Theory on the Design of Experiments

Brandon Campbell
Department of Economics,
Texas A&M University

April 26, 2013

Abstract

An experiment is a structured model to measure the validity of a hypothesis. There is extensive literature on how to design an experiment, derived by taking a already run experiment as a starting point, and determining what could be done better to get results that are more conclusive. As a result, the concept of an experiment has been treated informally, but if we seek to *design* an experiment, we need a formal basis. This paper formalizes the concept of an experiment, and what it means to test a hypothesis. It defines what an experiment is, what a hypothesis is, it defines the concept of construct validity, which is similar to identification or internal validity, and establishes basic experimental design features necessary to get conclusive results. Future work on this paper will consider the case where not everything is controllable, and we have to test things indirectly.

JEL Classification: C90, D78, D82

Keywords: Consistency, Construct Validity, Control, Independent.

1 Introduction

A man had a theory, and decided one day he was going to test it. He created his hypothesis, and collected the resources he would need for the experiment. A large box, a measuring tape, a pencil, an ink pad, a needle and something to heat it up, and a cockroach. He chose one end of the box to be the starting point. He then took the cockroach and pressed its feet to the ink pad, and placed it down in the box at the starting line. He then heated up the needle, and prodded it to the cockroaches bum. The cockroach jumped, and landed some distance away, leaving behind some inky footprints. Our Professor took out his measuring tape, and measured the distance. A whole 26 inches.

He then took the cockroach, pulled off two of its legs, one from each side, and repeated the experiment. He pressed the cockroaches remaining feet on the ink pad, and placed it on the start line. He again prodded it with the hot needle, and measured the foot print left behind. This time it was only 16 inches. He noted it down, and pulled two more legs off the cockroach. Now with only two legs, the cockroach only jumped a distance of 8 inches. He noted this down as well, and pulled off the remaining two legs. This time he pressed the cockroaches body against the ink pad, and then placed him at the starting line. He applied the hot needle to the cockroaches bum, but it didn't move. The professor checked to make sure the needle was hot - and it was plenty hot - and tried again. The cockroach still moved a sum of 0 inches on the as trial. Satisfied, the professor was proud to confirm his hypothesis that cockroaches with no legs feel no pain, and full legged cockroaches feel less pain the more legs that are removed.

Much like the above story, in economics we attempt to affirm or falsify a hypothesis, and sometimes use improperly designed experiments to do so. We can arrive at false conclusions, that are later falsified by other experiments. This system works, its very similar to evolution, but it works only *slowly*. If we could find a way to design an experiment properly from the outset, we could find results faster, even *conclusive* results, which has mostly evaded much of the empirical research in economics.

What this paper is not is a step-by-step manual to design an experiment properly. Such an algorithm can not exists without a proper foundation and structure to explain what an experiment is, and under what criteria we know we have conclusive results. That last part is what this paper addresses. It seeks to lay down the foundation. It creates an abstract mathematical environment inclusive of all forms of experiments, and it specifies structural requirements needed for an experiment to have conclusive results. This creates for us the foundation from which we can then next begin discussing the actual design of experiments in detail.

The paper proceeds by discussing basic properties of theories in section 2. In section 3 I introduce the model and the necessary and sufficient results for construct validity. In section 4, I begin introducing some earlier results on using indirect controls and observation in experiments, and in section 5 I extend the experiment diagnoses for when we use optimal theory predictions, the most common predictions we test.

2 Truth And Theory

This section is short because it is introducing the most primitive properties of a theory and an experiment. An experiment is used to establish validity, by directly observing results. However, to gain better control, theory became a bigger part of the picture in the early 1900's, and is particularly useful in allowing us to extend our experiments to test things that are not directly observable. As a result, in modern science, Theory and science walk hand in hand and often thought of as two sides of a coin, whereas in truth science was simply the empirical approach at first. The two properties we will introduce are consistency, and validity.

Definition 1 (Consistency)**.** *A statement is consistent if it is not possible to derive a contradiction with logic from the statement.*

Theorem 1 (Internal Consistency of Truth)**.** *If we believe there is an a single absolute truth, and that logic transforms true premises into true statements, then each piece of the single absolute truth must be consistent with the other parts.*

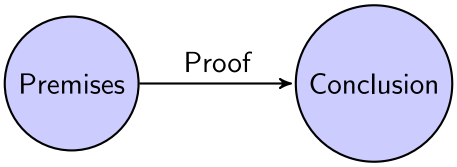
Proof. If not, then this means that using logic, it is possible to arrive at a contradiction using true premises derived from the absolute truth. With a contradiction, you can prove anything using logic. This includes using logic that starts with true premises and concludes with a falsehood. This is a contradiction of how logic operates. Therefore, if something is true of the natural world, than it must be consistent with other true facts of the natural world. □

The concept of consistency is important, because it is the fundamental concept behind analysis. If we plan to design something, for example an experiment, we now can use analyses to remove possible bad designs if they are inconsistent. This is the justification behind designing games or mechanisms in economic theory as well as other theories. While it fails to reject all false theories, since a theory can be consistent but still false from being based on false premises, it will never reject a true theory. As a result, it is useful to remove false theories.

Definition 2 (Validity)**.** *A premise is valid if it is true in the context you are making the statement.*

The context of a statement is the setting, environment, i.e. its relationship with all the other factors specified in the *true* model. If you are developing a model to explain patterns observed in reality, than the *true* model would be reality, what I call the natural world. A premise of a model to explain reality is valid if the premise is true in the natural world. As a result, I can now affirm a theory if I prove it consistent, and prove its premises are true.

The last primitive thing to consider is theory. Theory is useful as it is the bases of our hypothesis, and we will be using it to determine if an experimental design can give us conclusive results. A theory in its most base format in logic is: going from a precedent, proof, to conclusion. Without loss of generality, we can transform this to a set of the premises (P), the proof of logic (L), and the set of conclusion/ results (R), such that



Therefore, a theory is consistent if its proof that transforms the premises to the conclusion is consistent, and the set of premises is consistent with each other. If it is consistent, then it is possible for the statement or model to be a true statement, and we are left with only one major step, determining the validity of the premises.

3 Experimental Design to get Conclusive Results

If we look at a very simple model of an experiment, we can decomposed it into a few parts. but abstract ones.

Let $\gamma : A \times \Theta \rightarrow \Omega$ be the data generation process, which transmits the action vector a and the state of the world θ into some observable result ω. Let A, Θ,Ω, and Γ represent the set of all possible action vectors, states of the world, observable results, and all possible data generating processes, respectively. Further, specify $A = \times^n[0,1]$, such that each column in the vector can assume a range between 0 and 1, and action profile has $n \in \mathbb{N}$ columns or variables.

We can think of the action profile as all the choices people make that influences the environment. In a standard economics experiment, this would subjects choices in a game and their choice over beliefs they have. However, it also includes choices made by the experimenter, such that all fields of science have something in their action profile. The subjects endowment, preferences, how nature chooses when it is used as a randomizing device, would all be elements of the state of the world. The data generating process would include the mechanism, but also how things interact in the natural world, such as how sitting on a chair suspends the body above the floor. These are basic components, but we also need to account for three different ways in which they can categorized: the natural world, the hypothesis, and the experiment. We have talked about the natural world before, but below is its formal definition.

Definition 3 (Natural World)**.** *The design and values of the variables that occur when no experimental controls are exercised*[1]*. The value of a variable in the real world will be subscripted by f, such that a_f is the real world values of the action profile.*

Let us define a **hypothesis** as h, a subset of $\Gamma \times A \times \Theta$, with a range in Ω. The smaller the range in Ω is, the easier it would be to test the hypothesis. Let $C(h)$ consist of all variables specified and controlled in the hypothesis h. A variables is **controlled** in an experiment when an experimenters can choose the value for said variable directly. You **partially control** a variable when you can reduce the values a variable can take on, but not down to a single variable.

Let $I(h)$ consist of all variables not specified in h, that must instead be held independent. A variable is held **independent** if it has no affect on the observable result, no matter what value it realizes. Finally, let us define the **Consistent Interpretations function**. This is a correspondence $g(\tilde{h}, \omega | C(\hat{h})) \rightarrow \Gamma \times A \times \Theta$ that maps the experiment and the observed outcome into a set of possible hypothesis consistent with the result, conditional on the controls specified in the tested hypothesis.

Now, the true data generation process is usually unknown, but a experimenter can control everything in the action profile, parts of the data generation process, and even the state of the

[1] This is the same things as true data generating process, state of the world, and action profile.

world in lab setting. In this section, we will assume the experimenter can control everything but the outcome and observe only the outcome, but I will relax these assumptions later. A variable is **observed** if its value is known or can become known by an already established method. Let $\hat{\gamma}$ represent the hypothesized data generation process, $\hat{a}$ the hypothesized action vector, and $\hat{\theta}$ the hypothesized state of the world, and let $\hat{\omega}$ represent the derived hypothesized observable result. Thus, $\hat{\gamma}(\hat{a}, \hat{\theta}) = \hat{\omega}$, and the hypothesis can look like:

$$h_0 : \ \hat{\omega} = \omega_0$$

or,

$$h_0 : \ \hat{\gamma}(\hat{a}, \hat{\theta}) = \omega_0.$$

In order for the hypothesis to potentially be true, it must be consistent as established earlier in the paper. We can verify this with analysis, but not observation. As a result, the function of the experiment is to test for validity. If we transform the hypothesis into abstract of a theory, then we have two parts to test the validity of, the premises or the conclusion [2]. To test either requires an experiment.

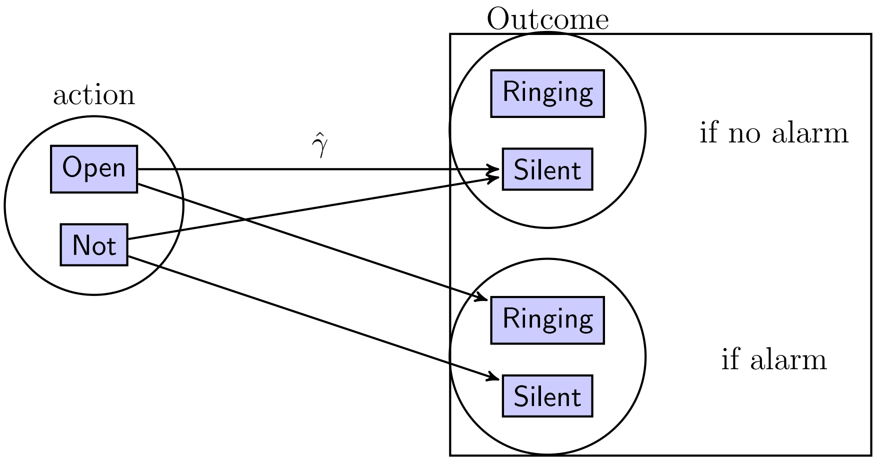
Definition 4 (Experiment)**.** *An experiment is the triplet of $\tilde{h} = (\tilde{\gamma}, \tilde{a}, \tilde{\theta})$ such that $\tilde{a} = (\hat{a}_1, ..., \hat{a}_{i-1}, a_i, ..., a_n)$ and if we decomposed the vector, we would get $\tilde{a}_{\{1,...,i-1\}} = \hat{a}_{\{1,...,i-1\}}$, and $\tilde{a}_{\{i,...,n\}} = a_{\{i,...,n\}}$. The same is likewise done for $\tilde{\gamma}$ and $\tilde{\theta}$.*

An experiment can not consist of an exact replication of the hypothesis, because otherwise we could not test the validity of anything. Therefore, the experiment must contain some point at which it allows freedom to for an element to be wrong, that the hypothesis specificises to be true. For example, suppose that at $\hat{a}_i = 0$ in the hypothesis. We design the experiment now to test if $\hat{a}_i$ is correct, which means the rest of the experiment is controlled for, but we

[2] You can not determine consistency with an experiment

do not control for $\hat{a}_i$, such that $\tilde{a}_i \in [0, 1]$, rather than specified as equal to 0. Since the test occurs in the natural world, this is the same as saying $\tilde{a}_i = a_i$.

For example, imagine you are leaving the movie theatre after a film, and consider using the emergency exit into the parking lot to avoid the line as people head out the normal way. You are concerned about setting off an alarm if you open the emergency exit, but would prefer going out that way otherwise. So let us set up an experiment to determine if the door is set up to an alarm. Your action profile is simple, to open the door or not. The state of the world is two states: there is an alarm, or not. The data generating process simply connects how the actions relate to the alarm going on or off, depending on the state of the world. The hypothesis would look like:



The experiment is a simple one, just open the door. If we open the door, then if there is an alarm it will ring, and if not, it will be silent. Now this experiment appears solid, but it is missing some important features. One point is we simply assumed the data generating process, such that if it is completely different, we did not learn anything. A possible difference is if it was a *silent* alarm. Thus, this example does not have construct validity, the most important property we will talk about for an experiment.

If an experiment has Construct validity, then $\tilde{\gamma}(\tilde{a}, \tilde{\theta}) = \hat{\omega}$, if the hypothesis is true. Thus, if running the experiment results in $\tilde{\omega} \neq \hat{\omega}$, then this falsifies the hypothesis.

Now **construct validity** means to measure what you mean to measure. This is an intuitive definition, but not a very useful one in a formal derivation. The following definition is formal, and allows an experiment to give conclusive results.

Definition 5 (Conclusive Results at observed ω)**.** *The results of an experiment are conclusive if their does not exist any alternative explanations that are consistent with the experimental controls, and the observed result. In terms of mathematics, this means that at some* $\omega \in \Omega$ *we get* $g(\tilde{h}, \omega|\hat{h}) = \{h\}$*, then you have conclusive results at* ω*, testing hypothesis* $\hat{h}$ *with experiment* $\tilde{h}$*.*

In fact, if in the above definition $h = \hat{h}$ then at ω you conclusively affirm the hypothesis, and would be considered testing the hypothesis $\hat{h}$. As a result, this is the formal definition of construct validity. In order to get conclusive results, I can establish one condition as necessary, that all non-hypothesized variables are held independent. This occurs because in a theory, variables not included do not affect the outcome, or in other words, in the hypothesis non-hypothesized variables are treated as independent by default. As a result, we will need to treat them as such in reality when running an experiment as well.

Assumption 1 (Independence of non-hypothesized variables[3])**.** *For all the variables of the action profile, the state of the world, or the data generating process, that are not specified in the hypothesis, must be held independent in the experiment to get conclusive results. In terms of mathematics, it means* $I(\hat{h}) = I(\tilde{h})$*.*

In other words, if for $\hat{a}_i$ is not hypothesized, then $\tilde{a}_i$ must be held independent. Another way of saying this, is that no matter what value $\tilde{a}_i$ has, it will not affect the observable result (otherwise it would be a bias, and allow alternative explanations). This is a hard property to satisfy.

[3] This will usually be referred to as (A1) in proofs to be auto-replaced in finished copy.

An example of the violation of the Independence of non-hypothesized variables in an experiment would be the early experiments on the prisoner's dilemma. In the prisoners dilemma, the dominant action for both players is to squeal, but in the early experiments this was not observed. The reason? They ran a repeated prisoner's dilemma, where other factors become relevant and agents might not optimize by squealing. In more recent experiments, the problems of social norms can be thought as a variable difficult to hold independent in experimental design, that can influence people's choices and reactions in an experiment.

Lemma 1. *If an experiment has conclusive results at observed ω and tests the hypothesis $\hat{h}$, then it satisfies independence of non-hypothesized variables at ω.*

Proof. Suppose that an experiment has conclusive results at ω, but does not have independence of non-hypothesized variables.

By the failure to meet the independence of non-hypothesized variables, this means there exists a variable not in the hypothesis that affects the experimental results. Without loss of generality, let this be $a_i \notin C(\hat{h}) \cap I(\tilde{h})$. Then by its failure to meet (A1), this means that $\exists x_1, x_2 \in [0,1]^2$ such that $\tilde{\gamma}(\tilde{a}_{-i}, \tilde{a}_i = x_1, \tilde{\theta}) \neq \tilde{\gamma}(\tilde{a}_{-i}, \tilde{a}_i = x_2, \tilde{\theta})$. Now since this fails (A1) at ω, then that without loss of generality, we can let $\tilde{\gamma}(\tilde{a}_{-i}, \tilde{a}_i = x_1, \tilde{\theta}) = \omega$, which means $\tilde{\gamma}(\tilde{a}_{-i}, \tilde{a}_i = x_2, \tilde{\theta}) \neq \omega$.

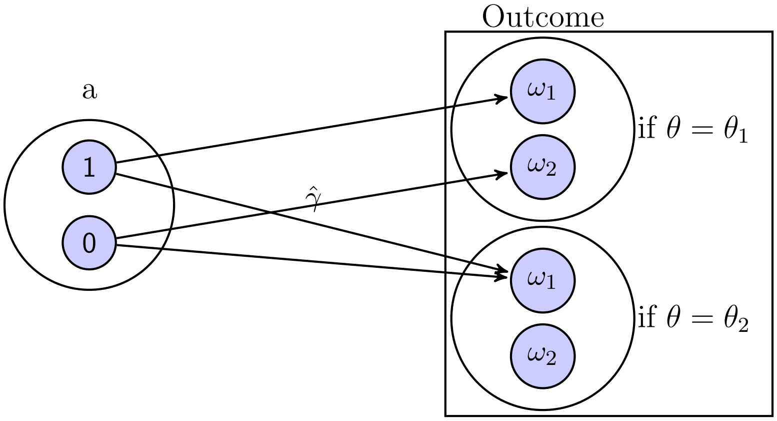
As a result, if ω occurs, then that means $g(\tilde{h}, \omega|\hat{h}) = \{h\}$ by conclusive results. However, by failure to meet (A1) at ω, then $h \neq \hat{h}$, because the hypothesis $\hat{h}$ does not explain a_i relationship and effect on the outcome. As a result, if $h = \hat{h}$, then $\hat{h}$ could not be the only consistent hypothesis, resulting in $g(\tilde{h}, \omega|\hat{h}) = \{\hat{h}, h_1\}$ where $h_1 \neq \hat{h}$, and thus this would contradict the assumed conclusive results at ω.

Therefore, we must proceed under the assumption $h \neq \hat{h}$. So this fails to affirm the hypothesis $\hat{h}$, but does it falsify it? The answer is no, because the hypothesis specified that $a_i \in I(\hat{h})$, which did not occur in our experimental controls. As a result, it is entirely possible that $\hat{h}$ is correct, but the effect of a_i dominates it in magnitude. In other words, we have admitted bias into our experiment, which allows us to falsify a true hypothesis. As a result, this experiment does not test the hypothesis $\hat{h}$, a contradiction. Therefore, we must have (A1) if we seek to test the hypothesis $\hat{h}$ with experiment $\tilde{h}$, and we get conclusive results at ω. □

Now a similar property to concern ourselves with is the possibility in the design of the experiment, we stop controlling too many variables, such that even with (A1), we do not get conclusive results because we have admitted too many possible hypothesis. Thus, It is possible to show that controlling all but one variable, and satisfying (A1) is sufficient to get a conclusive result, but it is not necessary.

Lemma 2 (Sufficiency for Conclusive Results)**.** *If an experiment controls all but one variable, and satisfies (A1), then the experiment has a conclusive result.*

A counter example to this considers the case where only the data-generating process is controlled, but the action profile and states of the world are not. In the counter example, we are able to get a conclusive result, and thus prevents the above sufficiency results from being necessary and sufficient.



In the above picture, it shows the counter example. In this case, if we run the experiment, and observe that we get ω_2, then we know two things *conclusively*, the agents action was

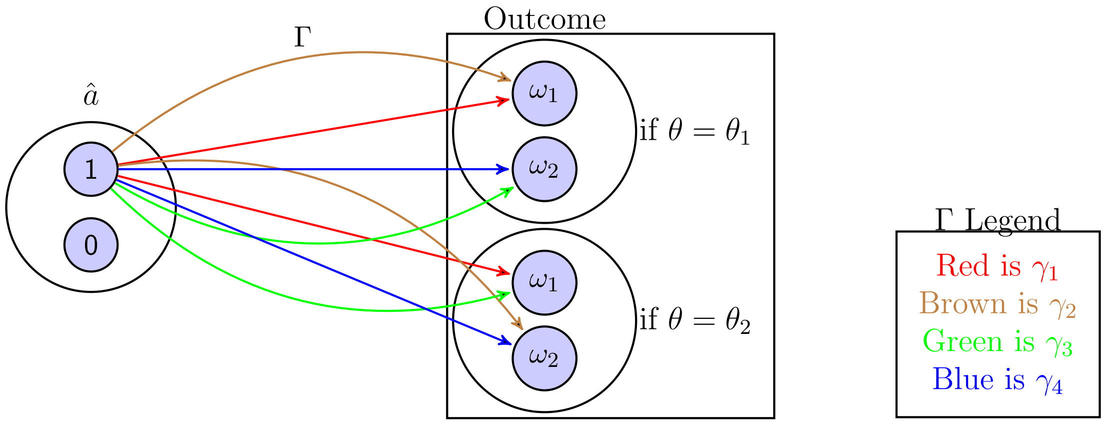
$a = 1$, and the state of the world is $\theta = \theta_1$. This is the only conclusive result in the experiment drawn above, but it still shows that it is not necessary to control all but one variable to get conclusive results. In this experiment, I controlled the data generating process γ, but not the actions chosen, a, nor the state of the world, θ. Thus, it is a fitting counterexample.

However, it is worth pointing out, if we are to get conclusive results, the counter example had to be done by controlling the data generating process. We can not get conclusive results by just controlling the action profile or state of the world by itself. This goes to show that the data generating process is the single most important thing to control in an experiment. This is good news for economists doing experiments, because we frequently control the data generating process in our field or lab experiments when we control the mechanism or institution, but bad news in general because the data generating process is usually not observable or easy to control fully.

An example to sketch out how conclusive results are unlikely without controls on the data generating process, or control on everything but the data generating process is as follows. In the example below we control the action profile by choosing $a = 1$, but not the state of the world, or the data generating process. As a result, the size of possible data generating process makes it impossible to get a single conclusive result.

Theorem 2 (Impossibility with Limited Controls)**.** *If you run an experiment, and you do not control the data generating process, and only partially control the action profile or the state of the world, then it is impossible to get conclusive results.*

Proof. Let us not control the data generating process, such that if $\gamma \in \Gamma$ then $\gamma \in \tilde{h}$. Without loss of Generality, suppose we do not control the state of the world, but control the action profile. Then, by the state of the world not being controlled and at best being partially controlled, this implies $\exists \theta_1, \theta_2 \in \tilde{h}$, and similarly, the set of possible observable outcomes must be at least two. As a result, we could always find the following sub-problem in everything mechanism problem with the above restrictions.



Now to see the complication, suppose we observe ω_1. Then the set of consistent hypothesis, or possible explanations, includes $g(\tilde{h}, \omega_1|\hat{h}) = \{(\gamma_1, 1, \theta_1), (\gamma_1, 1, \theta_2), (\gamma_2, 1, \theta_1), (\gamma_3, 1, \theta_2)\}$. As a result, no matter which hypothesis was the one we where testing for, we do not have conclusive results. The same goes for if we look at $a = 2$. As a result, since this is a sub-problem of every experiment which does not control the data generating process, and only partially controls the state of the world (or action profile), we can not get any conclusive results.

□

Thus, if we do not control the data generating process, we must control all the variables in the action profile and the state of the world to get conclusive results. As a result, controlling the data generating process is the most valuable of all of the controls. If we do not control the data generating process, then we most control the state of the world, and the action profile, which means we are conducting a lab experiment.

Corrolary 1 (Sufficiency with Control of Data Generating Process)**.** *If we control the data generating process, we can always test a hypothesis, but not all hypotheses.*

4 Testing something Indirectly

Previously, we assumed that only the outcomes are observable, but every variable of the hypothesis is controllable. This is different, for a lot of hypotheses we seek to test, we are testing something indirectly. In this case, we are usually testing a theory's conclusion to learn about the premise, rather than simply testing the premise directly. As a result, the interpretation of the results, and what they falsify or affirm, depend on what is being observed. If we observe the premise, and falsify the premise, the conclusion can still be true, but if we affirm the premise, the conclusion *has* to be true. The following table outlays the logic to determine whether you falsified or affirmed the premise or conclusion of a theory, depending on what was observed, and whether the experimental results showed the hypothesized results to be true or not. This is done assuming that the experiment has construct validity.

Observe Premise:

	Observe False	Observe True
Premise	Falsify	Affirm
Conclusion	Nothing	Affirm

Observe Conclusion:

	Observe False	Observe True
Premise	Falsify	Nothing
Conclusion	Falsify	Affirm

An example of a question we can answer when we test something we can observe something directly: "What happens if I do ________?"

An example of a question we can answer when we test something we can not observe directly: "Why does ________ do that when I do ________? "

This latter one is of particular interests to theorists, who seek to design theories that explain the purpose of certain structure in achieving certain predictions. Specifically, we want to know how subjects think and make choices, and what subjects is not observable.

Definition 6 (Influence Variable)**.** *A variable of an experiment which the experimenter can induce a specific value for indirectly, using other known and controlled variables.*

5 Interpreting Results with Optimal Theory Predictions

Now if we observe a result different from the hypothesized result, this can occur for three reasons:

1. The proof is wrong.

2. The experiment does not control properly to test the hypothesise.

3. The hypothesis is false in reality.

So the first reason is something you have to check by hand, and if it is found to be a consistent proof, then it must be the other two. We want to somehow disprove item two to leave only item three.

If the hypothesis is an optimal theory, as is common in economics, then we can determine if the results are biased under certain circumstances. If we look at the subjects actual payoffs,$\$^{actual}$, compared to the optimal payoffs,$\$^{optimal}$, then if:

$\$^{actual} > \$^{optimal}$ then we know that the problem is a lack of experimental controls without using any analysis.

$\$^{actual} < \$^{optimal}$ then we can not disprove a fault in the experimental controls, nor the possibility that the hypothesis is false without using analysis.

This is especially useful in going back to look over past experimental results. In the second outcome, if we can show that the experiment has construct validity, then we have shown that the tested hypothesis is false. However, with perfect experimental design at the outset, we would never observe actual payoffs greater than optimal payoffs, allowing us to conclude a hypothesis is false when $\$^{actual} < \$^{optimal}$ occurs (in fact, you should be able to conclude a specific premise is false).

6 Conclusion

This paper establishes a formal framework with which to design a perfect experiment from its outset. It gives a rigorous definition to construct validity, so we know when we have conclusive results, and that we are testing the hypothesis we seek to test. It also gives a rigorous definition to hypothesis, which turns out to be the integral definition for us to be able to also define what an experiment is. As a result, it allows us to transform the real life problem of designing an experiment, into a math problem.

The paper goes further, and establishes some necessary conditions and some sufficient conditions that an experiment's design must satisfy to test a hypothesis. It shows that it is necessary in an experiment hold all non-hypothesized variables independent, a difficult assumption that is likely the main culprit in most experiments failure to be perfect. Yet, even if we can not make a perfect experiment, it allows us to identify it, thus allowing us to get as close as possible. For the sufficient results, I showed that the data generating process is the most important variables to control in the experiment, and if we can do so, we can always test a hypothesis.

Finally, I begin introducing content on extending these results. Important assumptions to be dropped is that we can control everything, and that we can only observe the outcome. The results are preliminary, but are an important consideration because we usually can not control all variables in the action profile, such as agent's beliefs, or all the variables in the state of the world, even in a lab setting. Finally, I extend some of the results to a more immediately applicable case where we are testing optimal predictions that are common with economic theory. It gives an easy rule of thumb to find out if we are missing experimental controls, but shows we still need to establish construct validity if we want to conclusive falsify a hypothesis.

References

Bailey, Rosemary A, *Design of comparative experiments*, Vol. 25, Cambridge University Press, 2008.

Campbell, D.T., “Convergent and discriminant validation by the multitrait-multimethod matrix,” *Psychological Bulletin*, 1959, *56*, 81–105.

Hinkelmann, Klaus and O Kempthorne, *Design and Analysis of Experiments: Volume 1, Introduction to Experimental Design*, Wiley, 2008.

Roth, Alvin E and John Henry Kagel, *The handbook of experimental economics*, Vol. 1, Princeton university press Princeton, 1995.