# Chapter 1

# Guiding Principles

> "But let a man venture into an unfamiliar field, or where his results are not continually checked by experience, and all history shows that the most masculine intellect will ofttimes lose his orientation and waste his efforts in directions which bring him no nearer to his goal, or even carry him entirely astray. He is like a ship in the open sea, with no one on board who understands the rules of navigation. And in such a case some general study of the guiding principles of reasoning would be sure to be found useful."
>
> —Charles Peirce, 1877, The Fixation of Belief (1)

Experiments frequently yield unexpected results. Learning the most from these results, and communicating them effectively, requires understanding what went wrong. Without a method that does this, disconfirming surprises will be seen as incomprehensible and diffuse, as indicated by Chapters Four and Five. Thus, a sound method of identifying the causes of error should both protect against distorted inferences and promote the sharing of valuable but disconfirming data.

The prescriptive analysis is divided into two parts. First, to learn from disconfirmation, one's hypothesis needs a well-defined *ceteris paribus* clause. If this is not done, then surprising data will be perceived as being caused by an experiment that could have yielded any results, thus making the data not worth sharing, as indicated by Chapter Four. Although logically valid, these uniform error models do not allow predictions that can help debug the experiment. The approach described in Chapter Seven gives form to error, so failed predictions can be pinpointed to specific causes; that is, error models can be made non-uniform. This is done by elaborating on five *guiding principles* extracted from four important philosophers of science (Popper (2), Kuhn (3), Lakatos (4), and Mayo (5)). They are:

1. Theories must be sufficiently axiomatized (Popper).

2. Tests must be conducted carefully and tenaciously, isolating sources of error (Mayo).

3. An interesting theory should be retained in the face of disconfirmations, as long as it can be modified to make interesting new predictions (Lakatos and Kuhn).

4. Commit to tracking and making probabilistic statements about "what went wrong." (Mayo and (6))

5. When a competitor theory clearly wins, it may be time to give up on it and engage in other pursuits (Lakatos and Kuhn).

This approach provides guidance on how to design and interpret experiments so that unexpected results are seen as informative rather than incomprehensible mere facts that turn into discarded anomalies. Like Duhem's Simplicism, the approach is "born and matured in the daily practice of science" (pg. 3) (7).

Next, Chapter Eight discusses methods for documenting and communicating data. Chapter Three proposed that that any pedagogical data sharing strategy relies on the sharer knowing a lot about what the learner knows. Ethical and practical rules for data sharing, discussed in Chapter Two, involve imposing minimal conventions on the reader. Descriptively, in the Wason rule discovery task of Chapter Five, participants' judgments of error were uncorrelated with actual error and less likely to be shared with another person when financial penalties were absent. The prescriptive approach described in Chapter Eight proposes a method of carefully documenting the conventions used in collecting and reporting data, allowing better communication and providing the required structure so penalties can be implemented.

# Chapter 2

# Breeding Orchids

> "The conduct of subtle experiments has much in common with the direction of a theatre performance," says Daniel Kahneman, a Nobel-prize winning psychologist at Princeton University in New Jersey. Trivial details such as the day of the week or the colour of a room could affect the results, and these subtleties never make it into methods sections. Bargh argues, for example, that Doyen's team exposed its volunteers to too many age-related words, which could have drawn their attention to the experiment's hidden purpose. In priming studies, "you must tweak the situation just so, to make the manipulation strong enough to work, but not salient enough to attract even a little attention", says Kahneman. "Bargh has a knack that not all of us have." Kahneman says that he attributes a special 'knack' only to those who have found an effect that has been reproduced in hundreds of experiments. Bargh says of his priming experiments that he "never wanted there to be some secret knowledge about how to make these effects happen. We've always tried to give that knowledge away but maybe we should specify more details about how to do these things."
>
> —Ed Yong, 2012 (8)

Data attributed to error are not shared with others, and these error attributions are unduly affected by whether results are affirming or disconfirming. Although Chapter Five found that financial penalties for incorrect error attributions helped participants identify error, it is not possible to provide these incentives in the real world, as error attributions are usually an implicit part of research. Additionally, Chapter Five also found that even when penalties promoted substantial accuracy of error attributions, data sharing policies were still inappropriately affected by whether the feedback was affirming or disconfirming. Furthermore, Chapter Four found that unexpected results are seen as caused by error, and the degree to which this error makes predictions from the same experiment seem uniform is positively associated with decisions to not share data. This chapter proposes methods of making precise, non-uniform error models, thus protecting data sharing policies by promoting accurate

perception of error.

As with all methodologies, there is no attempt at logical proof. Instead the approach analyzes important problems that lead to experimental surprises, and proposes methods to both avoid and learn from them. Most of the chapter has no empirical tests. However, to the greatest extent possible, examples are given by applying the proposed method to the methodology and results described in Chapter Five.

The approach outlined here blends work on scientific discovery by Herbert Simon and colleagues (9; 10; 11), work on causal induction by Griffiths, Tenenbaum and colleagues (12), work on noisy causal inference by Scheines and colleagues (13; 14), and work by various social science methodologists (Rosenthal and Rosnow (15); Shadish, Cook, and Campbell, (16); Luce and Narens (17)).

To explain the approach, consider the process of breeding orchids. Successful breeding requires a delicate balance of conditions, which may only occur with the care, cleverness, and patience of the breeder, in limited environments, and with specialized tools. Experimentation in the social sciences often follows a similar process. In order to study a phenomenon that interests us, investigators labor to find the delicate conditions under which it can be most reliably observed. They then conduct experiments manipulating theoretically interesting variables, within the constraints of that microcosm. This is the "direction of a theatre performance" articulated by Kahneman (8).

Discovering these delicate conditions can provide great benefits in the opportunities that it creates to replicate basic patterns at will, and to compare results observed in conditions that vary in controlled ways. However, it also creates the risks of limiting studies to an experimental monoculture, producing fragile results that disappear outside the "hothouse," or represent theories without clear predictions for more complex settings. Researchers concerned about these risks will invest in testing the boundary conditions for their prized results.

Within these carefully constructed normal science microcosms there is a partial solution to the problem of deciding whether to share ambiguous, disconfirming data with others: as normal science develops, data become unambiguously informative about the hypotheses of interest. Thus, the problem of data sharing is closely tied to the problem of managing experimental uncertainty.

Paul Meehl and his associates (e.g., (18; 19; 20)) laid the foundations on the process of experimentation, which they conceptualized as the repeated choice and generation of methodological and statistical tools to learn about the core and auxiliary hypotheses of a research program. A *core hypothesis* is a set of sentences in a suitable formal language, such as first-order logic, coupled with an ontological statement about the elements of the world and how they

are related (18; 21). An *auxiliary hypothesis* is a series of sentences, in that formalism, that is used in conjunction with the core hypothesis to derive its observable consequences (18).

The proposed framework develops the core and auxiliary hypotheses through a series of five testing stages, each developed to solve a specific epistemic problem. It is roughly patterned on that of multi-phasic medical clinical trials, whose structure is partially mandated by regulatory requirements, driven by the need to ensure that research results are characterized well enough to allow decisions about their application in matters of life and death.

- In Stage 1, a *representation* of the core and auxiliary hypotheses is formed out of one's entire corpus of knowledge.
- In Stage 2, *pretesting*, the auxiliary hypotheses needed to derive observable predictions from a core theory are tested directly, in ways that allow assessing and reducing their failure probabilities.
- In Stage 3, *pilot testing*, the core is assumed to be true and its predictions are tested along with the auxiliary hypotheses known to be imperfect.
- In Stage 4, *testing*, the auxiliary hypotheses are assumed to be true and different core hypotheses are tested against each other.
- Finally, in Stage 5, *evidence synthesis*, the cumulative weight of evidence is examined, and a decision is made to either return to stages 1-4 or terminate the endeavor.

I lay out these stages, discuss the problems in each stage, and propose methods to solve them.

# Part I

# Stage One: Representation

The first stage develops and represents the core hypothesis that will be tested along with the auxiliary hypotheses needed to test it. Rather than reasoning analytically from premises to conclusions (deduction), or inferring general principles from instances (induction), Stage One is *abductive*, generating plausible explanations for phenomena and formulating a precise *ceteris paribus* clause under which the proposed explanation will hold.

# Chapter 3

# Core Hypotheses

Griffiths and Tenenbaum's Theory-Based Causal Induction (12) provides a method for developing inductive causal theories with structured background knowledge. The approach works by specifying an *ontology*, a set of *plausible relations* between ontological elements, and precise mathematical *functional forms* that these plausible relations can take, either deterministically or stochastically. These three elements are all formalized as structured prior knowledge (what Levi (22) calls the corpus of knowledge), making Bayesian computation possible.

## 3.1 Ontology

An ontology is the skeleton of any theory. It is an explicit commitment to the entities that exist in the theory, their properties, and how these different entities (types) relate to one another (23). The ontology is usually specified in one of a number of ways, including mereology (part-whole relationships) and topology (connectedness relationships) (24). For example, one theory may involve the amygdala as a part of the brain (a mereological specification), which is physically connected to the insula (a topological specification).

Theory-Based Causal Induction (12) uses a simplified version of an ontology. It involves entities or *types*, each with their own specific properties. For example, in Newton's mechanics, the relevant types were mass, velocity and acceleration. The ontology also specifies how many of each type we are considering, or a stochastic distribution of the number of entities of each type. For example, we may examine two masses colliding (a fixed number) or an unknown or variable number of masses colliding (stochastic). The next element of the ontology is a set of *predicates* that describe the possible causal relationships between the types, and range of values that each predicate can take (e.g., Boolean, continuous, etc.). For example, if mass A interacts with mass B through the collision predicate, then the velocity and acceleration of both masses change.

If no predicate relates two entities, then they cannot be directly causally connected.

## 3.2 Plausible Relations

Next are *plausible relations*. If types are connected by predicates then they are related in some way, but some relationships may be more plausible than others. For example, "a lamp is more likely than a fan to produce a spot of light, that a fan is more likely than a tuning fork to blow out a candle, and that a tuning fork is more likely than a lamp to produce resonance in a box" (pg. 664) (12). The plausibility of the relationship is likely closely related to the presence and strength of mechanistic explanations of how a cause can produce an effect (25). The set of plausible relationships can be represented as *Directed Acyclic Graphs* (DAGs), or causal graphical models with typed variables as vertices (14; 26).

## 3.3 Functional Form

Finally, a functional form specifies the exact mathematical relationship between cause and effect described by the plausible relation. For example, in the case of Newton's mechanics, acceleration, mass and force are related by a specific functional form: $F = ma$. If outcomes are stochastic, the functional form is a conditional probability density function. A very simple case, the noisy-OR function, proposes that the probability of an effect is merely the weighted independent sum of its causes:

$$P(effect) = w_0 + w_1 cause_1 + w_2 cause_2 + ... + w_n cause_n \tag{3.1}$$

In this function, $w_i$ is the strength of each cause $i$ to produce the effect. In more complex problems, the functional form may be as complex as a system of partial differential equations.

## 3.4 Example Gene Expression Experiment

Griffiths and Tenenbaum provide an example of the Theory-Based Causal Induction approach using a gene expression experiment with lab mice. The ontology, plausible relations, and functional form are summarized in the table below.

The plausible relation is that an injection of chemical $C$ into mouse $M$ will lead to an expression of gene $G$ in mouse $M$, and this relationship is expected to be true for all mice $M$ with probability $p$ for each $C, G$ pair:

$$Injected(\text{C, M}) \rightarrow Expressed(\text{G, M}), \text{p } \forall\{C, G\} \tag{3.2}$$

Ontology

| Type | Number | Predicates | Values |
|---|---|---|---|
| Chemical | $N_c \sim P_c$ | *Injected*(Chemical, Mouse) | Boolean: {T, F} |
| Gene | $N_g \sim P_g$ | *Expressed*(Gene, Mouse) | Boolean: {T, F} |
| Mouse | $N_m \sim P_m$ | | |

The functional forms specify that the injection of mouse $M$ with chemical $C$ is an exogenous event, determined by a coin flip with known bias (i.e., randomization):

$$\textit{Injected}(\text{C, M}) \sim \text{Bernoulli}(.) \tag{3.3}$$

The expression of gene $G$ in mouse $M$, on the other hand, is modeled using the noisy-OR function, where the probability of expression is modeled as a coin flip with bias determined by $v$ which is in turn determined both by the base rate of gene expression ($w_0$) and whether the mouse was injected with chemical $C$:

$$\textit{Expressed}(\text{G, M}) \sim \text{Bernoulli}(v) \tag{3.4}$$

$$v = w_0 + w_1 \textit{Injected}(\text{C, M}) \tag{3.5}$$

*Wason Task.* Take the Wason task, described in Chapter Five, as a second example. The core hypothesis in these experiments was that feedback attributed to error would be less likely to be shared.

Formulating the core hypothesis uses a many-sorted logic, meaning there are different *sorts* or *types* of things. What are these types? There are only two types of things participants (P) and triples (TR). Each type is related through multiple predicates:

- *Feedback*(P, TR): A participant (P) receives feedback (F) on a triple (TR) with boolean values ({FIT, DNF}).
- *Attributes*(P, F): A participant (P) attributes feedback (F) to error with boolean values ({Error, Not Error}).
- *Share*(P, TR): A participant (P) shares a triple (TR) with boolean values ({Share, No Share}).

Using this typed logic allows us to preclude impossible statements, such as attributing a participant to error. The ontology is summarized in the table below.

Based on the core hypothesis, the first plausible relation is that the feedback affects the error attribution for each participant $P$ and each triple $TR$ with probability $p_1$:

$$\textit{Feedback}(\text{P, TR}) \rightarrow \textit{Attributes}(\text{P, F}),\ p_1 \forall \{P, TR\} \tag{3.6}$$

| Ontology | | | |
|---|---|---|---|
| Type | Number | Predicates | Values |
| Triple | $N_t \sim P_t$ | *Feedback*(P, TR) | {FIT, DNF} |
| | | *Share*(P, TR) | {Share, No Share} |
| Participant | $N_p \sim P_p$ | *Attributes*(P, F) | {Error, Not Error} |

The second plausible relation is that the error attribution affects data sharing for each participant $P$ and each triple $TR$ with probability $p_2$:

$$\textit{Attributes}(\text{P, F}) \rightarrow \textit{Share}(\text{P, TR}),\ p_2 \forall \{P, TR\} \tag{3.7}$$

The DAG (27; 28; 29) for these plausbile relations is the structural model that the experimenter (me) is trying to learn:

Figure 3.1: DAG of a core hypothesis for the Wason task. $TR$ is a triple, $F$ is feedback, $\epsilon$ is exogenous error, $ATT$ is the attribution of the trial to error, and $SH$ the sharing decision. The two plates show that judgments are specific to a triple, and triples are specific to a participant.

The functional forms are as follows. First the feedback $F$ is jointly determined by the triple ($TR$) and error ($\epsilon$). The double box around this node indicates that it is completely determined by its parents:

$$\text{F}= \left\{ \begin{array}{ll} FIT & \text{if } \epsilon = F \text{ and } TR = FIT \text{ or } \epsilon = T \text{ and } TR = DNF \\ DNF & \text{if } \epsilon = T \text{ and } TR = FIT \text{ or } \epsilon = F \text{ and } TR = DNF \end{array} \right\}$$

$$\epsilon \sim \text{Bernoulli}(0.2) \tag{3.8}$$

The attribution of the feedback to error depends on the feedback, as indicated by the following plausible relation shown in the DAG:

$$\textit{Feedback}(\text{P, TR}) \rightarrow \textit{Attributes}(\text{P, F}) \tag{3.9}$$

Thus, the attribution is a bernoulli random variable:

$$\textit{Attributes}(\text{P, F}) \sim \text{Bernoulli}(\alpha_{p,t}) \tag{3.10}$$

The parameter $\alpha_{p,t}$ determines the tendency to attribute feedback to error, and can be estimated from the data. The example below shows the functional form of $\alpha_{p,t}$ using the logistic likelihood function. The tendency to attribute feedback to error is a function of the feedback ($\beta(F_t)$) and a subject-level

intercept ($\alpha_p$) to account for the fact that some participants are more likely to make (unconditional) error attributions than others:

$$L(\alpha_{p,t}) = \frac{1}{1 + e^{\beta(F_t)+\alpha_p}} \tag{3.11}$$

The data sharing judgment follows a similar pattern, but is instead jointly determined by both feedback and attribution, as indicated by the following plausible relations shown in the DAG:

$$\textit{Attributes}(\text{P, F}) \rightarrow \textit{Share}(\text{P, TR}) \tag{3.12}$$

And:

$$\textit{Feedback}(\text{P, TR}) \rightarrow \textit{Share}(\text{P, TR}) \tag{3.13}$$

Again, the sharing judgment is a bernoulli random variable:

$$\textit{Share}(\text{P, TR}) \sim \text{Bernoulli}(\sigma_{p,t}) \tag{3.14}$$

The parameter $\sigma_{p,t}$ determines the tendency to share data, and can also be estimated from the data. The example shows the functional form of $\sigma_{p,t}$, again using the logistic likelihood function. The tendency to share trials is a function of the feedback ($\beta_{sh}(F_t)$), a subject-level intercept ($\sigma_p$) to account for participant-level willingness to share, and whether the feedback was attributed to error ($\alpha_{sh}$):

$$L(\sigma_{p,t}) = \frac{1}{1 + e^{\beta_{sh}(F_t)+\sigma_p+\alpha_{sh}}} \tag{3.15}$$

This concludes the formal representation of the core hypothesis. Using this approach as a guide can help researchers develop suitably formalized psychological theories, as Popper required (2). Using a formal representation is especially helpful in dealing with the problem of changing definitions in response to refutation and blaming a theoretician for misinterpreting a theory, Popper's second and fourth conventionalist stratagems. If definitions are clearly specified ahead of time in the ontology, then this stratagem requires more rigorous defense. If our theory is clearly laid out in a standard format, then we guard ourselves against being attacked for theoretical misinterpretation or incoherence. Some of this may already be done implicitly by social scientists (which Griffiths and Tenenbaum provide evidence of). It is also important to note that formalizing a theory in this manner should come after, not precede, the content of the theory. That is, the formalization can "serve as a means of precise exposition, but not as a guarantee of soundness for the conceptions incorporated in the axiomatized theory" (pg. 250) (30).

# Chapter 4

# Auxiliary Hypotheses

The Theory-Based Causal Induction approach provides a framework for developing hypotheses. However, this is not enough for those who not only specify theories, but also experimentally test them. To test any theory, an experimenter must make a variety of assumptions about the experimental test, usually implicit, to fairly test the theory. This *ceteris paribus* clause is a conjunction of auxiliary hypotheses that allow the core hypothesis to make well-defined predictions.

Unfortunately, there is no general solution to the problem of listing all the relevant auxiliary hypotheses: we have always forgotten something. Luckily, there are types of auxiliary hypotheses that are used repeatedly, so it is helpful to specify them for every project so they do not have to be reconceptualized from scratched for each experiment. There are also some auxiliary hypotheses that are common in social science research.

## 4.1 General Auxiliary Hypotheses

The general auxiliary hypotheses proposed here are ones that experimenters, of any discipline, encounter. This typology was designed to include Meehl's (18) instrumental auxiliaries, theoretical auxiliaries, and experimental auxiliaries, and Mayo's (5) experimental models, and data models.

There are five general auxiliary hypotheses:

1. *Ontological*: Have we omitted any element from our ontology?
2. *Choice of evidence*: Do we have all the relevant evidence?
3. *Interpretation of evidence*: Are the theories that we use to interpret the evidence accurate?
4. *Instrumental*: Do our instruments work as intended?

5. *Experimental* Is the experiment free from confound?

### 4.1.1 Ontological

The first general auxiliary hypothesis is that the ontology we've specified in the Theory Based Causal Induction "core" is correct. Predictions may fail because we've left out an important entity from our ontology. For example, Newton's first law, that an object will maintain constant velocity unless a force acts upon it, would ostensibly be violated if one omitted atmospheric friction as a force. This is an ontological problem, closely related to, but more general than, omitted variables. This is also called problem framing (31; 32), where one must consider whether one has included all the relevant possibilities, hypotheses, and model structures when quantifying uncertainty. It has been speculated that severe theoretical failures are due to incorrectly specified ontologies rather than incorrect functional forms (33). The ontological auxiliary hypothesis is that we've included the relevant entities and predicates in our ontology, or that those that are omitted do not affect our inferences.

### 4.1.2 Choice of Evidence

The second general auxiliary hypothesis is that, when we've formed our theory and experiment, we've included all relevant evidence and nothing more (31; 34). When mustering evidence in support of or against the core hypothesis, we can consider a wide variety of evidence varying in relevance and quality (EPA, 2009) (35). We can include direct empirical evidence, such as direct measurement of the phenomenon we are interested in (e.g., a Randomized Controlled Trial; a laboratory experiment). We can also include semi-empirical evidence, such as measurement of a phenomenon but under different conditions that desired. We can also use data from variables that we think are related to entities we are interested in. If no empirical evidence is available, we can use theory to fill in the gaps. Finally, if there is no empirical evidence and no relevant theory, we can submit our own insight and opinions as evidence (e.g., a thought-experiment; guesswork).

With a well-defined theory, one can look at each entity, plausible relation, and functional form, and categorize the evidential support, or lack thereof, for each element (36). For example, we might not have an exact value (empirically verified) for the mass of a billiard ball that is involved in a collision, but may use theory about the size of the ball and the density of its material constituents, to make an approximation. When evidence is missing, this should be carefully noted.

### 4.1.3 Interpretation of Evidence

The meaning of any observation is not a self-evident truth or axiom. Instead, every observation requires an interpretive theory of evidence (4), or theoretical auxiliary hypothesis (18). For example, a theory of optics is required to interpret evidence using a microscope. Theoretical auxiliaries build on the results and derivations of other theories and evidence, hence depend on the strength of that science. For example, the use of an aggression scale to test a theory of the relationship between aggression and organizational climate has the theoretical auxiliary hypothesis that the scale measures aggression. If one fails repeatedly to find a predicted relationship between aggression and climate, it might mean that the prediction is wrong or that the scale is invalid. Any psychometric theory would be an auxiliary of interpretation.

### 4.1.4 Instrumentation

Auxiliary hypotheses of instrumentation are "the accepted theory of devices of control (such as holding a stimulus variable constant, manipulating its values, or isolating the system with, e.g., a soundproof box or white-noise masking generator, or of observation" (pg. 110) (18). The instrumental auxiliaries commonly used in psychology, for example, are computers for online surveys and stimulus presentation, pencils. A failed computer, broken pencil, ripped survey, a typographical error in instructions *etcetera* would be failed auxiliary hypotheses of instrumentation. While the meaning attributed to observations derived from these instruments depends on the auxiliary hypotheses of interpretation, the actual process that was implemented is one of instrumentation.

### 4.1.5 Experimentation

Finally, auxiliary hypotheses of experimentation concern the internal validity of the experiment or the "experimentally realized conditions" (18). Examples include the assumption that volunteers and non-volunteers for the experiment do not systematically differ in their responsiveness to experimental treatments (volunteer bias; (37)), that the task is not too cognitively demanding ((38)), and that the stimuli used in the experiment capture the important elements of the constructs they represent (stimulus sampling, (39)). These can also include issues in Mayo's experimental models and data models, including sample size and test statistics, protocols, descriptions of materials, and how they are related to possible errors.

## 4.2 Specific Auxiliary Hypotheses

There are also six specific auxiliary hypotheses that social science researchers typically invoke when designing and interpreting experimental evidence. Some combination of these auxiliary hypotheses are usually discussed in separate textbooks on survey design, research methodology, or psychometrics. The classification proposed here unifies them with a specific purpose as auxiliary hypotheses, rather than just topics of study in their own right. I categorize them by the acronym MIMECC:

1. M*otivation*: Are participants motivated to behave accurately?
2. I*nternal*: Are causal inferences free from confounds?
3. M*easurement*: Do measurements meet required assumptions?
4. E*xternal*: Do inferences in the sample generalize to the population?
5. C*onstruct*: Are the concepts used valid?
6. C*ommunication*: Do participants and researcher agree on what is expected?

Motivation is usually addressed by survey researchers to get participants to respond to survey requests (40). Internal validity, external validity, and construct validity are canonical parts of introductory social science research methods (16). Measurement is usually addressed as a topic in psychometrics (41; 42). Finally, communication is usually addressed as a separate topic related to risk communication and survey research (43; 44).

### 4.2.1 Motivation

The first specific auxiliary hypothesis is motivation. Even with a perfectly logical experiment, if participants don't care about the task then it is not possible to get good data from them. Motivating participants can be achieved by increasing the benefits of participation, decreasing the costs of participating, establishing trust with the researchers (40), and giving participants an incentive to respond carefully and truthfully (incentive compatibility or mechanism design) (45). Importantly, tasks that participants find fun are likely to result in high quality, engaged responses (46). For example, Fold-it (`http://fold.it/portal/`) is a protein folding game that provides very high quality data, from motivated and careful participants, that could be used to discover human discovery strategies (47). These are sometimes called serious games (48; 49; 50; 51), gamification (52; 53), or games with a purpose (46). See Dillman (40) for more on motivating participants in surveys and Pink (54) for motivation more generally.

- Benefits: Are the benefits enough to motivate participants?
- Costs: Are the costs of participation so high as to demotivate participants?
- Trust: Do the participants trust the researchers?
- Incentive Compatibility: Are participants incentivized to respond truthfully?
- Fun: Do participants enjoy performing the task?

*Wason Task.* In the rule-discovery task of Chapter Five, participants were offered $5 (Benefits) for 30 minutes of their time (Costs). They were given an informed consent document that specified the research was from a university (Trust). Participants were not offered money for accurate answers until Experiment Three (Incentive Compatibility). No assessment was made as to whether the task was fun (Fun).

### 4.2.2 Internal

The second category of specific auxiliary hypothesis is internal validity. This is by far the largest category, and the most extensively studied. Whenever an experiment is conducted, one conjectures the auxiliary hypothesis that no other factors aside from our intervention vary between the experimental and control groups. Random assignment is the first step in doing this: those who are randomly assigned to control and treatment group are expected to be equivalent, in the long run, on every variable they could differ on. However, if the participants know the condition they are in, or know the hypothesis of the experiment, then they would differ. As a result it is important that participants and researchers are blinded to both the condition the participant is in and the hypothesis of the experiment.

Another severe problem of internal validity is incomplete outcome data. Some participants may not complete the experiment. If the process that leads to missing data is random, then participants in the treatment and control group will, in the long run, not differ on other variables. However, if there is systematic tendency for some participants to drop out in a way that is related to the effect one is trying to measure, then bias will result.

It is also possible that experimental interventions have subtle side-effects. For example, the mere novelty of the intervention (i.e., putting the participant in a new situation) or knowledge that the participant is in a study (the Hawthorne effect), can cause differences between groups. See Shadish, Cook and Campbell (16) for more on threats to internal validity.

- Assignment: Was the assignment to condition adequately (randomly) generated?
- Condition Blinding: Was the method of condition assignment adequately concealed?
- Hypothesis Blinding: Was unnecessary knowledge of the assigned condition adequately prevented during the study?
- Incomplete Outcome Data: Were incomplete outcome data adequately addressed?
- Researcher Expectancies: Has the researcher intentionally or unintentionally influenced the participants?
- Novelty: Is the treatment intrusive and novel?
- Disruption: Is the treatment disruptive?
- Compensatory Rivalry: Does the control group know of the treatment group and try to outperform them?
- Resentful Demoralization: Does the control group know of the treatment group and perform worse as a result?
- Treatment Diffusion: Is the control group partially or fully exposed to the treatment?
- Instrumentation: Are measurement instruments stable over time?
- Testing: Does repeated measurement affect the measurements?
- Instability: Is the measurement process reliable over time?
- Selection: Are there systematic differences in respondent characteristics between groups?
- History: Has an event occurred between treatment and measurement?
- Maturation: Are results affected by the procession of time, such as boredom?
- Regression: Are units selected for extremeness and thus regress to their mean?
- Attrition: Have respondents been lost to treatment or measurement?
- Selection-Interaction: Are participants in different treatment groups differentially exposed to internal validity threats?

- Testing Interactions: Are there reactive measures wehre asking a question changes the response to the question?
- Identification: Does each causal factor lead to a different distribution of data?

*Wason Task.* In the rule-discovery task of Chapter Five, random assignment was automated by Qualtrics (Assignment) and this automation prevented me from knowing which participant received what treatment (Condition Blinding). I was not blinded to the hypotheses, and it is unclear whether the participants were (Hypothesis Blinding). Some participants did not complete the task (Incomplete Outcome Data and Attrition). It is possible that the development of the materials in the study unintentially conveyed the purpose of the study to participants or that expectancies affected statistical analyses, as they were not blinded (Researcher Expectancies). It is not clear how novelty or disruption effects would appply in this context (Novelty/Disruption). Although participants were blinded to alternative conditions, they were not blinded to their own conditions (in Experiments Two and Three). If they somehow were told by a friend the contents of the task, then several threats are possible (Treatment Diffusion, Resentful Demoralization, and Compensatory Rivalry). The instruments were computerized so should not have degraded, unless I made a programming mistake (Instrumentation). Having participants make error attributions may subsequently affect their response to data sharing decisions, where if the participant had not explicitly made the attribution there may have been no such association (Testing). The measurement process was stable over time as it was completely computerized, unless Qualtrics broke (Instability). Participants were randomly assigned so there would no be selection (Selection). I know of no events that occurred either after the beginning of the entire research programme or after each participant began the survey itself (History). Participants could get tired, frustrated, or bored with the task, as it was long and difficult (Maturation). Participants were not selected for extremeness (Regression).

### 4.2.3 Measurement

Any empirical study involves measurement, and the underlying formal structure of these measurements determines the type of inferences that can be made from them. Thus, auxiliary hypotheses of measurement propose that the mathematical assumptions needed to represent our data are correct. The important elements of this auxiliary hypothesis are representation, uniqueness, meaningfulness, and scaling.

Representation specifies the conditions under which we can create a mathematical (numerical) model of our measurements. For example, two bushels

of wheat added to three bushels of wheat results in five bushels of wheat in exactly the same way that $2 + 3 = 5$ (41). In this case, the arithmetic operator addition holds true for bushels of wheat. If empirical relationships hold in a manner that is identical to a set of numerical relationships, then the latter is a mathematical representation of the former. This is called an isomorphism.

Uniqueness is the degree to which the numbers we assign to observations can be exchanged for other numbers while preserving the properties of the representation. For example, a transitive (ordered) scale is preserved under any monotonic function. Interval scales are unique only up to an affine transformation, which preserves both order and distances between items in the ordered sets. The set of transformations that preserve the relational properties of the system (e.g., ordering, fixed differences) are the set of admissible transformations for the measurement system (42).

Meaningfulness delimits the valid assertions that can be made based on the representation and uniqueness of a measurement. The conclusions we draw should not change if we make an admissible transformation to our numbers. Any statement that does not change based on admissible transformations is meaningful; those that do change are not meaningful.

Finally, Scaling acts about the practical transformation of measurements into numerical scales, with possible errors. DeVellis (55) argues that items sharing a common cause are a scale, if they share a common consequence they are an index and if they are just part of a superordinate category then they are an emergent variable.

- Representation: What are the relational properties of the measurement system (e.g., transitivity, completeness)?
- Uniqueness: What are the set of homomorpisms or isomorphisms equivalent to our measurement system?
- Meaningfulness: What transformations can we apply to our data while preserving their meaning?
- Reliability: Are measurements test-retest, inter-rater, and internally reliable?
- Stability: Are measurements stable to split-half, parallel forms, and alternative-forms reliability?

*Wason Task.* In the rule-discovery task of Chapter Five, error attributions, feedback, and data sharing decisions were all nominal variables, whereas probability judgments were supposed to be absolute (Representation). Any one-to-one transformation is admissible for a nominal variable (42), but there are no admissible transformations to probability judgments if they are interpreted as

absolute probabilities (Uniqueness, Meaningfulness). There was no attempt at measuring reliability (test-retest, interrater, internal) and stability (split-half, parallel forms, alternative forms) of measurement instruments (Reliability and Stability).

### 4.2.4 External Validity

External validity specifies the difference between the experiment we've conducted in its ideal form and the real world circumstance to which we intend to generalize. External validity is important because if we choose an experimental situation that differs from the world we are trying to extrapolate to, then we may find no effect in the lab when there is an effect in the real world, or vice versa. To help with this, we'd like to randomly sample from the population we are interested in extrapolating to, giving them the same intervention that they would receive in the real world, and measuring the actual outcome of interest, rather than a proxy or surrogate. See Turner *et al.* (56) for a careful definition of external validity. For issues of recruitment (57; 58; 59; 60).

- Population: Are study subjects in the idealized study drawn from a population identical to the target population with respect to age, sex, etc?
- Intervention: Is the intervention in the idealized study identical to the intervention used?
- Control: Is the control group in the idealized study the same as the control group used?
- Outcome: Is the study outcome the same as the idealized study outcome?

*Wason Task.* In the rule-discovery task of Chapter Five, participants in the study were either CMU students or MTurk participants, and thus were not drawn from the population of interest (working scientists; Population). The interventions used in the study were not similar to that experienced by real scientists, either in terms of incentives ($5 or $100 is nothing like tenure or a pharma job) or penalties (earning less than $5 is nothing like being fired, e.g., Stapel) (Intervention). The data sharing outcome was nothing like a publication decision (Outcome).

### 4.2.5 Construct Validity

Construct validity is probably the most difficult to understand and vaguely defined auxiliary hypothesis in all of social science. Construct validity is a statement about an observed pattern of correlations in data and unobserved

causes or latent variables proposed by our theory (61). This is different from criterion validity which relates a pattern of correlations among observed or operationally defined measurements. It also differs from content validity because "no criterion or universe of content is accepted as entirely adequate to define the quality to be measured" (pg. 282) (61).

Construct validity is epistemic and empirical rather than ontological and theoretical (21). If our measures do not correlate with what they are supposed to they do not have convergent validity. If they are correlated with things they aren't supposed to be related to then they do not have discriminant validity. If either convergent or discriminant validity are violated then construct validity is poor.

- Convergent: Does the construct correlate well with other constructs it should be related to?
- Discriminant: Is the construct uncorrelated with constructs it shouldn't be related to?
- Necessity: Are there any dimensions contained in the constructs that are not contained in the measures?
- Sufficiency: Are there any dimensions contained in the measures that are not contained in the construct?
- Construct Stability: Does the dimensionality of the measures change across treatment conditions?
- Method Stability: Would the fit of the measures to the construct change if conducted using a different method (format)?
- Level Stability: Is the level of the treatment administered large enough to produce an effect?
- Process Accuracy: When responding to the measures, are participants using the same process that the construct supposes?

*Wason Task.* In the rule-discovery task of Chapter Five, the constructs of error attribution and data sharing were not validated. There was no attempt to correlate them with other judgments that should be related (Convergent Validity) or with other judgments that shouldn't (Discriminant Validity).

### 4.2.6 Communication

The most well understood but overlooked auxiliary hypothesis is communication: that the participant and experimenter agree on the meaning of the

experimental instructions and expected behaviors of the participant. This can fail in a number of ways. If instructions are written or read aloud, participants may not have the cognitive capability or reading ability to comprehend the instructions. If they are asked to perform a task that is complex, they may not be able to integrate and use the information provided in the task as required. Participants may gloss over key instructions if they are not made salient and the participant's attention is not drawn to them. They also may not encode instructions even if they briefly attend to them.

There are also problems with the language games played whenever natural language communication is involved. If participants have only basic knowledge or values, they will construct responses to specific questions that don't reflect what they actually believe or want (62). If excess information is conveyed, participants may feel that they are supposed to use the information, and if too little information is conveyed, they may try to fill in the gaps with what they think the researcher wants (44). If participants are deceived they may not find the researcher trustworthy and not attend to information or use it in the expected manner. See Fischhoff, Brewer, and Downs (43) for guidance on drafting high quality communications.

- Comprehension: What is the reading level of the instructions? What is the propositional content of the instructions?
- Attention: Do participants perceive the relevant communications? Do they register relevant information in long-term memory?
- Use: Do participants understand how to act on communicated information? Do they use prospective memory effectively, using information when they need to?
- Quantity (a): Is to little information communicated to participants?
- Quantity (b): Is too much information communicated to participants?
- Quality: Is false or insubstantiated information conveyed to participants?
- Manner: Is information commununicated to participants ambiguous, prolix, or disorganized?

*Wason Task.* In the rule-discovery task of Chapter Five, the instructions were all modified to be at a 6th to 8th grade reading level (Comprehension). Questions were asked both during and at the end of the task to make sure they understood and attended to the instructions (Attention), and whether they understood how to use the information (Use). Attempts were made to avoid deceiving or omitting any information (Quantity a,b; Quality), however too little information was provided about the probability judgments (Quantity a).

Information was communicated in a concise way, although some participants found it sterile (Manner).

### 4.2.7 Paradigmatic Auxiliary Hypotheses

Finally, there have two auxiliary hypotheses that are necessary for the construction of a paradigm: 1) sparsity of effects and 2) ideal interventions.

### 4.2.8 Sparsity of Effects

Any paradigm involves a fixed set of background factors that do not vary, and a set of interventions or experimental factors that do (63). The auxiliary hypotheses of experimentation discussed previously deals with the possibility that the factor that varies is not the only one that varies between treatment and control group. However, the sparsity of effects assumption is the complement to this: that the background set of factors that do not vary have no special effect on the outcome of the experiment. This is the assumption that we haven't created a syzygy, where the effect of our experimental manipulation is really the product of particular quirks of the paradigm, quirks which are not part of the theory we are using that interact with the experimental manipulation. Sparsity of effects for paradigms assumes that items not manipulated are not causing higher order interactions.

*Wason Task.* In the rule-discovery task of Chapter Five, the sparsity of effects auxiliary hypothesis is that items not manipulated are not causing higher order interactions, for example that the labeling of FIT/DNF is not causing higher order interactions with other variables, such as tendencies to choose triples that are expected to be affirming or disconfirming.

### 4.2.9 Ideal Intervention

Finally, experimental manipulations are formally entailed by changes to causal graphical models (26). For example, if we want to determine whether B has a causal effect on A, then if we manipulate B directly, all possible causes of B are eliminated, rendering it exogenous. This auxiliary hypothesis is that the manipulation of B is ideal, in the sense that if we try to change the state of B it responds perfectly, not stochastically, and that our intervention does not affect any other factors involved in the system we are investigating. However, there are a number of ways interventions can go afoul. For example, an intervention may be noisy, and only produce the desired effect some of the time. An intervention may also cause other effects than just the one we desire. Scheines (13) shows that choosing an ideal intervention is the same as choosing an instrumental variable. This can be seen as the construct validity of the intervention and is typically validated using a manipulation check.

*Wason Task.* In the rule-discovery task of Chapter Five, the interventions were the feedback, incentives and penalties. No manipulation checks were used for the feedback or penalties, but there was a manipulation check for the incentives using an open-ended response.

### 4.2.10 Creating and Representing a Paradigm

Stage One lays the formal foundation of core and auxiliary hypotheses needed for theory testing. From the above analyses we have a formalized theory along with auxiliary hypotheses that are nominally classified as being relevant or irrelevant to the task. Once these are carefully articulated, it is exciting to get to the critical experiment. However, failed predictions from this premature experiment usually indicate that Stage One provides nowhere near the necessary preparation to perform normal science. We don't know if the auxiliary hypotheses are actually relevant, how likely they are to fail, and what failure effects they may have. While Stage One helps protect against the need for invoking auxiliary hypotheses ad-hoc, or misusing a theory (Popper's first and fourth conventionalist stratagems), more is needed.

Researchers are all familiar with the concepts of pretests, pilot tests, and experimental tests. However, the distinctions among them are not typically made sharply in their training nor documented systematically in research reports. As a result, there is greater risk of a fortuitous result from a pretest being "promoted" to the status of an experimental test or, conversely, an experimental test being "demoted" to a pilot test when it produces unexpected (or unwanted) results. In order to take full advantage of pleasant or unpleasant surprises, experimenters need a systematic empirical approach to dealing with auxiliary hypotheses, so that their validity is neither over- nor understated. To do this, we clarify the following four stages of testing, each with their intended purpose and assumptions.

# Part II

# Stage Two: Pre-testing

The second stage, *pre-testing*, addresses the problem of getting information from experiments when we are uncertain both about whether our core hypothesis is true or false and whether our experimental design satisfies the necessary auxiliary hypotheses we've laid out. For example, most experiments make the auxiliary hypothesis that participants understand the instructions provided to them. A pre-test would propose these instructions then follow with a quiz to test comprehension. If participants can successfully introspect about their understanding, there should be predictable differences in responses among those who do and do not interpret the instructions as intended. Testing that assumption about introspection regarding the instructions requires a separate experiment, with its own complications, perhaps reduced by the strength of the general science regarding those issues.

Thus, the goal of pretesting is to collect data to assess and minimize the risk of failed auxiliary hypotheses. It helps solve the problem of choosing the experimental design that yields true auxiliary hypotheses (e.g., the participant understands the instruction), or minimizes the chance that the auxiliary hypothesis will be false.

The proposed method works as follows. Suppose one is concerned that a participant does not understand some concept communicated in the instructions. By performing a cognitive interview, or giving the participant a quiz, one can estimate the individual-level failure probability of that communication (64).

*Wason Task.* Here is an example taken from my pretesting of the Wason rule-discovery task used in Chapter Five. First, I considered the specific assumptions required to test my core hypothesis using *pre-posterior analysis*. Pre-posterior analysis is an important kind of suppositional reasoning (65), where one supposes that an outcome occurred and then entertains the set of serious possible causes of that outcome. In this way, one pre-empts the data by making causal attributions beforehand. The intent is to minimize the regret of not having considered a threat to the validity of our experiment beforehand.

By imagining cases where I got unexpected results, I came up with 16 auxiliary hypotheses (listed below) that I felt were necessary to give a proper test to my core hypothesis. As can be seen, they all focus on the comprehension auxiliary hypothesis:

1. Participant does not understand how the error works.
2. Participant doesn't understand the Actual Rule concept.
3. Participant doesn't understand the triple concept.
4. Participant doesn't understand the Your Rule concept.
5. Participant doesn't understand their task in general.
6. Participant doesn't understand what FIT/DNF means.

7. Participant doesn't understand what the feedback means.
8. Participant doesn't understand what it means to create a new triple to test Your Rule.
9. Participant doesn't have a hypothesis that they believe. Participants don't know how to confirm or disconfirm their hypothesis.
10. Participant doesn't understand how to change the error attributions.
11. Participant doesn't understand how to record error on the spreadsheet.
12. Participant doesn't understand how to use the spreadsheet.
13. Participant doesn't understand that evidence is disconfirming or confirming.
14. Participant doesn't understand that they only get one guess.
15. Participant doesn't understand the guessing.
16. Participant doesn't understand what the error attribution task is.

Cognitive interviews were then conducted to examine whether these auxiliary hypotheses were satisfied by the experimental design. This was done using think-aloud protocols, where a participant is asked to "think-aloud" while interpreting the instructions. They were also asked to respond to retrospective probes (66; 67). The participants were sixteen members of the Pittsburgh community recruited through a web advertisement through the Center for Behavioral Decision Research. They were paid $5 for 30 minutes of their time.

The list below shows the retrospective probes that were intended to examine the participants' understanding of the task. These were asked after the participant completed the think-aloud portion of the interview.

1. Can you explain, in your own words, what the "Error" is and its purpose in the task?
2. On each trial, how likely is it that an error will occur?
3. Can you explain, in your own words, what the "Actual Rule" is and its purpose in the task?
4. Can you explain, in your own words, what the "triple" is and its purpose in the task?
5. Can you explain, in your own words, what "Your Rule" is and its purpose in the task?

6. Can you explain, in your own words, the purpose of the task, in general?

7. Can you explain, in your own words, what "FIT" is and its purpose in the task?

8. Can you explain, in your own words, what "Does not fit" is and its purpose in the task?

9. Can you explain, in your own words, what the "feedback" is and its purpose in the task?

10. Can you explain, in your own words, what the purpose of the "new triple" is in the task?

11. How much do you believe your hypothesis?

12. Can you explain, in your own words, what it means to "change your mind about which trials you received false feedback."

13. Can you explain, in your own words, how to record error on the spreadsheet?

14. Can you explain, in your own words, how to use the spreadsheet in general?

15. Can you explain, in your own words, what "Fit" or "Does not fit" means for Your Rule?

16. How many chances do you have to guess?

17. Can you explain, in your own words, what "guessing" is and its purpose in the task?

18. Can you explain, in your own words, what the "Error" is and its purpose in the task?

Figures 7.1-7.3 show the failure rates for the sixteen auxiliary hypotheses for three iterations of the instructions in three sessions. The blue squares are posterior predictions of the failure probability, along with 95% credible intervals. The red squares are the actual (observed) failure probabilities. Each interview session included a series of 4 participants. After each session, the instructions were revised to reduce the probability of failure for each auxiliary hypothesis. A Beta$(1, 3)$ distribution was taken as the prior distribution for each failure rate, where Beta$(F, S)$ represents failures $(F)$ to comprehend and successful $(S)$ comprehension.

Overall, for any participant, the failure of at least one auxiliary hypothesis among the sixteen was very likely. Across the three iterations, 43 of the

48 (90%) observations fell within the 95% credible interval, indicating slight overconfidence but overall good calibration. It is important to note from this graph, that after only 4 iterations of the materials, I was able to have well-calibrated 95% credible intervals, with actual failure probabilities falling within the posterior predictions roughly 95% of the time. Between the fourth and sixth iterations, the accuracy of point estimates also greatly increased.


Figure 4.1: Posterior predictions and observed failure rates for iteration 4 of the Wason task design, Experiment One.

From the figures one can see that there was a consistent and intractable failure of auxiliary hypotheses 8 (the meaning of creating a new triple), 10 (changing their error attributions at the end of the task), and 15 (their final answer). The process determined not only that these auxiliary hypotheses are a problem, but also produced calibrated failure probabilities for each auxiliary

hypothesis. By knowing the probability of failure, it is possible to adjust inferences appropriately, as will be explored in the evidence synthesis approach described in Stage Five.

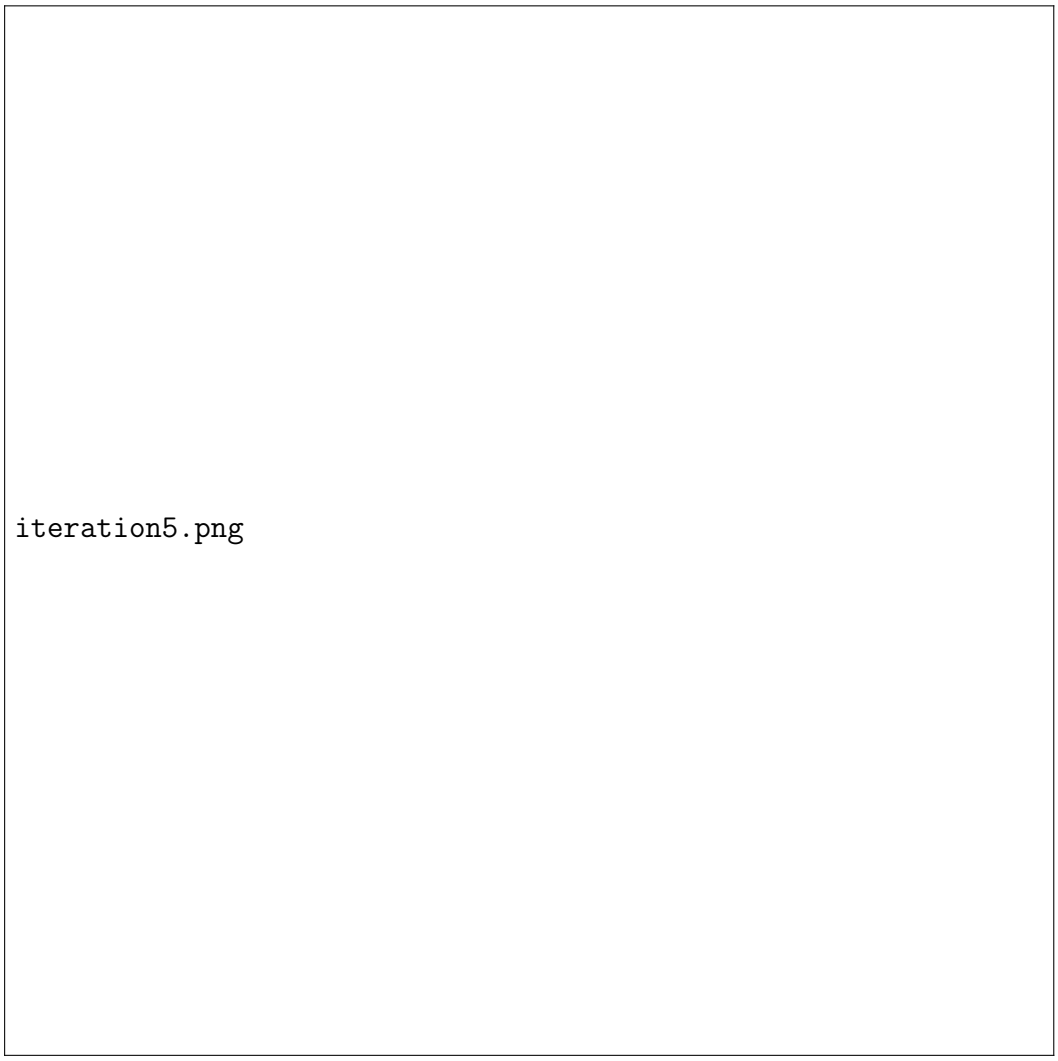


Figure 4.2: Posterior predictions and observed failure rates for iteration 5 of the Wason task design, Experiment One.


Figure 4.3: Posterior predictions and observed failure rates for iteration 6 of the Wason task design, Experiment One.

# Chapter 5

# Acceptance Sampling

Given a list of auxiliary hypotheses (specifications) and the ability to repeatedly adjust the design of the experiment to try to minimize their failure risks, pre-testing is in a form that is amenable to industrial design techniques, such as statistical quality control. For example, suppose we have designed an experiment with a specific set of instructions, questions, *etcetera*, as was done in the Wason example above. We have also agreed, provisionally, on ways of measuring the success or failure of each auxiliary hypothesis (e.g., open-ended responses on a cognitive interview). We now want to choose a sample size of participants to measure the auxiliary hypotheses, where if few enough failures occur we would accept the experimental design and move to Stage Three, but if too many failures occur, we redesign the experiment hoping to reduce the failure risks. This sequential testing process is equivalent to acceptance sampling that is used in industrial design and statistical quality control.

Here is simple example using literacy and motivation. A participant given instructions roughly between 6th and 8th grade reading level performs reasonably well on both open-ended and quiz-based tests of comprehension of the instructions. Thus, the comprehension auxiliary hypothesis is made within the tolerance level by the readability of the instructions. If we test 10 participants with these instructions, and only 1 of them fails to understand them, then we can apply acceptance sampling to get an idea of the risk of failures if we expand our examination more generally and accept those instructions as our experimental method. This is the problem of choosing a sample size to assure a specific quality of the product. Although the risks are logically infinite, they tend to actually be quite small, that is, detectable in about 20 participants (68).

Additionally, it is also possible to estimate the risks of new, previously not considered auxiliary hypotheses. The latter can be modeled by a poisson or exponential distribution. The key is that eventually the surprises will converg to zero: we may come up with at most 20 or 30 auxiliary hypotheses before we stop noticing surprises.

By making auxiliary hypotheses concrete, specific, and measurable, it is possible to create a precisely designed product that is a scientific experiment. The eventual aim of Stage Two is to manufacture parts of an experiment that meet specified auxiliary hypotheses exactly, just as the parts of each modern car is produced almost exactly the same (within some small margin of error)(69).

# Part III

# Stage Three: Pilot-Testing

Although Stage Two measures and reduces the risk of failed auxiliary hypotheses, we can never be sure our measurements are right. Suppose, for example, that we thought a question about the participant's comprehension of the task was a valid measure. To test that, however, would require a separate experiment to validate the measure. It is easy to see that testing our tests leads to an infinite regress. Thus, once we settle on tests we consider reasonable, and, according to those tests used in stage two our auxiliary hypotheses are measured to have low risk of being violated, we can conduct an experiment to see if we are right.

Stage Three does this. Instead of infinitely testing our tests, as would be required if we were committed to staying in Stage Two, Stage Three fuses the auxiliary hypotheses with the core hypothesis to make experimental predictions. However, the key feature of Stage Three that differentiates it from Stage Four is that it uses Lakatos' *negative heuristic* (4): if something goes wrong with the predictions, we consider that our Stage Two pre-testing didn't reveal all the problems with our auxiliary hypotheses; that is, we protect our core hypothesis and direct refutation at the fallible auxiliary hypotheses.

Why is this stage necessary, since we can directly test auxiliary hypotheses in Stage Two? First, we may not have thought of all necessary auxiliary hypotheses. Directly testing them in Stage Two does not guarantee that we find all important auxiliary hypotheses, but instead helps us estimate the failure probabilities of ones we've already considered. Second, there are limitations to the methodologies that can be employed in Stage Two (e.g., cognitive interviews). By using experimental manipulation instead of interviews and other measurements, and assuming that the core hypothesis is correct, one can draw the conclusion that a failed prediction in an experiment is an indication of a lurking omitted auxiliary hypothesis or a failed auxiliary hypothesis we have already considered. The form of the failed prediction can help us generate a plausible auxiliary hypothesis to explain it. That is, Stage Three uses failed prediction to generate ideas about possible ways our auxiliary hypotheses may have failed and possible ways of satisfying them. New auxiliary hypotheses generated in this way are part of what Mayo calles the *error repertoire* (5).

*Wason Task.* In Experiment One there was an unexpected lack of association between data sharing and both the feedback and whether the trial was attributed to error. The open-ended answers strongly suggested that participants didn't understand or pay attention to this part of the task. Tharing judgments were at the end of the task when participants were ready to quit. Participants also indicated that they wanted to propose new trials or state trials rather than triples. Their confusion with the data sharing measure was not considered beforehand, and the maturation threat was considered but not taken seriously enough.

Experiment Two, in contrast, did find a strong relationship between error

attributions, feedback, and data sharing. However, the response format that was used left two doubts in my mind. One was a testing artifact, where making the error attribution before the sharing judgment may cause some association between the two. First, since True and Share were both buttons on the left side of the screen, and False and Do not share on the right side, any tendency to merely click left or right would make a strong association between the two judgments. Second, since they were right after each other on the task, the participant may have inferred that they should be related, a demand effect. There was also no effect of the incentive on their scores, attribution patterns, or data sharing, even though the participants worked substantially harder.

Experiment Three used both trial-by-trial and end-of task data sharing judgments. While this doesn't account for the order effect or demand effect, it does allow comparison of end-of-task judgments to trial-by trial judgments, the former being less susceptible to these two artifacts. However, differences between trial-by-trial and end-of-task judgments could be caused both by differences in bias and differences in strategy, where participants do not know their final hypothesis during the task, making their sharing judgments ambiguous, but do know their final hypothesis at the end of the task, making a sharing strategy more worth developing. Thus, Experiment Three does not account for these artifacts of trial-by-trial judgments. Second, participants were given the perverse or compatible incentive to share, both giving them motivation to work toward finding a specific hypthesis–one that is convincing–and to sharing specific data that are consistent with this hypothesis. The data were exactly opposite of what I predicted: those in the perverse incentive condition tended to share all of their data. This was an interesting anomaly that didn't fit readily into either the specific auxiliary hypotheses mentioned in Stage One, nor those generated from the preposterior analysis in Stage Two. Instead I was completely dumbfounded.

More generally, although keeping track of the mistakes people generally make (i.e., the specific auxiliary hyptoheses), the mistakes I expect (from the preposterior analysis), and the mistakes I've previously made (defined implicitly in the experimental design that avoids the mistakes or explicitly in what Mayo calls the *error repertoire*, the errors I identified in all three experiments were ones that I hadn't considered beforehand. This could be because generating a new explanation is more interesting than attributing failed predictions to old errors (from the specific auxiliary hyptoheses, etc.), or that new explanations seem more convincing once new data are found (the hindsight that was expected, but not found, in the Experimental Surprises research).

# Part IV

# Stage Four: Testing

In Stage Four, the testing phase, the researcher is confident enough in her auxiliary hypotheses that she is willing to admit falsification of the core hypothesis if that is what the data suggest. After extensive pre-testing and pilot-testing in Stages Two and Three, we come to an experimental design that we believe has low enough risk of failed auxiliary hypotheses that if a failed prediction occured we would be willing to discard our hypothesis rather than invoke an auxiliary hypothesis to explain the failure.

There are two requirements to get to Stage Four testing. The *less* important requirement is that one is confident in one's auiliary hypotheses. The *more* important requirement is that one is willing to admit falsification of the core hypothesis under *any* circumstance. If this is not the case, then the core hypothesis is not appropriate for scientific investigation, for example if it is 'bubba psychology' (70). This occurs when experiments are "not to test our hypotheses but to demonstrate their obvious truth" for hypotheses that are "so clearly true (given the implicit and explicit assumptions)." If this is the case then,

> "the experiment tests is not whether the hypothesis is true but rather whether the experimenter is a sufficiently ingenious stage manager to produce in the laboratory conditions which demonstrate that an obviously true hypothesis is correct. In our graduate programs in social psychology, we try to train people who are good enough stage managers so that they can create in the laboratory simulations of realities in which the obvious correctness of our hypothesis can be demonstrated."

To be able to reach Stage Four testing, one must ask oneself at the outset of the research endeavor whether one would be willing to admit falsification of the core hypothesis given a perfect test. If not, then Stage Four testing can never occur.

As Stage Four testing is what most researchers have in mind when they develop methodological and statistical tools, most of the mainstream tools developed for the social sciences are applicable. This is because Stage Four testing assumes that one has already debugged the experiment, and that the only thing that could result in a failed prediction of a theory is that the theory is wrong, and an alternative is correct. This is the typical assumption used in most mainstream statistical approaches.

## 5.1 Wason Task

The Wason Task did not reach this stage. After reflecting on whether I would be willing to give up any of my core hypotheses, I think I would not be willing to give up the core hypothesis that feedback and error attributions

determine data sharing. I believe these to be true. Thus, they are poor core hypotheses. However, the core hypothesis that these data sharing policies are either normatively justified or effective are ones that I am willing to accept any answer on.

# Part V

# Stage Five: Evidence Synthesis

Unfortunately, if one completes the previous four stages, one will likely have a series of experiments that differ in complicated ways. At best, this is a nonstationary stochastic process. At worst, the experiments may seem completely unrelated to each other. How can one make sense of all this data? Is there a point where the 'warm-up' period of stages two and three ends and the 'real science' of Stage Four begins?

The common problem is this: one completes a series of experiments and finally one finds a set of auxiliary hypotheses where the core hypothesis makes a correct prediction. One has made a discovery. How should the previous experiments weigh on the judgment of the validity of the final discovery? If they are seen as related, they should cast doubt on our final result, requiring additional replications before firm conclusions can be drawn. If they are seen as unrelated, then the final experiment can be treated as the first member of its class. Similarly, if one is compelled to report these experiments to the scientific community, how far back should one go? To pilot-tests, pre-tests, or even thought experiments?

This is a very difficult problem. It is a typical criticism of meta-analysis; that is, the studies are unique, different, and thus not exchangeable. Treating them as exchangeable in that case is seen as flawed reasoning rather than reasonable assumption.

The solution I sketch here can be called *Generalized Meta-Analysis* (GMA). It is general enough to take into account arbitrary relationships between core hypothesis, auxiliary hypotheses, and experiments, including probabilistic risk analysis of auxiliary hypotheses. It can implement the Theory-Based Causal Induction framework directly. The approach uses Hierarchical Bayesian Models (HBMs) in the stochastic programming language Church (71).[1]

The approach works as follows. The hypothesis under examination is called the core hypothesis and is evaluated as a stochastic program along with auxiliary hypotheses that are required for it to make predictions about each experiment. While the core hypothesis remains the same over the experiments, the auxiliary hypotheses need not remain the same. Each set of auxiliary hypotheses allow each experiment to be interpretable in light of the core hypothesis. The core hypothesis makes probabilistic predictions for each experiment conditional on the auxiliary hypotheses.

There are three difficulties with this approach: 1) knowing the appropriate set of auxiliary hypotheses for each experiment, 2) knowing how each auxiliary hypothesis modifies the predictions of the core hypothesis, and 3) knowing how likely each of the auxiliary hypotheses are to be met conditional on the

[1] An important warning should be made before beginning discussion. The construction of the HBM should not interfere with critical and dynamic thinking. What I describe here is merely a formalized representation and way of performing computations on what one already believes. One should not rigidly adhere to it or think that the numbers are in any sense correct any more than one's subjective beliefs are correct.

experimental design. Although empirical evidence may be available on each of these three difficulties, the approach has no direct answer. Instead, it allows the researcher to make formal guesses and evaluate the consequences of these guesses in light of other data.

*Wason Task.* To make the method clear, consider an example from the Wason task in Chapter Five. In this task, I proposed a core hypothesis that there was some correlation between the receipt of disconfirming feedback and judgment that the feedback was error. This correlation can be called phi ($\phi$), which is a correlation coefficient for a $2 \times 2$ binary contingency table. Thus, my core hypothesis is that $\phi$ is positive, most likely above 0.3. This can be modeled as a scaled beta distribution, with $2 * \text{Beta}(a, b) - 1$ covering the interval $[-1, 1]$, the range of $\phi$. The core hypothesis proposes a specific distribution, parameterized by {a,b}, that is consistent with $\phi$ being above 0.3. Suppose, based on intuition, we set the hypothesis with parameters $a = 7$, $b = 3$, giving a mean $\phi$ of $2 * 0.7 - 1 = 0.4$. If we were to take this to be our naive model of the data (that is, assuming all auxiliary hypotheses are met), then we can take the observed sampling distribution of $\phi$ and compare it to our model to yield both a likelihood and posterior distribution of our hypothesis.

However, suppose we have some estimated risk of violating an auxiliary hypothesis about communication. Suppose we think that if this auxiliary hypothesis is violated, then participants will not understand what they are expected to do, and will respond randomly. Thus, if the auxiliary hypothesis is violated, we would expect $\phi$ to be distributed uniformly over the interval $[-1, 1]$. Suppose that, based on our measurements in Stage Two, we found that of four participants interviewed, three participants understood the instructions, and one did not, so our posterior estimate of the probability of this auxiliary hypothesis being violated is distributed $\text{Beta}(1, 3)$, assuming improper $\text{Beta}(0, 0)$ priors (we could also use Jeffreys' invariance prior or an informative prior if we want). Now, suppose our second experiment uses a different set of instructions which we estimated reduced this risk to $\text{Beta}(1, 9)$; that is, out of 10 participants, only 1 failed to comprehend the instructions.

To integrate the two experiments together with the core hypothesis, all we need to do is examine the following cases. If our hypothesis is true and the auxiliary hypothesis is true in both experiments, then our hypothesis makes predictions $2 * \text{Beta}(7, 3) - 1$ for $\phi$ for both experiments. The expected value is $\phi = 0.4$. However, this is weighted by the probability that the auxiliary hypotheses are true for both experiments, which is the product of the two beta random variables (two coin flips) which is, on average, equal to $3/4 * 9/10 = 28/40$. Next, it is possible that the communication auxiliary hypothesis was violated for the first experiment, but not the second. This would happen, in expectation, with probability $1/4 * 9/10 = 9/40$. Likewise, it is possible that the communication auxiliary hypothesis was violated for

the second experiment and not the first, this would happen, in expectation, with probability $3/4 * 1/10 = 3/40$. Finally, the auxiliary hypothesis could be violated in both experiments with probability $1/4 * 1/10 = 1/40$.

As a result, if our core hypothesis is true, it makes the following predictions for the two experiments. With probability 28/40, it predicts $\phi \sim 2*\text{Beta}(7,3)-1$ for both experiments. With probability 9/40, it predicts $\phi \sim U[-1,1]$ for the first experiment and $\phi \sim 2*\text{Beta}(7,3)-1$ for the second. With probability 3/40 it predicts $\phi \sim 2*\text{Beta}(7,3)-1$ for the first experiment and $\phi \sim U[-1,1]$ for the second experiment. Finally, with probability 1/40 it will predict $\phi \sim U[-1,1]$ for both experiments.

What does this tell us? In general, the predictions of our hypothesis should be more diffuse than if we were naive, as the predictions are mixed with the uniform distribution for 12/40 cases, in expectation.

Now, what if we want to 'debug' our experiment, and pinpoint which of two, non-equivalent auxiliary hypotheses were more likely to be violated? Suppose that in both experiments we also had an alternative auxiliary hypothesis that participants inferred from the order of the error judgment and feedback that they should be highly related. This would create a $\phi$ coefficient much higher than we would otherwise expect, say $\phi \sim 2 * \text{Beta}(19,1) - 1$. Just as with the other auxiliary hypothesis, in Stage Two pretesting we found the violation of this hypothesis was $\text{Beta}(2,1)$ for the first experiment and $\text{Beta}(1,4)$ for the second. That is, the first experiment used methods that were more likely to be susceptible to this order effect than the second.

Using this, we can derive the following qualitative analysis. If $\phi$ is very high in experiment one, this is likely due to the second auxiliary hypothesis (no order effects) being violated. However, if it is very high in experiment two, we are more likely to expect that our theory was correct. Negative $\phi$ coefficents for either experiment strongly suggest that participants didn't understand the instructions, although this is more true for the first rather than second experiment. Importantly, the degree to which we update our core hypothesis will depend on the risk of these auxiliary hypotheses.

We can derive posterior probabilities for our core hypothesis if we have specific alternatives in mind by setting some values for {a,b} in the $\phi \sim 2 * \text{Beta}(a,b) - 1$ distribution. Alternatively, if we do not have any in mind, we can simulate them by drawing random {a,b} values of equal total to the one we consider (e.g., $a + b = 10$). That is, we first draw a value for $a \sim U[0,10]$ and then calculate $b = 10 - a$. By doing this repeatedly for $N$ models, we can create $N$ random alternative hypotheses to the one we've considered. The prior probability of any of these hypotheses will be dirichlet distributed. This procedure can be extended to arbitrary numbers of theories, auxiliary hypotheses and experiments. However, computation slows down rapidly.

Each of the following examples will use a simple example. The prior for

auxiliary hypotheses are the same in experiment one, about 50% as estimated from the Stage Two pretesting (Beta$(1, 1)$). After experiment one, we've refined the methods, and now our Stage Two pretesting of experiment two reveals an estimate of each auxiliary hypothesesis being true 75% of the time (Beta$(3, 1)$).

## 5.2 Example 1: Two Disconfirmations

Naively, we can consider two disconfirmations in both experiments being $\phi_1 = 0.1$ and $\phi_2 = 0.1$. As seen in the table below, the generalized meta-analysis using the metropolis-hastings algorithm tells us that our core hypothesis is actually slightly more likely after receiving two 'disconfirmations' ($\phi_1 = \phi_2 = 0.1$). XX

| Hypothesis | Likelihood | Prior | Posterior |
|---|---|---|---|
| Naive Core Hypothesis | $2 * \text{Beta}(14, 6) - 1$ | Beta$(1, 1)$ | ?? |
| GMA Core Hypothesis | $2 * \text{Beta}(14, 6) - 1$ | Beta$(1, 1)$ | ?? |
| Auxiliary 1 Exp 1 | $U[-1, 1]$ | Beta$(1, 1)$ | ?? |
| Auxiliary 2 Exp 1 | $2 * \text{Beta}(19, 1) - 1$ | Beta$(1, 1)$ | ?? |
| Auxiliary 1 Exp 2 | $U[-1, 1] - 1$ | Beta$(3, 1)$ | ?? |
| Auxiliary 2 Exp 2 | $2 * \text{Beta}(19, 1) - 1$ | Beta$(3, 1)$ | ?? |

## 5.3 Example 2: Confirmation and Disconfirmation

In example two, confirmation is observed in Experiment One ($\phi_1 = 0.4$) and disconfirmation is observed in Experiment Two ($\phi_2 = 0$). xx

| Hypothesis | Likelihood | Prior | Posterior |
|---|---|---|---|
| Naive Core Hypothesis | $2 * \text{Beta}(14, 6) - 1$ | Beta$(1, 1)$ | ?? |
| GMA Core Hypothesis | $2 * \text{Beta}(14, 6) - 1$ | Beta$(1, 1)$ | ?? |
| Auxiliary 1 Exp 1 | $U[-1, 1]$ | Beta$(1, 1)$ | ?? |
| Auxiliary 2 Exp 1 | $2 * \text{Beta}(19, 1) - 1$ | Beta$(1, 1)$ | ?? |
| Auxiliary 1 Exp 2 | $U[-1, 1] - 1$ | Beta$(3, 1)$ | ?? |
| Auxiliary 2 Exp 2 | $2 * \text{Beta}(19, 1) - 1$ | Beta$(3, 1)$ | ?? |

## 5.4 Example 3: Disconfirmation and Confirmation

In example three, disconfirmation is observed in Experiment One ($\phi_1 = 0$) and confirmation is observed in Experiment Two ($\phi_2 = 0.4$). xx

The relationship is not symmetric. A naive meta-analysis would give the same overall result with $\phi_1 = 0$ and $\phi_2 = 0.4$ as the reverse. However, GMA allows us to weight the better second experiment more, thus yielding more confirmation of our hypothesis than the first. This is because we believe that auxiliary hypothesis one was likely violated in Experiment One.

| Hypothesis | Likelihood | Prior | Posterior |
|---|---|---|---|
| Naive Core Hypothesis | $2 * \text{Beta}(14, 6) - 1$ | $\text{Beta}(1, 1)$ | ?? |
| GMA Core Hypothesis | $2 * \text{Beta}(14, 6) - 1$ | $\text{Beta}(1, 1)$ | ?? |
| Auxiliary 1 Exp 1 | $U[-1, 1]$ | $\text{Beta}(1, 1)$ | ?? |
| Auxiliary 2 Exp 1 | $2 * \text{Beta}(19, 1) - 1$ | $\text{Beta}(1, 1)$ | ?? |
| Auxiliary 1 Exp 2 | $U[-1, 1] - 1$ | $\text{Beta}(3, 1)$ | ?? |
| Auxiliary 2 Exp 2 | $2 * \text{Beta}(19, 1) - 1$ | $\text{Beta}(3, 1)$ | ?? |

## 5.5 Example 4: Two Confirmations

In example four, confirmation is observed in Experiment One ($\phi_1 = 0.4$) and confirmation is observed in Experiment Two ($\phi_2 = 0.4$).

We can see that, although disconfirmation of experiment one doesn't harm our hypothesis, it also doesn't help much once we have confirmation from Experiment Two. This is again in contrast to a naive approach which would have much stronger results with two confirmations than one.

| Hypothesis | Likelihood | Prior | Posterior |
|---|---|---|---|
| Naive Core Hypothesis | $2 * \text{Beta}(14, 6) - 1$ | $\text{Beta}(1, 1)$ | ?? |
| GMA Core Hypothesis | $2 * \text{Beta}(14, 6) - 1$ | $\text{Beta}(1, 1)$ | ?? |
| Auxiliary 1 Exp 1 | $U[-1, 1]$ | $\text{Beta}(1, 1)$ | ?? |
| Auxiliary 2 Exp 1 | $2 * \text{Beta}(19, 1) - 1$ | $\text{Beta}(1, 1)$ | ?? |
| Auxiliary 1 Exp 2 | $U[-1, 1] - 1$ | $\text{Beta}(3, 1)$ | ?? |
| Auxiliary 2 Exp 2 | $2 * \text{Beta}(19, 1) - 1$ | $\text{Beta}(3, 1)$ | ?? |

From this setup, we can condition on any observed $\phi$ values for Experiment One and Two and derive the posterior probability of the auxiliary hypotheses being violated in either or both experiment. With this formalism, we can then determine what modifications to make to our experiment in Stage Three when our predictions are disconfirmed. Interesting patterns can emerge, where

negative results have no effect on the belief in our core hypothesis, or even increase our belief.

# Part VI

# Conclusion

Using the Theory-Based Causal Induction approach of Griffiths and Tenenbaum (12) we can create carefully constructed theories that satisfy Popper's requirement of sufficient axiomatization for falsification. We also defend against two conventionalist strategems of modifying our ostensive definitions or blaming the theoretician. Additionally, by using a standardized set of both general and specific auxiliary hypotheses, along with additional ones generated for the specific topic, we satisfy Mayo's requirement for a precise ceteris paribus clause and error repertoires. This sets us up for the severe testing necessary in Stages Two, Three and Four.

Stage Two provides a novel approach to pre-testing. By systematically imagining scenarios that would lead our predictions to be false (preposterior analysis) along with consideration of the general and specific auxiliary hypotheses from Stage One, we can design experiments to iteratively test, refine, and estimate the risk of these auxiliary hypotheses failing. Using acceptance sampling and statistical quality control techniques, we and choose sample sizes for each iteration of the experimental design to estimate this risk without wasting resources.

Stage Three presents pilot-testing. It assumes Lakatos' negative heuristic and directs empirical refutation at the auxiliary hypotheses. Failed experiments are used to try to pinpoint the cause of the failure. If the cause was one of the auxiliary hypotheses considered in Stage Two, we can use the Generalized Meta-Analysis to pinpoint the auxiliary hypotheses with the highest posterior probability of failing. If the cause was not one of the auxiliary hypotheses in Stage Two, strange error patterns will emerge and this will encourage us to consider new auxiliary hypotheses and return to Stage Two to estimate their failure risk.

Once we are quite sure that our auxiliary hypotheses are met, and our core hypothesis successfully predicts experimental results, we can compare it to an important alternative hypothesis in Stage Four. At this point, the failure of the hypothesis indicates strong reason to reject it. This is Mayo's severe test.

Finally, the evidence synthesis approach, Stage Five, allows the researcher to deal with the problem of 'warm-up' experiments. Those risky experiments will be weighted properly by this scheme. Disconfirming evidence will do more to indicate that an auxiliary was violated than the core hypothesis was false.

"Although the number of works upon Methodeutic since Bacon's Novum Organum has been large, none has been greatly illuminative. Bacon's work was a total failure, eloquently pointing out some obvious sources of error, and to some minds stimulating, but affording no real help to an earnest inquirer. THE book on this subject remains to be written; and what I am chiefly concerned to do is to make the writing of it more possible."

—Charles Peirce, 1931, The Collected Works, Vol 2, 109 (72)

"When I was young, no remark was more frequent than that a given method, though excellent in one science, would be disastrous in another. If a mere aping of the externals of a method were meant, the remark might pass. But it was, on the contrary, applied to extensions of methods in their true souls. I early convinced myself that, on the contrary, that was the way in which methods must be improved; and great things have been accomplished during my life-time by such extensions. I mention my early foreseeing that it would be so, because it led me, in studying the methods which I saw pursued by scientific men, mathematicians, and other thinkers, always to seek to generalize my conception of their methods, as far as it could be done without destroying the forcefulness of those methods. This statement will serve to show about how much is to be expected from this part of my work."

—Charles Peirce, 1931, The Collected Works, Vol 2, 110 (72)

# Bibliography


[1] C. Peirce, "The fixation of belief," 1877.

[2] K. Popper, *The logic of scientific discovery*. Psychology Press, 2002.

[3] T. Kuhn, *The structure of scientific revolutions*. University of Chicago press, 1996.

[4] I. Lakatos, J. Worrall, and G. Currie, *The methodology of scientific research programmes*, vol. 1. Cambridge Univ Pr, 1980.

[5] D. Mayo, *Error and the growth of experimental knowledge*. University of Chicago Press, 1996.

[6] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. Murdock, E. Nyberg, J. Prager, *et al.*, "Building watson: An overview of the deepqa project," *AI Magazine*, vol. 31, no. 3, pp. 59–79, 2010.

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

[8] E. Yong, "Bad copy," *Nature*, vol. 485, no. 7398, pp. 298–300, 2012.

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

[10] D. Klahr and H. Simon, "Studies of scientific discovery: Complementary approaches and convergent findings.," *Psychological Bulletin*, vol. 125, no. 5, p. 524, 1999.

[11] C. Schunn and D. Klahr, "The problem of problem spaces: When and how to go beyond a 2-space model of scientific discovery," in *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, pp. 25–26, 1996.

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

[13] R. Scheines, "The similarity of causal inference in experimental and non-experimental studies," *Philosophy of Science*, vol. 72, no. 5, pp. 927–940, 2005.

[14] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search*, vol. 81. The MIT Press, 2000.

[15] R. Rosenthal and R. Rosnow, "Essentials of behavioral research: Methods and data analysis," *New York*, 1991.

[16] W. Shadish, T. Cook, and D. Campbell, "Experimental and quasi-experimental designs for generalized causal inference," 2002.

[17] L. Narens and R. Luce, "Measurement: The theory of numerical assignments.," *Psychological Bulletin*, vol. 99, no. 2, p. 166, 1986.

[18] P. Meehl, "Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it," *Psychological Inquiry*, vol. 1, no. 2, pp. 108–141, 1990.

[19] P. Meehl, "The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions," *What if there were no significance tests*, pp. 393–425, 1997.

[20] P. Meehl and N. Waller, "The path analysis controversy: A new statistical approach to strong appraisal of verisimilitude.," *Psychological Methods*, vol. 7, no. 3, p. 283, 2002.

[21] D. Borsboom, G. Mellenbergh, and J. Van Heerden, "The concept of validity.," *Psychological Review; Psychological Review*, vol. 111, no. 4, p. 1061, 2004.

[22] I. Levi, *The covenant of reason: rationality and the commitments of thought*. Cambridge Univ Pr, 1997.

[23] 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.

[24] A. Varzi, "Basic problems of mereotopology," *Formal Ontology in Information Systems. IOS Press (this volume)*, 1998.

[25] C. Luhmann and W. Ahn, "Buckle: A model of unobserved cause learning.," *Psychological review*, vol. 114, no. 3, p. 657, 2007.

[26] J. Pearl, *Causality: models, reasoning, and inference*, vol. 47. Cambridge Univ Press, 2000.

[27] W. Gilks, A. Thomas, and D. Spiegelhalter, "A language and program for complex bayesian modelling," *The Statistician*, pp. 169–177, 1994.

[28] M. Lee and E. Wagenmakers, "A course in bayesian graphical modeling for cognitive science," *Unpublished manuscript. http://users. fmg. uva. nl/ewagenmakers/BayesCourse/BayesBook. pdf*, 2009.

[29] D. Spiegelhalter, "Bayesian graphical modelling: a case-study in monitoring health outcomes," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 47, no. 1, pp. 115–133, 1998.

[30] C. Hempel, "Formulation and formalization of scientific theories," *The Structure of Scientific Theories*, vol. 2, pp. 244–55, 1977.

[31] L. Phillips, B. Fasolo, N. Zafiropoulos, *et al.*, "Benefit-risk methodology project work package 2 report: applicability of current tools and processes for regulatory benefit-risk assessment 31 august 2010 ema/549682/2010-revision 1."

[32] M. Morgan and M. Henrion, "Uncertainty: a guide to the treatment of uncertainty in quantitative policy and risk analysis," *Uncertainty: a guide to the treatment of uncertainty in quantitative policy and risk analysis*, 1990.

[33] D. Spiegelhalter and H. Riesch, "Don't know, can't know: embracing deeper uncertainties when analysing risks," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 369, no. 1956, pp. 4730–4750, 2011.

[34] G. Bammer and M. Smithson, *Uncertainty and risk: multidisciplinary perspectives*. Earthscan/James & James, 2008.

[35] J. Van Der Sluijs, M. Craye, S. Funtowicz, P. Kloprogge, J. Ravetz, and J. Risbey, "Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: The nusap system," *Risk Analysis*, vol. 25, no. 2, pp. 481–492, 2005.

[36] B. Fischhoff, W. de Bruin, Ü. Güvenç, D. Caruso, and L. Brilliant, "Analyzing disaster risks and plans: An avian flu example," *Journal of Risk and Uncertainty*, vol. 33, no. 1, pp. 131–149, 2006.

[37] R. Rosenthal and R. Rosnow, "The volunteer subject.," 1975.

[38] C. Cannell, P. Miller, and L. Oksenberg, "Research on interviewing techniques," *Sociological methodology*, vol. 12, pp. 389–437, 1981.

[39] G. Wells and P. Windschitl, "Stimulus sampling and social psychological experimentation," *Personality and Social Psychology Bulletin*, vol. 25, no. 9, pp. 1115–1125, 1999.

[40] D. Dillman, *Mail and internet surveys: The tailored design method*. John Wiley & Sons Inc, 2007.

[41] P. Suppes and J. Zinnes, "Basic measurement theory," *Handbook of mathematical psychology*, vol. 1, no. 1-76, 1963.

[42] C. Coombs, R. Dawes, and A. Tversky, "Mathematical psychology: an elementary introduction.," 1970.

[43] B. Fischhoff, N. Brewer, J. Downs, U. S. Food, and D. Administration, *Communicating Risks and Benefits: An Evidence-based User's Guide*. Food and Drug Administration, 2011.

[44] N. Schwarz, "Self-reports: How the questions shape the answers.," *American psychologist*, vol. 54, no. 2, p. 93, 1999.

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

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

[47] F. Khatib, S. Cooper, M. Tyka, K. Xu, I. Makedon, Z. Popovi, D. Baker, and F. Players, "Algorithm discovery by protein folding game players," *Proceedings of the National Academy of Sciences*, vol. 108, no. 47, pp. 18949–18953, 2011.

[48] D. Michael and S. Chen, *Serious games: Games that educate, train, and inform*. Muska & Lipman/Premier-Trade, 2005.

[49] B. Bergeron, *Developing Serious Games (Game Development Series)*. {Charles River Media}, 2006.

[50] J. Carroll, "Beyond fun," *interactions*, vol. 11, no. 5, pp. 38–40, 2004.

[51] R. Koster, *A Theory Of Fun In Game Design*. Paraglyph press, 2005.

[52] J. McGonigal, *Reality is broken: Why games make us better and how they can change the world*. Penguin Pr, 2011.

[53] B. Reeves and J. Read, *Total engagement: using games and virtual worlds to change the way people work and businesses compete*. Harvard Business School Press, 2009.

[54] D. Pink, *Drive: The surprising truth about what motivates us*. Canongate, 2010.

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

[56] R. Turner, D. Spiegelhalter, G. Smith, and S. Thompson, "Bias modelling in evidence synthesis," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 172, no. 1, pp. 21–47, 2009.

[57] S. Treweek, M. Pitkethly, J. Cook, M. Kjeldstrøm, T. Taskila, M. Johansen, F. Sullivan, S. Wilson, C. Jackson, R. Jones, *et al.*, "Strategies to improve recruitment to randomised controlled trials," *Cochrane Database of Systematic Reviews*, vol. 4, 2010.

[58] P. Edwards, I. Roberts, M. Clarke, C. DiGuiseppi, R. Wentz, I. Kwan, R. Cooper, L. Felix, and S. Pratap, "Methods to increase response to postal and electronic questionnaires (review)," 2009.

[59] R. Prescott, C. Counsell, W. Gillespie, A. Grant, I. Russell, S. Kiauka, I. Colthart, S. Ross, S. Shepherd, D. Russell, *et al.*, "Factors that limit the quality, number and progress of randomised controlled trials.," *Health technology assessment (Winchester, England)*, vol. 3, no. 20, p. 1, 1999.

[60] J. Watson and D. Torgerson, "Increasing recruitment to randomised trials: a review of randomised controlled trials," *BMC medical research methodology*, vol. 6, no. 1, p. 34, 2006.

[61] L. Cronbach and P. Meehl, "Construct validity in psychological tests.," *Psychological bulletin*, vol. 52, no. 4, p. 281, 1955.

[62] B. Fischhoff, "Value elicitation: is there anything in there?," *American Psychologist*, vol. 46, no. 8, p. 835, 1991.

[63] C. Schunn and D. Klahr, "A 4-space model of scientific discovery," in *Proceedings of the seventeenth annual conference of the Cognitive Science Society*, pp. 106–111, 1995.

[64] D. Vose, *Risk analysis: a quantitative guide*. John Wiley & Sons Inc, 2008.

[65] I. Levi, *For the sake of the argument: Ramsey test conditionals, Inductive Inference, and Nonmonotonic reasoning*. Cambridge Univ Pr, 1996.

[66] G. Willis *et al.*, "Cognitive interviewing: A âhow toâ guide," *Research Triangle Park, NC: Research Triangle Institute. Retrieved January*, vol. 21, p. 2004, 1999.

[67] K. Ericsson and H. Simon, "Verbal reports as data.," *Psychological review*, vol. 87, no. 3, p. 215, 1980.

[68] G. Morgan, B. Fischhoff, A. Bostrom, and C. Atman, *Risk Communication: A Mental Models Approach*. Cambridge University Press, 2001.

[69] W. Shewhart, *Statistical method from the viewpoint of quality control*. Dover Publications, 1939.

[70] W. McGuire, "The yin and yang of progress in social psychology: Seven koan.," *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology*, vol. 26, no. 3, p. 446, 1973.

[71] N. Goodman, V. Mansinghka, D. Roy, K. Bonawitz, and J. Tenenbaum, "Church: a language for generative models," in *Uncertainty in Artificial Intelligence*, vol. 22, p. 23, 2008.

[72] C. Peirce, "The collected works of charles sanders peirce (8 vols)," 1931.