group correctly identified 44 of 99 actual errors and incorrectly identified 60 of 458 non-errors as error. For the perverse incentive

[5] For affirming feedback in the compatible condition, they correctly attributed 3 of 5 trials to error, and incorrectly attributed 8 of 258 trials to error, $\chi^2(1) = 3.9$, $p < 0.057$, $\phi = 0.24$. For disconfirming feedback they correctly attributed 88 of 241 trials to error, and incorrectly attributed 0 of 5 trials to error, $\chi^2(1) = 10$, $p < 0.05$, $\phi = 0.32$. For affirming feedback in the perverse condition, they correctly attributed 5 of 7 trials to error, and incorrectly attributed 20 of 271 trials to error, $\chi^2(1) = 2.9$, $p = 0.088$, $\phi = 0.16$. For disconfirming feedback they attributed 95 of 241 trials to error correctly, and incorrectly attributed 0 of 10 trials to error, $\chi^2(1) = 16$, $p < 0.05$, $\phi = 0.29$.

condition, participants correctly identified 61 of 121 actual errors and incorrectly identified 66 of 474 non-errors as error. This accuracy also slightly improved in judgments made at the end of the task for both the compatible and perverse incentive conditions ($\chi^2(1) = 121$, $p < 0.001$, $\phi = 0.45$; $\chi^2(1) = 167$, $p < 0.001$, $\phi = 0.57$, respectively).

Data Sharing

Participants in the compatible incentive condition shared 169 of 243 trials when the feedback was disconfirming ($74\%, SE = 5.9\%$) and 245 of 259 when it was affirming ($97\%, SE = 1.2\%$), $t(499) = 6.4$, $p < 0.001$, $d = 0.28$. Similarly, they shared 48 of 102 trials when they attributed feedback to error ($53\%, SE = 9.2\%$) and 437 of 471 when they judged it to be accurate ($97\%, SE = 1.3\%$), $t(570) = 8.8$, $p < 0.001$, $d = 0.37$.

At the end of the task, participants shared 112 of 180 trials that they judged to be an error ($71\%, SE = 16\%$) and 391 of 414 when they judged it to be accurate ($99\%, SE = 1.5\%$), $t(515) = 2.5$, $p = 0.037$, $d = 0.11$. However, there was also significant variation across participants in how much data they shared when they perceived the feedback to be an error, $\chi^2(1) = 62$, $p < 0.001$. As can be seen in Figure 1, most participants in the compatible incentive condition shared all of the trials they attributed to error at the end of the task, while a significant proportion shared none of those trials. However, there was no such variation for data sharing in response to disconfirming feedback $\chi^2(2) = 1.4$, $p < 0.63$, or error attributions during the task, $\chi^2(2) = 1.5$, $p < 0.65$.

Unexpectedly, participants in the perverse incentive condition did not share trials at lower rates than those in the compatible incentive condition. They shared 193 of 246 trials when the feedback was disconfirming ($88\%, SE = 5.6\%$) and 239 of 277 trials when it was affirming ($97\%, SE = 1.6\%$), $t(520) = 1.7$, $p < 0.19$, $d = 0.074$. They also shared 85 of 124 trials when they judged the feedback to be an error during the task ($82\%, SE = 8.1\%$) and 347 of 399 trials when they judged the feedback to be accurate ($96\%, SE = 2.1\%$), $t(520) = 2.9$, $p < 0.012$, $d = 0.13$. At the end of the task, they shared 74 of 130 trials that they judged to be an error ($81\%, SE = 20\%$) and 333 of 354 trials that they judged to be accurate ($100\%, SE = 0.34\%$), $t(481) = 2.1$, $p = 0.08$, $d = 0.098$.

As seen in Figure 5.3, there was significant variation across participants in their decisions to share data after receiving disconfirming feedback, $\chi^2(1) = 12$, $p = 0.0051$, whether they shared data that they perceived to be error during the task, $\chi^2(1) = 10$, $p = 0.014$, and whether they shared data that they perceived to be error at the end of the task, $\chi^2(1) = 27$, $p < 0.001$. For all three judgments, most participants in the perverse incentive condition shared all of their trials, with a minority sharing less.

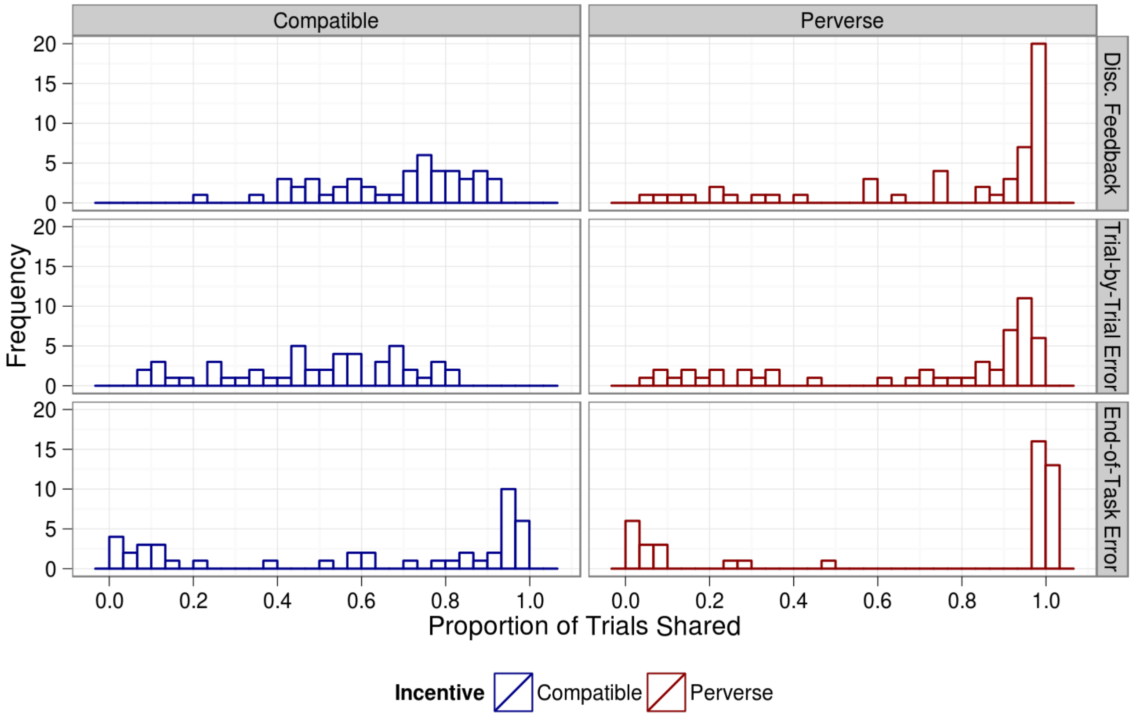


Figure 5.3: Proportion of trials shared by whether the trial was disconfirming (top row), whether participants attributed that trial to error during the task (middle row), and whether participants attributed the trial to error at the end of the task (bottom row).

Our prediction was that some participants would be seduced by the perverse incentive, thus deciding only to share trials that were consistent with their final answer. However, there was no difference between conditions in the probability of omitting data that were inconsistent with their final answer, $t(999) = 0.13$, $p = 0.79$. A second way that participants could produce these results while exploiting the perverse incentive would be to seek out only affirming data, knowing they data would make a simple and convincing story. One way to implement this weak testing strategy is to propose the (2,4,6) triple, knowing that they would receive affirming feedback unless the feedback is in error. However, participants in the two incentive conditions were equally likely to propose (2,4,6) triples, $t(1154) = 0.59$, $p = 0.67$.

As participants were both accurate and consistent in their error attributions, they may have been able to remove actual errors from the data they shared. Overall, at the end of the task participants shared 63 of 118 (53%) trials that were both actual errors and perceived as errors, 62 of 65 (95%) trials that were actual errors but not perceived as errors, 106 of 163

(65%) trials that were perceived as errors but not actual errors, and 615 of 650 (95%) trials that were neither perceived as error nor actual error. When including both main effects and the interaction between actual error and attribution of error to predict whether each trial would be shared at the end of the task, there was only a significant main effect of error attribution, and not actual error, for both compatible and perverse conditions ($t(509) = 4.7$, $p < 0.001$ vs. $t(479) = 5.2$, $p < 0.001$, respectively). This means that error attributions, but not actual errors, matter in determining whether data is shared.

The reason perceived and actual errors diverged was that disconfirmation had a systematic and additive effect on perceived error, even after controlling for actual error. Main effects of both actual error ($t(1025) = 7$, $p < 0.001$) and disconfirming feedback ($t(1025) = 7.3$, $p < 0.001$) increased the chance of attributing a trial to error at the end of the task, with no significant interaction between the two ($t(1025) = 1.6$, $p = 0.23$). Thus, affirming trials were shared more often, as they were less likely to be perceived as errors than disconfirming trials even when they were actually errors, whereas disconfirming trials were shared less frequently because they were inappropriately seen as errors when they were not.

5.3.3 Discussion

Participants with a compatible or perverse incentive to share data were equally likely to attribute disconfirming feedback to error. The financial penalty for making incorrect probability judgments and error attributions produced greater consistency and accuracy, compared to Experiments One and Two. Participants in both incentive conditions also shared fewer trials whose feedback was disconfirming or attributed to error, either during or at the end of the task. Although participants were successful in identifying actual errors, it was attributions of error that determined whether they shared trials, indicating that being able identify error does not preclude failing to share trials with accurate disconfirmations, while sharing ones with inaccurate affirmations.

We expected the perverse incentive to reduce the consistency and accuracy of error attributions, as well as to reduce the sharing of data attributed to error. However, such motivated reasoning was not observed. Rather, data sharing behavior in the two conditions differed in an unexpected way. Both for decisions made after each trial and at the end of the task, participants in the perverse incentive condition shared *more* data than those in the compatible incentive condition–thereby demonstrating a more ethical data sharing stance. While it is possible that higher stakes, such as those involved in pharmaceutical or academic research, would lead to motivated reasoning and data sharing policies, participants responded to the moderate stakes used in this research with reasoned and ethical behavior.

In decisions made at the end of all trials, however, some participants in the perverse incentive condition decided to share none of the data they attributed to error. Contrary to our prediction, participants in the perverse incentive condition did not omit more trials that were inconsistent with their final answer than those in the compatible incentive condition. Additionally, those in the perverse incentive condition did not try to produce a convincing story by taking as few trials as possible, in order to reduce the risk of collecting inconvenient data, either making their Final Answer less convincing or requiring selective reporting.

There are several possible explanations why participants in the perverse incentive condition shared trials at a higher rate than those in the compatible incentive condition. First, they may have thought that the learner knows they can hide data, even though the instructions indicated that only the trials they decided to share would be shared. Second, they may have believed that sharing more trials increases the learner's confidence, regardless of whether they are consistent with their Final Answer. Third, they may have been more strongly motivated to do the right thing and give the learner all the data available, even if that came at the cost of their own compensation. Four examples of such motivation:

1. "Yes. / It was an exercise in thinking about probabilities and cooperating with another. "
2. "I think there was some deception involved. This experiment may be about how willing the participant is to share money."
3. "seems more like a trust then a math problem. "
4. "I shared everything because, not knowing if the FIT/DNF response by the computer was correct, I didn't want to deliberately bias the info I passed on by being selective."

Thus, Experiment Three extends the positive test strategy to communication of results, seen in selective reporting, such that disconfirming data are seen as both caused by error and not worthy of sharing with others. Contrary to our prediction of motivated reasoning (104), the perverse incentive condition not only did not increase error attributions, but increased the sharing of data that were disconfirming or attributed to error.

5.4 General Discussion

Over 50 years of psychological research has found that hypothesis testing follows a positive test strategy (149), whereby people collect data that they expect to affirm their expectations and discount disconfirming data, should it nonetheless reach them. The present study asks how the positive test strategy affects data sharing. We use the Wason 2-4-6 rule discovery task (155), adding the possibility of error to simulate the uncertainty of actual research (109). In this task, participants seek to discover a rule by conducting 'experiments' to test their hypotheses about its answer, then receive affirming or disconfirming feedback, known to have a 20% error rate. We extended the task by adding several incentive schemes, then examining their effects on participants' decisions about sharing the feedback they received with another person. We also evaluated participants' performance in terms of the accuracy and consistency of their judgments of whether the feedback is error.

Experiment One replicated the pattern of results from previous studies, finding that disconfirming feedback is attributed to error more often than is affirming feedback (109). A new result is that participants' error attributions were generally consistent with their prior beliefs, in the sense of their being more likely to attribute affirmative feedback to error when they had strongly expected that the triple would not fit the rule, and being more likely to attribute disconfirming feedback to error when they had strongly expected the triple to fit the rule. However, their judgments of whether the feedback was in error were unrelated to its accuracy. Whether they shared trial results was unrelated to whether the feedback was disconfirming or attributed to error.

Experiment Two replicated Experiment One along with a new condition that provided participants with a large financial incentive for discovering the rule. As in Experiment One, participants attributed disconfirming feedback to error at a greater rate than affirming feedback in the control condition, but not the incentive condition. Those in the control condition again made error attributions that were somewhat consistent with their expectations but were quite inaccurate. In contrast, participants in the incentive condition were neither consistent nor accurate. Experiment Two elicited data sharing decisions after each trial, using a fixed-response format, unlike Experiment One which asked a single open-ended question at the end. Participants in both conditions were more likely to share feedback if it was affirming and perceived to be accurate.

Experiment Three introduced two incentive schemes for sharing data: (a) *compatible* incentives rewarded the sharer and receiver based on the receiver's success; (b) *perverse* incentives rewarded the sharer based on whether the receiver believed that the problem had been solved, and did not disclose when data were not shared. Both conditions penalized participants for making inaccurate probability and error judgments. As before, participants in both conditions were more likely to attribute feedback to error when it was disconfirming. The penalty increased both the accuracy and consistency of error attributions for participants in both conditions, compared to

Experiments One and Two. Contrary to prediction, participants with the perverse incentive shared more trials that were disconfirming or attributed to error than did participants with the compatible incentive. In both conditions, despite these participants' ability to identify error feedback, their perception of error was more important than actual error in determining their data sharing.

The present research has several internal and external validity limitations. In terms of internal validity, data sharing judgments were worded as "information sharing" possibly sending the message to participants in the perverse incentive condition that they should share rather than hide data. Many participants also discontinued their participation prematurely, explaining that the rule was too simple, not realizing that they had not identified it. For example:

> "If the rule did not seem initially clear to me, I would have gone through more trials. However, the number of trials I did and the results, given the instructions, seemed adequate to successfully complete the task."

Those who quit prematurely, proposing only a few trials, also proposed only trials that they expected to receive affirmation, and received only affirmation, except for rare errors that they were highly accurate in identifying. This confound limited participants' chance of obtaining disconfirming feedback. As disconfirming feedback is necessary for selective reporting, this confound causes the experiments to underestimate its magnitude.

The circumstances of the experiments differ from those of working scientists in several ways. First, scientists never know the exact error rates in their experiments, but have, instead, just a range of plausible values based on their experience and intuition. Those ambiguous error rates may be more readily modified to fit results than the fixed ones used in the experiments. Second, although the patterns observed here generally parallel those observed in real labs (150), the participants were either undergraduates or MTurk respondents, not scientists. The training and experience of working scientists may allow them to identify and report only accurate data, appropriately omitting errors that would confuse readers.

An additional experiment is needed to clarify the data-sharing results from Experiment Three. It found that participants given a perverse incentive behaved more ethically than participants in the compatible incentive condition, in the sense of sharing more trials that were disconfirming or attributed to error. One possible cause of this surprising result is that participants may have thought that the other participant knew they could hide trials, hence might become suspicious if data were too orderly. The second is that the sharers were genuinely willing to sacrifice their own pay to benefit others, with incentives that evoked ethical concerns. To determine which explanation is correct, Experiment Four will explicitly manipulate whether participants are told that the person receiving the data knows that the sharer does not have to include all the data, while also using more neutral language so the task is not perceived as being about cooperation or sharing. Additionally, all participants will be told the correct answer at the end of the task. They will then be allowed to modify the data they share, but not adjust their Final Answer. Thus, concern for ethics and altruism should lead participants to change the data they share to match the correct answer, even at the likely cost to their own payoff. However, if other concerns determine their data sharing, such as uncertainty about whether trials were errors or fear of being caught, then knowing the correct answer should allow participants to share only trials that are consistent with their Final Answer, especially for those who believe they cannot be caught.

Experiment Five will tie everything up with the best method derived from the previous experiments. For data sharing to matter, participants must not be able to solve the rule easily, as if they do solve the rule then there is no potential conflict between their Final Answer and the Actual Rule, and no opportunity for selective reporting. To do this, Experiment Five will use an alternative rule $(x, x^2, x^2 + 2)$ that should give more disconfirming feedback and be difficult to solve, thus encouraging participants to see the task as a challenge and complete more trials. Experiment Five will also use a more contextualized task, so that the terminology is easier to comprehend (e.g., true or false feedback).

The results of three experiments suggest that financial penalties are needed to help participants accurately evaluate their data. Without such penalties, Experiments One and Two elicited error attributions that were largely inaccurate and inconsistent with prior beliefs. In Experiment Three, adding a financial penalty for incorrect judgments substantially increased consistency and accuracy. However, they still shared data that were systematically biased by feedback, including inaccurate affirmations and excluding accurate disconfirmations. This selective reporting occurred even when poor data sharing could cost the sharer money, as in the compatible incentive condition.

The difficulty participants had when trying to avoid sharing errors shows that helpful selective reporting is not easy. One strategy participants could have used to achieve accurate selective reporting would be to use exact replications. Participants in all three experiments did not have the perfect accuracy in error attributions that would be required to selectively exclude errors from shared data. At the end of the task, exact replications would allow participants to clearly identify which trials were error and which were accurate, and, in turn, selectively report only accurate data.

Similar policies can help real scientists share data. Experiment Three found that penalties for incorrect probability judgments and error attributions greatly increased consistency and accuracy. One way to implement such a penalty would be to require that statistical analyses and experimental methods presented in published reports provide enough detail, in the paper or ancillary material, to be reproducible–with appropriate professional penalties for those who fail. As a protection, researchers can adopt the protocols of impartial organizations dedicated to independent replication of experiments and analyses (e.g., https://www.scienceexchange.com/). Another way of improving error identification is to encourage researchers to complete exact replications. These replications allow researchers to identify errors with high accuracy and make selective reporting of perceived errors highly accurate.

Part IV

Prescriptive

Introduction to the Prescriptive Analysis

The final part of the dissertation proposes methods of bringing human behavior, as determined by the descriptive analyses of Part Three, in line with normative standards, as proposed in Part Two. Chapter Two concluded that, although there is no logical ground for determining whether data or theory is faulty when they conflict, data sharing policies that omit disconfirming data are unethical because they impose conventions on the reader, thus deceiving them. However, Chapter Four found that surprising disconfirmations are perceived to be caused by error, and future observations that were seen as diffuse were judged to be less worthy of publication. Chapter Three concluded that disconfirmations are more likely to be errors than affirmations only when the selection of true hypotheses is common. However, participants in the Wason rule discovery task thought the opposite. With no penalty for incorrect error attributions, participants proposed triples that did not fit the rule (false hypotheses) more often than those that did fit the rule, but attributed error more often to disconfirmation than affirmation. Finally, in the rule discovery task, probability judgments improved with a financial penalty for incorrect answers. As judgments involving probability and statistics are always communicated in research, Chapter Six proposes methods of documenting data, methods, and statistical analyses so that penalties can be implemented when inferences are faulty.

Chapter 6

Open Communication

Up to this point, the dissertation has dealt mainly with the philosophical, mathematical, and psychological challenges to data sharing. In this chapter, I outline a simple procedure for implementing data sharing practically. It uses three technological solutions:

- Open-Data
- Open-Methods
- Open-Analyses

The hope is that, as these elements are laid out and standardized, journals are likely to change their policies to meet the standards, as indicated by one editor of the journal Nature (Spencer, 2010). Additional information on communication of research and uncertainty can be found in (159).

6.1 Open and Archived Data

To meet the criterion set out in Chapter Two of imposing minimal irrevocable conventions on the reader, a generic open-data convention is needed. Luckily, this has been done for us (OpenDefinition):

> "A piece of content or data is open if anyone is free to use, reuse, and redistribute it—subject only, at most, to the requirement to attribute and/or share-alike."

Along with this open-data definition, social scientists need to compile a list of conventions they consider important, and release a document like the CONSORT statement (95). Once these minimal conventions are agreed on then data can be documented and archived according to these conventions on a variety of websites, such as DataVerse and PsychFileDrawer.

All of the data and materials from Chapter 4 are here:

Davis, Alexander
“Surprises, Error, and Data Sharing”
http://hdl.handle.net/1902.1/14819
V3 [Version]

All of the data and materials from Chapter 5 are here:

Davis, Alexander
“Incentives, Error, and Data Sharing”
http://hdl.handle.net/1902.1/18699
V1 [Version]

6.2 Open and Archived Methods

The data are only half of the documentation process. The methods used to generate the data need to be as, or more, carefully documented. This can be seen as a problem of version-controlling one’s experiments (Perez, 2011), which can be dealt with using version controlling software like Git. This version controlling can be used for any part of the research process, from the development of methods, computational tools, and materials to the writing of papers and grants. References can be made to documentation of the run-up to discovery, or “warm-up period” using open lab notebooks such as OpenWetWare. Some pretesting from Chapters Four and Five are here Alex’s OpenWetWare.

6.3 Open and Reproducible Data Analysis

To avoid the need for forensic statistics, aimed at recreating the black box of what the authors must have done to their data (160), and to make statistical analyses truly reproducible, I use Sweave. Sweave integrates the free R statistical computing language with the free document preparation system LaTeX. All statistical analyses are coded in R directly into the Sweave document, which is embedded in a LaTeXdocument. Thus, any statistical analyses done can be easily read and reproduced along with the entire published paper with the Sweave document and the original data files. In fact, this entire dissertation was written this way, and the Sweave document can be obtained here:

Davis, Alexander
“The File-Drawer Problem (Dissertation)”
http://hdl.handle.net/1902.1/18786
V1 [Version]

I encourage readers to reproduce and check my code for errors, with rewards for those who succeed in finding errors.

Part V

Conclusion

Chapter 7

Recapitulation

This dissertation has evaluated the file-drawer problem—where disconfirming data are selectively excluded from published reports—from historical, normative, empirical, and prescriptive perspectives. The historical perspective suggests that incentives to publish only confirming data, as well as perceptions that disconfirming data are faulty, are the most likely causes of the file-drawer problem (Chapter One). These incentives threaten the scientific community and altruistic researchers, as the community can only identify valid data if penalties for selective reporting are high enough. Additionally, approaches that allow researchers to determine when data are faulty are based on possibly useful conventions that are not universally justified. Because researchers have some flexibility in determining the reporting conventions that are most appropriate for their circumstance, any ethical data sharing policy must not impose these conventions on other readers who must use their data, as this imposition is deceptive and, in turn, unethical (Chapter Two). Chapter Three analyzes two conventions that may be invoked to justify discarding disconfirming data: 1) that disconfirming data are less informative than affirming data, and 2) that disconfirming data are more likely to be faulty. The first conjecture was found to be true in the usual case social scientists face, where Type 1 errors are fixed at a much lower rate than Type 2 errors. The second conjecture, on the other hand, was found to be usually false, as the likelihood of error in disconfirming data depends mostly on whether one is generally good at choosing true hypotheses, which is most likely false when scientists are doing groundbreaking work. Chapter Four explores human judgments of error in scientific data in hindsight and foresight. Although the tendency to attribute disconfirming results to error was no greater in hindsight than foresight, these error attributions led to diffuse or uniform predictions for future data, and the greater the expected diffuseness of future data, the less likely participants were to see the data as worth sharing. Chapter Five found that, although participants were generally poor at picking triples that fit the rule, they expected disconfirming feedback to be error more than affirming feedback, violating the normative analysis in Chapter Three. Furthermore, participants shared disconfirming feedback less than affirming feedback, and this was strongly determined by their perception of fault in the data, rather than actual fault. Finally, Chapter Six proposes three simple technologies, open-data, open-methods, and open-analyses, that allow for minimal conventions to be imposed on those receiving data, promoting the ethical data sharing policies outlined in Chapter Two.

Chapter 8

Future Directions

This chapter discusses future directions that will follow the dissertation. The overarching goal is to promote logical and mathematical analyses of data sharing problems, examine how humans actually analyze and share data, and develop methods that can help identify error in data and effectively communicate results.

8.1 Normative

Two open questions remain from the normative analysis: 1) how do we establish the conventions of members of the scientific community, and 2) what general data sharing policies can be developed if one wants to maximize the informativeness of the shared data? The first part will explore the formal and practical implications of basing sharing policies on the explicit conventions of members of the scientific community, as well as create a basic structure, or ontology such as the OWL (161), such that "minimal conventions" can be precisely defined and standardized (162). The second part will integrate mathematical analyses of information sharing and social learning from Chamley (163) and Hirshleifer (8) to accommodate agents with bounded rationality in terms of confusion and attention.

8.2 Descriptive

To the maximum extent possible, future empirical research will be selected based on the ability to meet the following criteria:

1. The task should be a real problem, not a made-up one.

2. Solving the real problem should provide a real social good.

3. Solving the real problem should allow participants to learn.

4. Participants should be allowed to contribute or co-author the final paper.

8.2.1 Surprises, Error, and Data Sharing

Concluding studies

A final approach to examining the effect of hindsight on error attributions would be to use the more traditional hindsight paradigm, where participants are asked to 're-judge' their prior beliefs after receiving outcome knowledge. Specifically, they would be asked "how likely would you have been to identify this cause of the error?" versus having them actually do so in foresight.

Suppositional versus conditional causes and data

Experiments Two through Four found that observed data are expected to be more likely to replicate than supposed data (164), but the probability of the causes of supposed versus observed data do not change. This line of research will focus on discovering why there is this asymmetry between supposing and observing data versus supposing and observing causes.

Deciding on hypothesis generation versus evaluation

Experiment Five found that more 'natural' explanations that come to mind easily when designing an experiment are also seen as relatively more likely after observing the results compared to explanations that may have required more thought (and even empirical observation) to generate. Deeper probing in foresight may help, by making sure all explanations that are serious possibilities are considered before observing the results. Unfortunately, there is no limit to this time consuming and often frustrating process, so the termination of this process ultimately depends on a judgment that the so-far-considered explanations are 'good enough'. This line of research will look at how lay participants and scientists determine when hypothesis generation is good enough to cover all the serious explanations and possible errors of an experiment, and then engage in the experiment itself, rather than remaining in a purely reflective mode, as philosophers do.

The problem of old evidence

A third set of experiments will extend Experiment Five and examine the degree to which a hypothesis (either a core hypothesis or error model) not considered a priori can be believed after observing the data. Given that Bayesian calculations do not apply in this situation, what psychological processes are involved in determining the posterior convincingness of hypotheses? This is closely related to the problem of old evidence (165) and belief revision (166; 167), and is a common problem faced by the FDA (168):

> "Dr Schultz: I think the question related to new information from the outside. We have been in situations where new data (e.g., from other studies) have come to light as we were analyzing a specific study, which could influence the outcome either positively or negatively. This is a difficult problem, and I am not sure exactly how to deal with it. In particular, it is hard to suggest ignoring negative information that could impact our assessment of the safety of the product".

Polanyi's wild goose chase

A fourth set of experiments will examine how participants decide whether to pursue anomalies or continue on their research project; that is, whether and how they decide to engage in the 'wild goose chase'. This will also look at whether a 'warm-up' period is used and whether this period is flexibly defined to selectively report data.

8.2.2 Wason Rule Discovery Task

Networks and Communication

Behavioral data sharing policies can be compared to rational analyses of social learning (163) in simulated environments where multi-way communication is possible. For example, participants in the Wason rule-discovery task can engage in rule discovery in parallel and complementary ways, sometimes even in direct competition. They can also communicate their results back and forth to each other directly or through intermediaries (e.g., journals) that determine how the data will be disseminated and what rewards each researcher receives.

Debugging

An extension to the Wason paradigm that provides both greater external validity and practical application would be hypothesis testing and error identification in programming and debugging code. Every code is an attempt to solve a problem, a hypothesis, or a conjecture. When the code does not work, there are a number of ways it could fail. Importantly, a program could fail either because it doesn't actually solve the problem or because of a typo or bug in the code. This is a genuine hypothesis testing problem with the possibility of error, with real consequences, and can be used to solve real world problems as well as teach participants to program, a valuable skill.

Computational Modeling

Markov Decision Processes (169) provide an important modeling extension to the Bayesian analyses proposed so far. Computational models using Markov Decision Processes will be compared against actual human behavior.

Collaborative filtering and review

A field experiment will test whether an open and collaborative publication system could work, similar to ArXiv or Stack Exchange, by getting error models from the general public.

Crowdsourced discovery

'Crowdsourced' science can be used both to solve real world social science problems (170), possibly faster than an individual experimenter could, while simultaneously providing externally valid data on scientific reasoning and educating the public by engaging them in real science (170). The idea is to give MTurk participants a real social science/science problem and see what they do, following the process enumerated below:

1. I choose the problem.
2. The crowd comes up with the solution.
3. I conduct the actual study.
4. I give them feedback.
5. Repeat.

To really study scientific reasoning, participants have to make their own instruments, develop real hypotheses, and get real feedback. As most lay people could develop a survey to test a simple hypothesis, this task fits well. Thus, by selecting a real social science survey research problem and having participants develop their own surveys to test it, one can get data on real scientific reasoning in a controlled manner. It is also possible to compare this to a graduate student trying to solve the problem, to see who can come up with faster/better/more cost efficient solutions. It also educates the public about science directly, providing another social good.

8.3 Prescriptive

I intend to write a book, *Breeding Orchids*, that thoroughly discusses how benchwork is and should be done, focusing on pre-testing, pilot-testing, hypothesis formulation, and evidence synthesis. The Breeding Orchids approach will be applied to different research projects, including my own, as well as adapted to pharmaceutical drug research in the hopes of process improvement, as “the drug discovery process involves blunders, wrong turns, false hypotheses as well as successes, but none of the failures are being reported or published in the journals and other sources” (171). The meta-analytic approach will also be extended to include formal proofs, as well as tested for usability and practical application. Application would ideally be on research involving cutting edge experiments, such as drug discovery involving a pharmaceutical company or academic researchers.

8.4 Other Causes

There are also a variety of other causes of the file-drawer problem beside the two investigated in this dissertation. These are listed and briefly discussed below.

8.4.1 Ad-hoc Methods

One reason for the file-drawer problem in Psychology is that methods are rarely standardized or repeated, so there is much room for negative results. Psychologists almost always develop their own paradigms, including materials and procedures, to test novel hypotheses. The tinkering involved in this process involves many experiments that produce null results. Take Michael Gorman’s experience:

> “Behind virtually every published experiment is an extensive series of such pilot studies, where one tinkers with procedures and variables to find a promising combination. Such tinkering is never referred to in a published report, except perhaps as a footnote, but the most significant discoveries often occur when piloting” (pg. 81-82, 87) (102).

He leaves us with the impression that experimental psychology produces vastly more experiments than are reported, and these unreported experiments are inconsistent with the experimenter’s theory. As a result, one would expect that any experimental psychologist, if properly incentivized to be honest (for example, offering him or her tenure), would probably tell us that they conduct anywhere between 3 and 10 times as many experiments as they report. Unfortunately, empirically verifying this fact is very difficult.

8.4.2 No Exact Replications

Sterling (12) argues that publication bias causes a perverse cycle, where non-significant results are not reported, and without knowing this, others will replicate this failed result. In contrast replications do not occur for studies that get significant results. Thus, the profession proceeds in a wasteful cycle of replicating experiments that test false hypotheses that we should have known were false if null results were properly documented, and not replicating hypotheses that we believe are true because no one has a reason to do so.

8.4.3 Over-Optimism

Sometimes file-drawers emerge because an experiment seems easier than it actually is. In the case of the simple experiment to demonstrate cold fusion by Pons and Fleishman of the University of Utah, many poor experiments were conducted because "many were taken in by the seeming ease of the experiment only to discover that a palladium electrolytic cell was a deal more complicated than expected" (48).

8.4.4 Lack of Interest

Reysen (172) surveyed 237 faculty members in Psychology, finding that faculty members don't write up non-significant results because they see them as unpublishable, a waste of time, due to flawed in results, that they cannot understand the results, or that the results are useless. Similarly, non-significant results are "generally boring; it's difficult to get up the enthusiasm to write them up, and it's difficult to get them published in decent journals" (173). The boring, uninteresting work of writing up negative results is seen as telling us "nothing new or interesting", and as laughable as "a cover story in Nature trumpeting people Can't Fly!" (174). There also may be a generational component, as David Singer sees unwillingness to "benefit from a nice piece of research that looks like it tells us nothing" as an attitude that applies to older scientists, whereas "young scientists are right to insist we start publishing negative results" (175). Media reports are more likely to pick up on "interesting positive results showing cancerous effects of nuclear radiation than 'uninteresting' negative results showing no effect" (176). Timmer *et al.* (177) surveyed authors of abstracts to see whether they published their results subsequently. Most failed to follow up on negative results because they found them uninteresting, or too difficult to publish, or didn't have time. They did not find that studies with statistically significant results were more likely to be published than those that weren't significant, but those with positive results did have a higher impact (in terms of citations) post-publication. As in psychology, publishing data that suggest a medical treatment is ineffective is difficult. For example, journal editors will argue that such evidence is not “novel” (33). Failures to publish non-significant results are likely due to lack of motivation on the part of the researcher.

8.4.5 Low Power and Rare Discoveries

John Ioannidis became interested in the problem of biased medical research and unwarranted conclusions when “poring over medical journals, he was struck by how many findings of all types were refuted by later findings” (41). He uses a simple Bayesian formulation to demonstrate that, as long as the prior probability of discovering a truly effective medical treatment is low, statistical tests with very high sensitivity and specificity will still error more of the time when they lead one to conclude a discovery has been made compared to when they lead to the conclusion of no discovery (3). Additionally, when sample sizes are small, effect sizes are small, there is flexibility in procedures and definitions, financial conflicts of interests, and competition, bias is likely to greatly increase the proportion of false discoveries deemed true. Future directions could involve evaluating the rarity of true hypotheses and discoveries among those that are conjectured.

8.4.6 Competition

Hilary Spencer of the Nature publishing group remarked that that researchers report not sharing their data with others because they want to maintain a “competitive advantage in publication” (pg. 3). A recent anonymous editorial in Nature points out that a researcher may “hoard her samples out of fear of competition; another doggedly promotes his hypothesis long after the data have falsified it; negative results are hidden because of competing financial interests” (178). Negative data may be a type of public good “which helps other laboratories while not materially advancing their own reputations” (51). Future directions involving competition between researchers in simulated environments could examine whether incentives in the form of competition lead to the file-drawer problem.

8.5 Conclusion

The file-drawer problem has been a concern for social scientists over the last fifty years, and seems to be getting worse (55; 11). However, a recent burst of activity has focused researchers, across a variety of disciplines, on this issue, with transparent approaches to documenting data (63), statistical analyses (62), and methods (OpenWetWare). Hilary Spencer of the Nature publishing group expects journals to follow the lead of researchers when we clearly articulate our data sharing policies. The journal *Perspectives on Psychological Science* has a new special section dedicated to the file-drawer problem, thanks to Bobbie Spellman (68).

The file-drawer problem is closely related to the dilemma of trying to understand why failed predictions occur, either in replications of the experiments of others, of our own, or for new experiments. As discussed in Chapter Two, determining whether theory or data are false when they conflict relies on convention. However, Chapters Four and Five show that these conventions are determined by the perception of error, or a fallible 'psychology of observation', which in turn is determined by the outcome of the experiment. As Chapter Five shows, error attributions are more likely to be made, and data less likely to be shared, when the outcome of the experiment is disconfirming, even after accounting for actual error.

How can this knowledge of error perception and communication improve scientific methods and reporting? First, we can acknowledge that we see ourselves as more like Millikan (potential Nobel laureates) than Blondlot (fooling ourselves), but *behave* more like Blondlot than Millikan, as indicated by both error attributions and data sharing policies. Thus, as open-access data advocates have argued, the norm of sharing all of our data should allow the "story of the failures that make the successes possible" (137, p. 15) to be told when our intuition would prescribe otherwise. While this open approach helps others understand what we've done, it also helps make clear that the pattern of failed prediction and ad-hoc error attribution are an inevitable part of everyday scientific practice, regardless of the amount of pre-testing and pilot-testing we conduct, and thus education must include this process. Although sometimes briefly discussed in research methods courses, the methodology, logic, and appropriate reporting of pre-tests and pilot-tests is not clearly defined. This dissertation makes clear the need for educational programmes that include these practical elements of 'benchwork', that are currently implicit and hidden.

The dissertation is an example of evidence-based methodological research, using logical and mathematical analysis, empirical observation and experimentation. Hopefully, the recent willingness of psychologists and related social scientists to embrace serious methodological problems in their fields, such as the file-drawer problem, will make the solutions themselves an example of the better science that is needed. In the dissertation I tried to do this. The one spark that the file-drawer problem has ignited, both in myself and among other social scientists, is an affirmation that there is only one rule when doing science: *not fooling ourselves.* I was allowed to make this dissertation transparent, open, and skeptical, and this has kept my spark alive. Hopefully I can do this for others. I believe maintaining this spark will keep the field alive, and as it goes out, so too will the practical and ideological goals of social scientists extinguish.

Bibliography

[1] R. Feynman, “Cargo cult science,” *Engineering and Science*, vol. 37, no. 7, pp. 10–13, 1974. 2

[2] J. Probst, “Prisoners’ dilemma: The importance of negative results,” *Family Medicine-Kansas City*, vol. 38, no. 10, p. 742, 2006. 2, 8

[3] J. Ioannidis, “Why most published research findings are false,” *PLoS Medicine*, vol. 2, no. 8, p. e124, 2005. 2, 9, 13, 29, 112

[4] R. Rosenthal, “The file drawer problem and tolerance for null results.,” *Psychological Bulletin*, vol. 86, no. 3, p. 638, 1979. 3, 70

[5] R. Rosenthal and D. Rubin, “[selection models and the file drawer problem]: Comment: Assumptions and procedures in the file drawer problem,” *Statistical Science*, vol. 3, no. 1, pp. 120–125, 1988. 3

[6] G. Guyatt, A. Oxman, V. Montori, G. Vist, R. Kunz, J. Brozek, P. Alonso-Coello, B. Djulbegovic, D. Atkins, Y. Falck-Ytter, *et al.*, “Grade guidelines: 5. rating the quality of evidence–publication bias,” *Journal of Clinical Epidemiology*, 2011. 3

[7] T. Tse, R. Williams, and D. Zarin, “Reporting basic results in clinicaltrials. gov,” *Chest*, vol. 136, no. 1, pp. 295–303, 2009. 3

[8] D. Hirshleifer and S. Teoh, “Limited attention, information disclosure, and financial reporting,” *Journal of Accounting and Economics*, vol. 36, no. 1, pp. 337–386, 2003. 3, 106

[9] R. Steinbrook and J. Kassirer, “Data availability for industry sponsored trials: what should medical journals require?,” *BMJ*, vol. 341, 2010. 3

[10] R. DeVellis, *Scale development: Theory and applications*, vol. 26. Sage Publications, Inc, 2011. 3

[11] D. Fanelli, “Negative results are disappearing from most disciplines and countries,” *Scientometrics*, pp. 1–14, 2012. 3, 9, 13, 73, 112

[12] T. Sterling, “Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa,” *Journal of the American statistical association*, pp. 30–34, 1959. 3, 9, 73, 111

[13] M. Mahoney, “Publication prejudices: An experimental study of confirmatory bias in the peer review system,” *Cognitive Therapy and Research*, vol. 1, no. 2, pp. 161–175, 1977. 3, 5, 69

[14] J. Cohen, “A power primer.,” *Psychological Bulletin*, vol. 112, no. 1, p. 155, 1992. 3

[15] J. Cohen, “The statistical power of abnormal-social psychological research: A review.,” *Journal of Abnormal and Social Psychology*, vol. 65, no. 3, pp. 145–153, 1962. 3

[16] P. Sedlmeier and G. Gigerenzer, “Do studies of statistical power have an effect on the power of studies?,” *Psychological Bulletin; Psychological Bulletin*, vol. 105, no. 2, p. 309, 1989. 3

[17] J. Rossi, “Statistical power of psychological research: What have we gained in 20 years?,” *Journal of Consulting and Clinical psychology*, vol. 58, no. 5, p. 646, 1990. 3

[18] D. Bem, “Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect.,” *Journal of Personality and Social Psychology*, vol. 100, no. 3, p. 407, 2011. 4

[19] E. Wagenmakers, R. Wetzels, D. Borsboom, and H. Van Der Maas, “Why psychologists must change the way they analyze their data: the case of psi: comment on bem (2011).,” 2011. 4

[20] J. Simmons, L. Nelson, and U. Simonsohn, “False-positive psychology,” *Psychological Science*, vol. 22, no. 11, pp. 1359–1366, 2011. 4, 45

[21] E. Yong, “Bad copy,” *Nature*, vol. 485, no. 7398, pp. 298–300, 2012. 4

[22] D. DeMets and C. Meinert, “Data integrity.,” *Controlled Clinical Trials*, vol. 12, no. 6, p. 727, 1991. 4

[23] S. Reynolds, “Ori findings of scientific misconduct in clinical trials and publicly funded research, 1992–2002,” *Clinical Trials*, vol. 1, no. 6, pp. 509–516, 2004. 4

[24] M. Heger, “Clinical trial website struggles to serve as research data hub,” *Nature Medicine*, vol. 18, no. 6, pp. 837–837, 2012. 4

[25] E. Yong, "Replication studies: Bad copy.," *Nature*, vol. 485, no. 7398, p. 298, 2012. 4, 8

[26] K. Dickersin and Y. MIN, "Publication bias: the problem that won't go away," *Annals of the New York Academy of Sciences*, vol. 703, no. 1, pp. 135–148, 1993. 4

[27] E. Hasenboehler, I. Choudhry, J. Newman, W. Smith, B. Ziran, and P. Stahel, "Bias towards publishing positive results in orthopedic and general surgery: a patient safety issue?," *Patient Safety in Surgery*, vol. 1, no. 1, pp. 1–6, 2007. 4, 9

[28] B. Martinson, M. Anderson, and R. De Vries, "Scientists behaving badly," *Nature*, vol. 435, no. 7043, pp. 737–738, 2005. 4

[29] R. Lehman and E. Loder, "Missing clinical trial data," *BMJ*, vol. 344, 2012. 4

[30] J. Stern and R. Simes, "Publication bias: evidence of delayed publication in a cohort study of clinical research projects," *BMJ*, vol. 315, no. 7109, pp. 640–645, 1997. 4

[31] J. Ioannidis, "Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials," *JAMA: the Journal of the American Medical Association*, vol. 279, no. 4, pp. 281–286, 1998. 4, 8

[32] E. Turner, A. Matthews, E. Linardatos, R. Tell, and R. Rosenthal, "Selective publication of antidepressant trials and its influence on apparent efficacy," *New England Journal of Medicine*, vol. 358, no. 3, pp. 252–260, 2008. 4

[33] D. Vergano, "Filed under f (for forgotten)," *USA Today*, 2001. 5, 112

[34] N. Sussman, "The file-drawer effect: Assessing efficacy and safety of antidepressants.," 2004. 5

[35] G. Yamey, "Scientists who do not publish trial results are unethical," *BMJ: British Medical Journal*, vol. 319, no. 7215, p. 939, 1999. 5

[36] F. Godlee and E. Loder, "Missing clinical trial data: setting the record straight," *BMJ*, vol. 341, 2010. 5

[37] D. Madigan, D. Sigelman, J. Mayer, C. Furberg, and J. Avorn, "Underreporting of cardiovascular events in the rofecoxib alzheimer disease studies," *American Heart Journal*, 2012. 5

[38] L. Hedges, “How hard is hard science, how soft is soft science? the empirical cumulativeness of research.,” *American Psychologist*, vol. 42, no. 5, p. 443, 1987. 5

[39] A. Franklin, “Millikan’s oil-drop experiments,” *The Chemical Educator*, vol. 2, no. 1, pp. 1–14, 1997. 5, 43

[40] H. Collins, “Lead into gold: the science of finding nothing,” *Studies In History and Philosophy of Science Part A*, vol. 34, no. 4, pp. 661–691, 2003. 5, 6

[41] D. Freedman, “Lies, damned lies, and medical science,” *The Atlantic*, vol. 306, no. 4, pp. 76–84, 2010. 5, 112

[42] NA, “The sounds of silence: Negative clinical-trial results are underreported. but this may soon change,” *The Economist*, p. e124, 2004. 5

[43] F. Godlee, “Research misconduct is widespread and harms patients,” *BMJ*, vol. 344, 2012. 5

[44] J. Karlawish, P. Whitehouse, and R. McShane, “Silence science: the problem of not reporting negative trials,” *Alzheimer Disease & Associated Disorders*, vol. 18, no. 4, p. 180, 2004. 5

[45] J. Ioannidis, “Adverse events in randomized trials: neglected, restricted, distorted, and silenced,” *Archives of Internal Medicine*, vol. 169, no. 19, p. 1737, 2009. 5

[46] C. Begley and L. Ellis, “Drug development: Raise standards for preclinical cancer research,” *Nature*, vol. 483, no. 7391, pp. 531–533, 2012. 6, 9

[47] M. Anestis, “When nothing is something important: Why everyone should know about null findings,” 2010. 6

[48] H. Collins and T. Pinch, *The golem: What you should know about science.* Cambridge Univ Pr, 1998. 7, 111

[49] I. Langmuir, “Pathological science: Colloquium at the knolls research laboratory, december 18, 1953,” 1953. 8

[50] K. Liddle, “Accentuate the negative.” `http://www.thetwentyfirstfloor.com/?p=2278`, 2011. 8

[51] D. McCormick, “A positive need for negative data,” *Biotechniques*, vol. 43, no. 4, p. 389, 3007. 8, 112

[52] S. Rockwell, B. Kimler, and J. Moulder, "Publishing negative results: the problem of publication bias," *Radiation Research*, vol. 165, no. 6, pp. 623–625, 2006. 8

[53] U. Dirnagl *et al.*, "Fighting publication bias: introducing the negative results section," *Journal of cerebral blood flow and metabolism: official journal of the International Society of Cerebral Blood Flow and Metabolism*, vol. 30, no. 7, p. 1263, 2010. 9

[54] A. Khan, S. Khan, and W. Brown, "Are placebo controls necessary to test new antidepressants and anxiolytics?," *The International Journal of Neuropsychopharmacology*, vol. 5, no. 3, pp. 193–197, 2002. 9

[55] D. Fanelli, "Positive results increase down the hierarchy of the sciences," *PloS One*, vol. 5, no. 4, p. e10068, 2010. 9, 112

[56] D. Fanelli, "Do pressures to publish increase scientists' bias? an empirical support from us states data," *PLoS One*, vol. 5, no. 4, p. e10271, 2010. 9, 74

[57] K. Popper, *The logic of scientific discovery*. Psychology Press, 2002. 9, 17, 18

[58] H. Poincaré, *Science and hypothesis*. Science Press, 1905. 9, 17

[59] I. Lakatos, J. Worrall, and G. Currie, *The methodology of scientific research programmes*, vol. 1. Cambridge Univ Pr, 1980. 9, 14, 16, 19, 20, 42

[60] P. Shafto, B. Eaves, D. Navarro, and A. Perfors, "Epistemic trust: Modeling children's reasoning about others' knowledge and intent," *Developmental Science*. 9, 29, 86

[61] V. Stodden, "Reproducible research: Tools and strategies for scientific computing," *Computing in Science & Engineering*, pp. 11–12, 2012. 9

[62] V. Stodden, "Enabling reproducible research: Open licensing for scientific innovation," *International Journal of Communications Law and Policy, Forthcoming*, 2009. 9, 112

[63] G. King, "An introduction to the dataverse network as an infrastructure for data sharing," *Sociological Methods & Research*, vol. 36, no. 2, pp. 173–199, 2007. 9, 112

[64] F. Leisch, "Sweave. dynamic generation of statistical reports using literate data analysis.," 2002. 9

[65] D. Von Winterfeldt, W. Edwards, *et al.*, *Decision Analysis and Behavioral Research*, vol. 1. Cambridge University Press Cambridge, 1986. 9

[66] B. Fischhoff, "Judgment and decision making," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 1, no. 5, pp. 724–735, 2010. 9

[67] D. Bell, H. Raiffa, and A. Tversky, "Descriptive, normative, and prescriptive interactions in decision making," *Decision making: Descriptive, normative, and prescriptive interactions*, vol. 1, pp. 9–32, 1988. 9

[68] B. Spellman, "Introduction to the special section," *Perspectives on Psychological Science*, vol. 7, no. 1, pp. 58–59, 2012. 13, 44, 113

[69] L. Wasserman, "A world without peer review," 2012. 13

[70] R. De Vries, M. Anderson, and B. Martinson, "Normal misbehavior: Scientists talk about the ethics of research," *Journal of Empirical Research on Human Research Ethics: JERHRE*, vol. 1, no. 1, p. 43, 2006. 13

[71] T. Kuhn, *The structure of scientific revolutions*. University of Chicago press, 1996. 13, 14, 16, 17, 72

[72] M. Spence, "Job market signaling," *The Quarterly Journal of Economics*, vol. 87, no. 3, pp. 355–374, 1973. 14

[73] P. Laplace, "A philosophical essay on probabilities," 1806. 15, 24, 26

[74] A. Gelman and C. Shalizi, "Philosophy and the practice of bayesian statistics," *British Journal of Mathematical and Statistical Psychology*, 2010. 15, 25

[75] Z. Pylyshyn, *The robot's dilemma: The frame problem in artificial intelligence*, vol. 4. Ablex Publishing Corporation, 1987. 15

[76] M. Polanyi, *Personal knowledge: Towards a post-critical philosophy*. Psychology Press, 1998. 17

[77] D. Mayo, *Error and the growth of experimental knowledge*. University of Chicago Press, 1996. 17, 23, 71

[78] P. Duhem, *The aim and structure of physical theory*, vol. 13. Princeton Univ Pr, 1991. 17

[79] C. Radhakrishna Rao, “Ra fisher: The founder of modern statistics,” *Statistical Science*, vol. 7, no. 1, pp. 34–48, 1992. 20

[80] L. Savage, “On rereading ra fisher,” *The Annals of Statistics*, vol. 4, no. 3, pp. 441–500, 1976. 20, 21

[81] E. Lehmann, “The fisher, neyman-pearson theories of testing hypotheses: One theory or two?,” *Journal of the American Statistical Association*, pp. 1242–1249, 1993. 21

[82] S. Zabell, “Ra fisher and fiducial argument,” *Statistical Science*, pp. 369–387, 1992. 21

[83] R. Fisher, “Statistical methods and scientific inference.,” 1956. 21, 25

[84] R. Fisher, “The design of experiments.,” 1935. 21

[85] J. Neyman, “Outline of a theory of statistical estimation based on the classical theory of probability,” *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 236, no. 767, pp. 333–380, 1937. 21

[86] J. Neyman and E. Pearson, “On the problem of the most efficient tests of statistical hypotheses,” *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933. 22

[87] J. Neyman, “Frequentist probability and frequentist statistics,” *Synthese*, vol. 36, no. 1, pp. 97–131, 1977. 22

[88] A. Greenwald, “Consequences of prejudice against the null hypothesis.,” *Psychological Bulletin*, vol. 82, no. 1, p. 1, 1975. 23

[89] J. Kadane, *Principles of uncertainty*, vol. 92. Chapman & Hall, 2011. 24, 25

[90] T. Seidenfeld, “Why i am not an objective bayesian; some reflections prompted by rosenkrantz,” *Theory and Decision*, vol. 11, no. 4, pp. 413–440, 1979. 24, 26

[91] A. O’Hagan, C. Buck, A. Daneshkhah, J. Eiser, P. Garthwaite, D. Jenkinson, J. Oakley, and T. Rakow, *Uncertain judgements: eliciting experts’ probabilities*, vol. 2. Wiley Chichester, 2006. 25

[92] D. Danks and F. Eberhardt, “Explaining norms and norms explained,” 2008. 25

[93] H. Jeffreys, “An invariant form for the prior probability in estimation problems,” *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, 1946. 26

[94] E. Jaynes, “Information theory and statistical mechanics,” *Statistical Physics. Brandeis Lectures*, vol. 3, pp. 160–185, 1963. 26

[95] K. Schulz, D. Altman, D. Moher, *et al.*, “Consort 2010 statement: updated guidelines for reporting parallel group randomised trials,” *BMC Medicine*, vol. 8, no. 1, p. 18, 2010. 27, 100

[96] J. Overall, “Classical statistical hypotheses testing within the context of bayesian theory,” *Psychological Bulletin*, vol. 71, no. 4, pp. 285–292, 1969. 29, 37

[97] B. Fitelson, “The plurality of bayesian measures of confirmation and the problem of measure sensitivity,” *Philosophy of Science*, pp. 362–378, 1999. 30

[98] A. Tversky and D. Kahneman, “Belief in the law of small numbers.,” *Psychological Bulletin*, vol. 76, no. 2, p. 105, 1971. 37

[99] B. Fischhoff, N. Welch, and S. Frederick, “Construal processes in preference assessment,” *Journal of Risk and Uncertainty*, vol. 19, no. 1, pp. 139–164, 1999. 42

[100] W. Shadish, T. Cook, and D. Campbell, “Experimental and quasi-experimental designs for generalized causal inference,” 2002. 42

[101] D. Goodstein, “In defense of robert andrews millikan,” *Engineering and Science*, vol. 63, no. 4, pp. 30–38, 2000. 43

[102] M. Gorman, *Simulating science: heuristics, mental models, and technoscientific thinking*, vol. 31. Indiana university press Bloomington and Indianapolis, 1992. 43, 110

[103] N. Roese and J. Olson, “Counterfactuals, causal attributions, and the hindsight bias: A conceptual integration.,” *Journal of Experimental Social Psychology*, 1996. 43

[104] Z. Kunda, “The case for motivated reasoning.,” *Psychological Bulletin; Psychological Bulletin*, vol. 108, no. 3, p. 480, 1990. 43, 74, 93

[105] I. Klotz, “The n-ray affair,” *Scientific American*, vol. 242, no. 5, pp. 122–131, 1980. 43

[106] R. Wood, "The n-rays," *Nature*, vol. 70, pp. 530–531, 1904. 43

[107] M. Gorman, *Scientific and technological thinking.* Lawrence Erlbaum, 2005. 43, 73

[108] M. Gorman, "Error, falsification and scientific inference: An experimental investigation," *The Quarterly Journal of Experimental Psychology*, vol. 41, no. 2, pp. 385–412, 1989. 43, 74, 80

[109] D. Penner and D. Klahr, "When to trust the data: Further investigations of system error in a scientific reasoning task," *Memory & cognition*, vol. 24, no. 5, pp. 655–668, 1996. 43, 70, 74, 75, 78, 80, 94

[110] K. Dunbar, "What scientific thinking reveals about the nature of cognition," *Designing for science: Implications from everyday, classroom, and professional settings*, pp. 115–140, 2001. 43, 73

[111] C. Lord, L. Ross, and M. Lepper, "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence.," *Journal of Personality and Social Psychology*, vol. 37, no. 11, p. 2098, 1979. 43, 69, 73

[112] J. Klayman and Y. Ha, "Hypothesis testing in rule discovery: Strategy, structure, and content.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 15, no. 4, p. 596, 1989. 43

[113] R. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises.," *Review of General Psychology; Review of General Psychology*, vol. 2, no. 2, p. 175, 1998. 43

[114] H. Blank, S. Nestler, G. Von Collani, and V. Fischer, "How many hindsight biases are there?," *Cognition*, vol. 106, no. 3, pp. 1408–1440, 2008. 44

[115] J. Christensen-Szalanski and C. Willham, "The hindsight bias: A meta-analysis," *Organizational Behavior and Human Decision Processes*, vol. 48, no. 1, pp. 147–168, 1991. 44

[116] P. Slovic and B. Fischhoff, "On the psychology of experimental surprises.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 3, no. 4, p. 544, 1977. 44, 45, 46, 49, 50, 57, 64, 69

[117] S. Hawkins and R. Hastie, "Hindsight: Biased judgments of past events after the outcomes are known.," *Psychological Bulletin*, vol. 107, no. 3, p. 311, 1990. 44

[118] J. Holland, K. Holyoak, and R. Nisbett, *Induction: Processes of inference, learning, and discovery*. The MIT Press, 1989. 44

[119] M. Pezzo, “Surprise, defence, or making sense: What removes hindsight bias?,” *Memory*, vol. 11, no. 4-5, pp. 421–441, 2003. 44

[120] N. Roese, “Counterfactual thinking.,” *Psychological Bulletin*, vol. 121, no. 1, p. 133, 1997. 44

[121] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009. 44

[122] C. Glymour, “Learning, prediction and causal bayes nets,” *Trends in Cognitive Sciences*, vol. 7, no. 1, pp. 43–48, 2003. 44

[123] T. Griffiths and J. Tenenbaum, “Theory-based causal induction.,” *Psychological Review*, vol. 116, no. 4, p. 661, 2009. 44

[124] B. Fischhoff, “Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty.,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 1, no. 3, p. 288, 1975. 44

[125] N. Christensen, P. Shawhan, G. Gonzlez, *et al.*, “Vetoes for inspiral triggers in ligo data,” *Classical and Quantum Gravity*, vol. 21, p. S1747, 2004. 44

[126] N. Christensen, “Veto studies for ligo inspiral triggers,” *Classical and Quantum Gravity*, vol. 22, p. S1059, 2005. 44

[127] J. Crocker and M. Cooper, “Addressing scientific fraud,” *Science*, vol. 334, no. 6060, pp. 1182–1182, 2011. 44

[128] J. Bradley, “Open notebook science using blogs and wikis,” *Nature Precedings*, 2007. 44

[129] S. Everts, “Open-source science,” *Chemical & engineering news*, vol. 84, no. 30, pp. 34–35, 2006. 44

[130] N. Kerr, “Harking: Hypothesizing after the results are known,” *Personality and Social Psychology Review*, vol. 2, no. 3, pp. 196–217, 1998. 45, 70

[131] W. Mason and D. Watts, “Financial incentives and the performance of crowds,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 2, pp. 100–108, 2010. 46

[132] J. Horton, D. Rand, and R. Zeckhauser, "The online laboratory: Conducting experiments in a real labor market," *Experimental Economics*, vol. 14, no. 3, pp. 399–425, 2011. 46

[133] J. Downs, M. Holbrook, S. Sheng, and L. Cranor, "Are your participants gaming the system?: screening mechanical turk workers," in *Proceedings of the 28th international conference on Human factors in computing systems*, pp. 2399–2402, ACM, 2010. 46

[134] D. Oppenheimer, T. Meyvis, and N. Davidenko, "Instructional manipulation checks: Detecting satisficing to increase statistical power," *Journal of Experimental Social Psychology*, vol. 45, no. 4, pp. 867–872, 2009. 46

[135] G. Paolacci, J. Chandler, and P. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, 2010. 46

[136] R. Koenker, "Quantreg: quantile regression," *R package version*, vol. 4, 2009. 50

[137] J. Wooldridge, *Introductory econometrics: A modern approach*. South-Western Pub, 2009. 50

[138] B. Efron and R. Tibshirani, *An introduction to the bootstrap*, vol. 57. Chapman & Hall/CRC, 1993. 50, 78

[139] A. Martin, K. Quinn, and J. Park, "Mcmcpack: Markov chain monte carlo in r," *Journal of Statistical Software*, vol. 42, no. 9, pp. 1–21, 2011. 55

[140] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian data analysis*. CRC press, 2004. 58

[141] B. Efron, "Bayesian inference and the parametric bootstrap," 2011. 58

[142] T. Gilovich, "Biased evaluation and persistence in gambling.," *Journal of Personality and Social Psychology*, vol. 44, no. 6, p. 1110, 1983. 69

[143] G. Munro and P. Ditto, "Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information," *Personality and Social Psychology Bulletin*, vol. 23, no. 6, pp. 636–653, 1997. 69

[144] L. Ross, M. Lepper, and M. Hubbard, "Perseverance in self-perception and social perception: Biased attributional processes in the debriefing

paradigm.," *Journal of Personality and Social Psychology*, vol. 32, no. 5, p. 880, 1975. 69

[145] R. Wyer, D. Frey, *et al.*, "The effects of feedback about self and others on the recall and judgments of feedback-relevant information," *Journal of Experimental Social Psychology*, vol. 19, no. 6, pp. 540–559, 1983. 69

[146] M. Gorman, "How the possibility of error affects falsification on a task that models scientific problem solving," *British Journal of Psychology*, vol. 77, no. 1, pp. 85–96, 1986. 70, 74, 75

[147] A. Masnick and C. Zimmerman, "Evaluating scientific research in the context of prior belief: Hindsight bias or confirmation bias?," *Journal of Psychology of Science and Technology*, vol. 2, no. 1, pp. 29–36, 2009. 70

[148] B. Fischhoff, P. Slovic, and S. Lichtenstein, "Fault trees: Sensitivity of estimated failure probabilities to problem representation.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 4, no. 2, p. 330, 1978. 71

[149] J. Klayman and Y. Ha, "Confirmation, disconfirmation, and information in hypothesis testing.," *Psychological Review*, vol. 94, no. 2, p. 211, 1987. 73, 80, 93

[150] K. Dunbar, "How scientists really reason: Scientific reasoning in real-world laboratories," *The nature of insight*, vol. 396, 1995. 73, 95

[151] K. Dunbar, "Concept discovery in a scientific domain," *Cognitive Science*, vol. 17, no. 3, pp. 397–434, 1993. 73

[152] D. Klahr and K. Dunbar, "Dual space search during scientific reasoning," *Cognitive science*, vol. 12, no. 1, pp. 1–48, 1988. 73

[153] R. O'Connor, M. Doherty, R. Tweney, *et al.*, "The effects of system failure error on predictions," *Organizational Behavior and Human Decision Processes*, vol. 44, no. 1, pp. 1–11, 1989. 73

[154] J. Ioannidis and T. Trikalinos, "Early extreme contradictory estimates may appear in published research: The proteus phenomenon in molecular genetics research and randomized trials," *Journal of Clinical Epidemiology*, vol. 58, no. 6, pp. 543–549, 2005. 74

[155] P. Wason, "On the failure to eliminate hypotheses in a conceptual task," *Quarterly Journal of Experimental Psychology*, vol. 12, no. 3, pp. 129–140, 1960. 74, 94

[156] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*, vol. 1. Cambridge University Press New York, 2007. 78

[157] A. Gelman, Y. Su, M. Yajima, J. Hill, M. Pittau, J. Kerman, and T. Zheng, "arm: Data analysis using regression and multilevel/hierarchical models," *R package version*, pp. 1–3, 2010. 78

[158] D. Fudenberg and J. Tirole, "Game theory, 1991," 1991. 86

[159] B. Fischhoff, "Communicating uncertainty: Fulfilling the duty to inform," *Issues in Science and Technology*, vol. 28, no. 4, p. 63, 2012. 100

[160] K. Baggerly and K. Coombes, "Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology," *The Annals of Applied Statistics*, vol. 3, no. 4, pp. 1309–1334, 2009. 101

[161] M. Schneider and G. Sutcliffe, "Reasoning in the owl 2 full ontology language using first-order automated theorem proving," *Automated Deduction–CADE-23*, pp. 461–475, 2011. 106

[162] R. King, J. Rowland, S. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. Soldatova, *et al.*, "The automation of science," *Science*, vol. 324, no. 5923, pp. 85–89, 2009. 106

[163] C. Chamley, *Rational herds: Economic models of social learning.* Cambridge Univ Pr, 2004. 106, 108

[164] J. Zhao, V. Crupi, K. Tentori, B. Fitelsen, and D. Osherson, "Updating: Learning versus supposing," 2012. 107

[165] C. Howson and P. Urbach, *Scientific Reasoning: The Bayesian Approach.* Open Court Publishing Co, 1989. 108

[166] S. Suzuki, "The old evidence problem and agm theory," *Annals of the Japan Association for Philosophy of Science*, vol. 13, no. 2, pp. 105–126, 2005. 108

[167] C. Chihara, "Some problems for bayesian confirmation theory," *The British Journal for the Philosophy of Science*, vol. 38, no. 4, pp. 551–560, 1987. 108

[168] J. Woodcock, R. Temple, K. Midthun, D. Schultz, and S. Sundlof, "Fda senior management perspectives," *Clinical Trials*, vol. 2, no. 4, p. 373, 2005. 108

[169] M. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994. 109

[170] L. Von Ahn, "Games with a purpose," *Computer*, vol. 39, no. 6, pp. 92–94, 2006. 109

[171] V. Kundoor and A. Mueen, "Uncovering negative results: Introducing an open access journal: Journal of pharmaceutical negative results," *Journal of Young Pharmacists: JYP*, vol. 2, no. 4, p. 339, 2010. 110

[172] S. Reysen, "Publication of nonsignificant results: A survey of psychologists' opinions 1, 2," *Psychological reports*, vol. 98, no. 1, pp. 169–175, 2006. 111

[173] J. McDonald and U. of Delaware, *Handbook of biological statistics*, vol. 2. Sparky House Publishing Baltimore, MD, 2009. 111

[174] B. Dunning, "Defending the null hypothesis," 2011. 111

[175] R. Skloot, "Publication probity," *New York Times*, 2006. 111

[176] G. Koren and N. Klein, "Bias against negative studies in newspaper reports of medical research," *JAMA: the journal of the American Medical Association*, vol. 266, no. 13, pp. 1824–1826, 1991. 111

[177] A. Timmer, R. Hilsden, J. Cole, D. Hailey, and L. Sutherland, "Publication bias in gastroenterological research–a retrospective cohort study based on abstracts submitted to a scientific meeting," *BMC medical research methodology*, vol. 2, no. 1, p. 7, 2002. 111

[178] Anonymous, "No shame: The handling of results suggesting faster-than-light neutrinos was a model of fitting behaviour." 112