Planning, Implementing, and Reporting Quantitative Research in Education: A User's Guide

R. Brent Stansfield, Ph.D. (rbstansfield@umich.edu)

Larry D. Gruppen, Ph.D. (lgruppen@umich.edu)

University of Michigan

Introduction

Dr. Marcus is a mechanical engineer who, for three years, has been teaching a regular graduate level seminar on ecologically-minded structural design. Each semester, about a dozen graduate students take his course. They tend to do most of the readings, they attend most of his classes, and they perform fairly well on his tests. He's usually disappointed with the sophistication of the in-class discussions, but some semesters it's quite good. He requires each student to generate an original independent project, and he's surprised by the variety of ideas. Some students design something to improve the efficiency of a specific structure or process. Some write a literature review on some topic. Most of them choose projects directly related to their own work outside the class.

This teaching is satisfying, but Dr. Marcus wants the class to be better. He wonders whether the course is improving his students' skills. And if not, he wonders what changes he should make to the class. Should it be bigger or smaller? More lecture-focused or more discussion-focused? Should students be encouraged to do different types of projects? He knows he could conduct some educational research to answer these questions, but isn't sure how to proceed. He has some intuitions and even some strong opinions about how to improve the class, but he's empirically-minded enough to doubt the validity of even seemingly obvious solutions.

He meets with the mechanical engineering program director, Dr. Schneider, who is also planning a large scale evaluation of the departmental curriculum. She's concerned about a recent surge of complaints from students that they have too many required classes and not enough exposure to practical, applied solutions to real situations. She's aware of some other programs that have adopted a project-based learning curriculum, but isn't certain whether that kind of curriculum change would be worth the effort and cost.

Their interests are more than institutional. Dr. Marcus wants to improve his class, but he really wants to know *why* those improvements work, because without understanding the mechanisms that cause those improvements, further improvements will be just as

difficult. Dr. Schneider, similarly, wants to generate some statistics that distinguish her program from her competitors', but she also wants to make sure that those statistics represent important and true aspects of her department. She wants to adjust her curriculum, not for political or financial reasons, but for educational ones. In short, they are both empiricists, and know that true understanding cannot be achieved without experimentation. Together, they decide to plan a research program that addresses their institutional needs, improves their pedagogy, and contributes valid and useful knowledge to the engineering education literature.

Every strong research program requires rigorous testing of hypotheses generated from research questions in order to build useful causal models, and ultimately, scientific theories. No single study, no matter how carefully done, will generate a definitive answer to a research question. Only a large number of studies, using a variety of methods, addressing the same questions in different ways can falsify or support a theory. Drs. Marcus and Schneider must first conduct exploratory research to identify the causal models they want to test, develop a set of hypotheses crucial to those models, conduct experiments to test those hypotheses, and finally publish their results in peer-reviewed journals to disseminate their findings.

This document is designed to help education researchers plan similar research programs. Sections are organized by the chronology for developing a series of studies. First, we will discuss how exploratory research can suggest theories and causal models, but cannot test them. Second, we will describe testable research hypotheses and how to conduct adequately-powered, appropriately designed experiments. Third, we will discuss the treatment of data and how to choose the most appropriate statistical tests. Fourth, we will discuss how to interpret results and prepare them for publication.

There are few correct answers in research design and analysis: many methods can be used to address the same problem. Methodologists disagree, sometimes vehemently, about the appropriate use and interpretation of certain designs and statistics. The process of data analysis and the conventions for reporting results are constantly changing as scientists and editors adjust to new methodologies and fields of

study. However, the basic premises of the scientific method are constant. Throughout the document, several of these points are highlighted as cardinal rules. The cardinal rules in this document describe widely-agreed upon truths of research, some of which are commonly misunderstood by novice researchers.

Our focus is on quantitative research design, so qualitative methods and analysis will not be described. Measurement, the art of translating real world concepts into observable quantities, is an enormous field, too complicated to be treated thoroughly in this document. We will address some aspects of measurement important to research design, but will leave much of the theory of measurement and operationalization to other authors.

This document is intended to be a guide to help a researcher plan, implement, and report empirical studies that are informative and useful to a broader scientific audience. We assume the reader has some familiarity with statistics and some experience using a statistical software package or two, but we use no math in our descriptions. A wide variety of complex and obscure methodological and statistical methods is available to researchers, but the concepts described in this document are common, widely accepted, and used consistently in the social science research literature.

Exploratory Studies and the Development of Theories

Exploratory Research

Dr. Schneider is unsure about what sorts of experiments or studies she should conduct to understand and improve her program. Would a problem-based curriculum lead to more or less student satisfaction? To more or less efficient use of faculty time? Would students become better thinkers or planners? What aspects of the program are important to measure? What faculty, student, administrative behaviors indicate what they are really thinking? How should she measure the effects of her curricular reform?

In short, she is not being guided by any theory. A scientific theory is a broad understanding of how a real-world system works. A theory defines constructs and the ways in which those constructs interact. Imagine trying to study the physical sciences

without an atomic theory stating that matter is comprised of tiny particles, of which there are only a few dozen types. Would you focus on the properties of growing things versus inert things? Red things versus green things? Big things versus small? There are an infinite number of possible studies and experiments which would yield little or no information about the true nature of matter. The theory provides a set of constraints within which to think and a set of predictions that can be verified or falsified. The atomic theory did not arise in its entirety all at once, but was a philosophical stance for thousands of years before taking its modern form over the past couple centuries.

Since Dr. Schneider has no working theory about what causes her students to learn or not learn as they do, she decides to start her research program with an exploratory study. Later, she will use the results she obtains to form causal models, which will allow her to make predictions she can test with focused experiments. Only then, by repeatedly improving her ability to predict learning outcomes by using these causal models, will she begin to have a scientific theory of learning. But the first step must always be careful observation of the phenomena to be explained: an exploratory study.

The point of an exploratory study is to cast a wide net over a poorly understood problem and catch as many possible relationships between aspects of that problem. It's important to remember, however, that a net will always catch things you don't want along with the things you do. When Dr. Schneider and Dr. Marcus meet to list all the aspects of the program that they suspect may be affected by Dr. Marcus' course, they decide to start giving questionnaires to his students to measure their attitudes about environmentalism, learning styles, and Dr. Marcus' teaching. They want to know the students' plans for specialization within engineering, their expectations for career success, their self-perceived achievement, and personal commitment to environmentally-responsible design. They plan to collect the students' admissions data: undergraduate grade point averages both in general and for science-related courses, interview scores, graduate record examination scores. They plan to track these students' academic performance through the class, the program, and ultimately their job placement when they leave.

Some of these measures are existing data: undergraduate grade point averages are numbers that have already been computed, have meanings that are already well-understood, and are organized in a database kept by the admissions department. Most of these measures, however, are not yet usable. Some require a little research: students' learning styles is an area of education research with a long history, and many measures have been published, critiqued, and revised. Still others have yet to be built: students' personal commitment to environmentally-responsible design is probably too specialized a concept to have been measured before. The researchers may have to develop a questionnaire to measure that construct.

Before collecting any data, they will need to obtain permission to conduct the study from their Institutional Review Board (IRB). IRBs exist to protect the rights of human participants as subjects in scientific studies. Most academic institutions have at least one IRB[1] and every IRB can act somewhat independently from others. Since most exploratory studies do not involve manipulation or deception, IRBs tend to approve them quickly. Moreover, since most of this data collection is part of the normal educational process, IRBs will generally grant an educational exemption—a less restrictive set of responsibilities for the researchers. In any case, the researchers should identify and communicate with the IRB responsible for research within their department before collecting any data or performing any analyses.

Once the measures have been identified and a feasible, IRB-approved plan for collecting the data is written, the researchers need to decide how many students will be adequate for their purpose. This is a common question among inexperienced researchers and will be treated in greater detail in the Hypothesis Testing section below. In exploratory studies, the goal is simply to gather as much information as practical. Our first cardinal rule applies to exploratory studies:

Cardinal rule # 1: Exploratory studies can suggest, but cannot test theories.

Let's assume the researchers have collected their data and have computed a full correlation matrix, measuring how strongly each of their measures is related to each of the others. Studying this matrix, they may find that students' class participation correlates highly with self-perceived achievement. They may also find that rigorous study habits correlate highly with higher test scores. But they may be surprised to find that these two correlations are independent—that is, that neither class participation nor self-perceived achievement correlates with rigorous study habits and higher test scores. They would reasonably conclude that students have a quality they could call "confidence" which leads to class participation and self-perceived achievement, but that this "confidence" does not cause academic discipline or aptitude.

However, because this conclusion was created to explain unexpected results from an exploratory study, it must be tentative. The researchers cannot know whether this correlation pattern would happen again with a different set of students. In other words, these correlations may be accidental or they may reflect some underlying truth, but an exploratory study can never distinguish between these possibilities.

In order to be reasonably sure that student "confidence" is a real quality and that it does not actually cause studying or actual achievement, the researchers need to look for these patterns again in another group of students. If this second study finds similar correlations, it can be taken as evidence for the hypothesis. Similarly, if the second study does not find similar correlations, it can (assuming it has enough power) be taken as evidence against the hypothesis. That second study is an attempt to replicate a specific correlation. Replication includes an element of prediction and prediction is the basis of all scientific theories. A theory that makes no predictions is not empirically falsifiable and is therefore not scientific. A theory that makes consistently wrong predictions is an incorrect theory and must be revised or discarded.

Despite their inability to test hypotheses, exploratory studies are powerful and useful methods for developing theories and causal models and for defining hypotheses when the number of relationships between measures is very large. It is important to learn to find useful hypotheses. Different theories may be able to explain the same patterns of data, but there will always be a circumstance for which the theories make different

predictions. These differing predictions are testable hypotheses: the opposing theorists can agree that whoever's prediction comes true has the better theory.

For instance, Dr. Marcus may suspect that student aptitude is a stable trait that predicts students' ultimate job placement success; some students have it, and some don't. Dr. Schneider may have a different theory, believing that all their students have enormous potential but have different learning styles. She may believe that the system is biased towards visual learners, so verbal learners test poorly and fail to find jobs, but they would have learned more and gotten better jobs if the program had addressed their learning needs.

When the data from their exploratory arrives and they find that students who perform well on their undergraduate calculus tests tend to earn high post-degree salaries, their conclusions will differ. Dr. Marcus will say that those students are simply smarter and more studious, while Dr. Schneider will say that non-mathematical aptitudes such as environmental awareness are undervalued by their program and their field as a whole. Thus, when discussing admissions policy, Dr. Marcus will propose admitting only students with superior calculus scores. Dr. Schneider would disagree vehemently. The observed correlations cannot say who is right.

A useful hypothesis would address the differences between their underlying theories. Do verbal learners have equivalent aptitude as visual learners, when aptitude is defined with less bias toward mere technical skill? Would grading students based on their sensitivity to environmental and social issues reduce the correlation of undergraduate calculus scores with post-graduate earnings? Dr. Schneider's theory predicts "yes," whereas Dr. Marcus' theory predicts "no." These research questions raise differing hypotheses, which can be empirically tested.

The problem in exploratory studies is that people are very good at explaining observed results after the fact (*post hoc*). It is always tempting to react to the results of an exploratory study as if those results were obvious and inevitable. It is important to remember all the possible correlations that *did not* occur in the data and the likelihood that the observed correlations were entirely accidental. Changing your theory to incorporate a finding that was simply a chance occurrence is an error. It is such a

troublesome and common error that researchers call it a Type I error: Type I error is the conclusion that an effect exists when in fact it does not.

> Cardinal rule #2: Regard your results with skepticism. Type I error results from excess credulity.

People are prone to erroneous beliefs. A gambler who wears his new hat to the casino and wins a lot of money may decide the hat is lucky. He will start wearing the hat more. He will attribute wins to the hat and attribute his losses to other causes. What he will probably not do is a planned experiment: flip a coin each day to determine whether or not he will wear his hat and track his winnings across a few months. He is not skeptical enough of his own beliefs to bother with that. Planned experiments require discipline, patience, and resources. Data from planned experiments is valuable. Observation of things as they naturally happen is much easier and observational data is cheap. We are surrounded by observational data.

Unfortunately, a planned experiment is not always possible, practical, or ethical. If you hypothesize that men and women are predisposed to different learning styles, you will have to do an observational study because you can't randomly assign genders to your subjects. If you hypothesize that a complete reorganization of your curriculum would improve student performance dramatically, you will have to implement that curriculum for at least some of your students and allow those students to choose whether to switch back to the more conventional curriculum if they want.

In an experiment, the experimenter wants control. He wants to manipulate only certain things—whether or not the gambler wears the "lucky" hat, or which students attend the new curriculum, etc.—and observe the effects. He wants to be able to claim that those effects are due to the things he manipulated and to no other causes. Therefore, all other possible causes for those effects need to be as similar as possible between the experimental conditions, whether by direct control or by randomization. Obtaining this control is what makes experiments expensive.

In an observational design, the experimenter relinquishes control. Since the experimenter is not directly controlling anything, the distinction between cause and effect is blurred and causal assertions cannot be made definitively. This is a crucial cardinal rule of research design:

Cardinal rule #3 : Correlation does not indicate causation

If a researcher finds that students who said they used a computer-based learning resource tended to score higher than others in the class on a knowledge test, he cannot definitively conclude that the computer-based resource is a useful learning tool. It may be the case that only motivated students chose to use the resource; they would have gotten the highest test scores anyway. Perhaps performing well on the test caused students to overstate how thoroughly they prepared. The co-occurrence of high scores and self-reported computer studying is merely a correlation. In order to test the causality that the learning tool causes better performance, the researcher would have to assert control—assigning some students to a computer-using group and others to a non-computer-using group. That assignment is manipulation and that study is an experiment. In that case, the experimenter will want to make sure that the two groups are as evenly matched academically as possible, that they have comparable computer skills, comparable ages, comparable gender ratios, etc. The more evenly matched the two groups are on any factor that might influence their test scores, the more strongly the experimenter can claim that a difference in the groups' test scores is due to the computer-based resource.

It is true, however, that *if* a causal connection exists between two constructs, then there *should* be a correlation between them. Therefore, if the researcher's measures are reliable and valid, his sample size large enough, his sample representative enough, and he finds no correlation, there is likely no causation. For this reason, correlational designs can be an efficient way of ruling out many causal models.

When Dr. Schneider and Dr. Marcus have collected their data and started examining the various correlations and group differences in that data, they will start to infer some

causal models and start to build a theory about how their students learn within their curriculum. This theory will make predictions about other correlations and group differences that they can look for in this exploratory dataset. However, they will have to keep in mind that many of the correlations they find are due to chance, and that if they were to re-run the entire study, a somewhat different pattern would emerge. They will have to test their theory by experiment, and they certainly will have to revise the theory in light of unexpected experimental results.

Theories generate research hypotheses, which are tested by research. Research hypotheses can never be more important than the theories that raise them, and answers can never be better than the hypotheses they address. Exploratory research helps generate research hypotheses, but only planned studies can test them. Before beginning any planned study, it is vital to have clearly defined research hypotheses. Below, we outline the nature of theories, causal models, and hypotheses.

Theories, Causal Models, and Hypotheses

A scientific theory is a broad understanding of how a real-world system works. The theory defines constructs and the rules under which those constructs interact. For instance, Newtonian physics proposes that an object has a property called “inertia” and its surroundings exert a force on it called “friction.” Even though these properties are ascribed to concrete objects, the properties themselves are ephemeral. Science is agnostic as to whether these constructs actually exist in the world: truth is a philosophical, not a scientific, problem. Psychological constructs such as “mood” or “personality” or “motivation” are similarly ephemeral and may not map directly on to any concrete object or system—science doesn’t care. What science cares about is the accuracy of predictions the theory can make. A theory is only a theory if it makes predictions that may or may not come true. If a theory’s predictions do not reliably come true, scientists will reject the theory by modifying it so it makes better predictions or replacing it with a new theory entirely. Every theory, no matter how successful to date, will likely prove faulty in the future. A theory can only be improved using continuous rigorous testing by skeptical scientists eager, or at least willing, to reject it.

Dr. Schneider and Dr. Marcus have invested the time and energy to gather and analyze their exploratory data and pored over the results long enough to generate an appealing theory. They will understandably be emotionally and intellectually invested in this theory and may find themselves more excited about proving it true than about proving it false. They will be wise to remember our next cardinal rule:

> Cardinal rule #4: Theories can never be proven true; they can only fail to be proven false.

Useful research programs try to falsify theories. Successful researchers are skeptical of their own results and relentlessly seek alternative explanations. Many obvious and attractive statements—the earth is at the center of the solar system, the universe is infinitely large, heavy things fall faster than light things, animals pass acquired physical traits to their offspring—make predictions which do not come true in the real world. Ancient mathematicians worked hard, and fruitlessly, to explain why the visible planets change their speed and direction as they move across the night sky, ostensibly in their orbit around the Earth. An alternative explanation--that those planets orbit the sun and not the Earth--explains the planets' behavior much more parsimoniously. Despite the non-intuitiveness (and even heresy) of such a model, its superior predictive power is undeniable.

Every theory contains several causal models. For instance, Newtonian physics proposes that friction reduces inertia by converting kinetic energy into heat energy. The constructs and causality of this model are important to the understanding of matter: something called "friction" *causes* something called "energy" to change into heat by *causing* a reduction in something called "inertia." Without this structure, the model cannot provide explanation. A researcher who merely points at the complexity of the numerous relationships between the size, color, temperature, roughness, speed, and shininess of various objects sliding down various surfaces is not providing explanation. Without explanatory power, a model is useless.

Figure 1 is a diagram of a simple causal model where two causes (A and B) produce some effect. The physics causal model above has a similar structure in which the inertia of an object (cause A) and the friction it experiences (cause B) determine the amount of heat energy (effect). This model is not comprehensive: other factors determine the amount of heat energy as well. This causal model is part of the larger Newtonian physics and describes only one particular causal relationship. An experiment to demonstrate how friction and inertia cause heat energy will have to control for all other possible sources and drains of heat.

An educational researcher may draw a similar causal model. For instance, he may postulate that the social cohesion of a class of students (cause A) and their access to study materials (cause B) affect the amount of spontaneous collaborative learning (effect). Again, there are certainly other factors determining whether students form study groups, but this causal model is an empirically testable assertion. The researcher can manipulate the social cohesion of a class by encouraging more inter-student communication or asking the class to work as a team on some problem. The researcher can manipulate the students' access to the study materials. The causal model makes some predictions about how these manipulations should affect spontaneous collaborative learning. Such causal assertions are necessary for coherent theories and such predictions are necessary for scientific hypothesis testing.

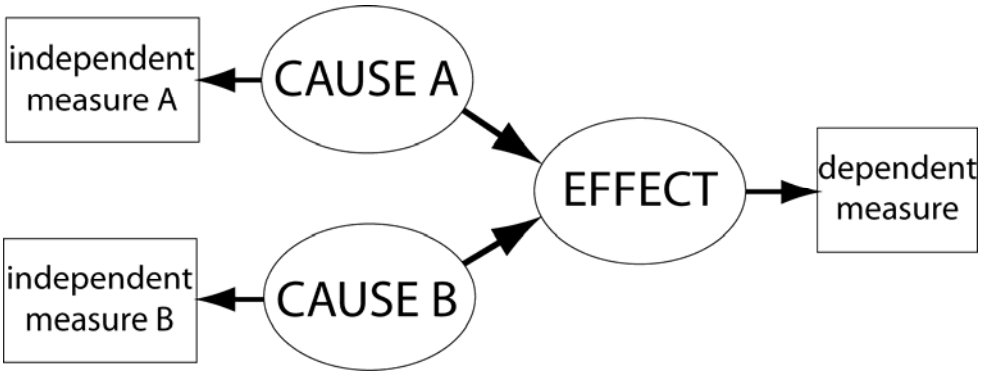

Figure 1: Schematic of a simple causal model in which two causes, measured by independent measures, produce an effect, measured by a dependent measure.

Constructs (ovals) exert causality (directional arrows) on each other and on the measures (boxes) a researcher uses to observe them.

Hypothesis Testing

Scientists should be deeply skeptical. Anyone asserting a scientific explanation bears an enormous burden of proof. No matter how well-designed and well-executed a planned study may be, there is always room for doubt. There are always confounds. There is always measurement error. There is always the possibility of a fluke result. When a researcher submits a research manuscript for publication, peer reviewers will always express these doubts and point out these flaws. The scientific community aspires to be an extremely tight filter, letting in only the more hardened truths. It is important to expect, and even enjoy, tough scrutiny when presenting a research finding to peers.

When a theory's predictions prove true consistently, that skepticism softens. For each individual, the point of acceptance of a theory is a subjective decision, but it is never complete. On any given day, any given physicist may find some piece of data which casts doubt on the law of gravity. If that result can be replicated to the satisfaction of enough other physicists, the theory of gravity as we know it would be discarded and rewritten. No scientific theory, no matter how credible, historical, or central to our understanding of nature is exempt from this rule: science is the constant endeavor to prove itself wrong.

Once researchers have some hypotheses, they can conduct hypothesis-testing studies. As opposed to exploratory studies, hypothesis-testing studies can prove a theory's predictions wrong, thereby disproving the theory itself. This is possible because the hypothesis exists before the study is conducted, so the study can be designed to falsify it. Good experiments are designed to falsify hypotheses. The failure to find a predicted relationship is not falsification: anyone can fail to find something that actually exists by looking in the wrong place, or not looking carefully. If Dr. Schneider believes

that students with poorer math skills can achieve superior job placement if the curriculum focuses on and rewards non-technical aspects of engineering, she is making a prediction that Dr. Marcus' theory (that the technically-apt students are simply better) cannot handle. If she finds empirical evidence to support her prediction, the power of the finding comes less from the support it lends her theory (after all, there may be other reasons why she found that effect) but rather more from the blow it deals to Dr. Marcus' theory. If the hypothesis endures an attempt to falsify it, the researcher may conclude that the hypothesis—and by extension, the theory that generated it—is valid (at least to some extent).

Some important caveats apply:

Cardinal rule #5 : No single study can prove a hypothesis false.

Even hypothesis-testing studies require replication for verification. If a causal model is true, it should be observable repeatedly. A researcher repeating the experiment using the same methods as the original study (an exact replication) should find the same results. If he does, he can believe that the results of the original study were not merely accidental. However, he might still believe in an alternative explanation for those results. For this reason, an even stronger replication will use different methods. If the causal model predicts results in a wide range of experimental situations, alternative explanations become less likely. A climatologist may be stunned to observe that the sun does not rise every morning. He may claim that our model of the solar system is wrong. However, a skeptical colleague may remind him that his data comes from the South Pole and suggest that he try to replicate his findings at the equator. Similarly, Dr. Schneider's initial adoption of a problem-based curriculum may receive rave reviews from faculty in the first year. However, without an attempt to replicate the finding in the next year, she cannot know whether the faculty are happy about the new curriculum *per se*, or are merely happy to see some sort of change. If faculty attitudes drop back to normal levels in the second year of the new curriculum, the original result will not have

replicated. The prediction that faculty prefer the new curriculum will have failed and an alternative explanation will have succeeded.

How to design a hypothesis-testing study

True experimental designs are the strongest tests of an hypothesis. The only way to prove causality is to: (1) intentionally change something and (2) show that something else is affected. This is a true experiment. In a true experimental design, the researcher actively manipulates one or more independent variables and measures any changes in one or more dependent variables. If the experimenter renders inoperative most other possible influences on the dependent measures, the experiment is well-controlled. In a well-controlled experiment, observed differences in the dependent measure can be reasonably attributed to the manipulation. Well-controlled experiments are very difficult to conduct, especially in the social sciences where many causal factors are beyond any researcher's control. Nonetheless, it is important to identify and control as much as possible when testing a hypothesis.

A true experimental design always contains a comparison between different values of the independent variables. The most powerful comparison is between an experimental condition and a control condition. Ideally, the only difference between subjects in the experimental and the control conditions is the independent variable. When Dr. Schneider measures the impact of the new curriculum, she compares faculty reaction to the new curriculum (the experimental condition) to that of the old curriculum (the control condition). Her choice of control condition is important: if she uses attitude measures taken in a year prior to the adoption of the new curriculum, she can only make strong claims if that year was the same as the current year in all aspects except the curriculum—the same faculty, same political and economic conditions, same student aptitude, or anything else that might impact faculty attitudes in general. If she uses a retrospective control condition, viz. asking this year's faculty to rate both curricula simultaneously, she can only make strong claims if faculty can assess accurately their prior attitudes. Again, these ideals are generally impossible in social science research, making replication by different methods very important.

Well-controlled studies are extremely rare because not everything can be controlled. To the extent that a researcher accepts a lack of control, she weakens her case that the manipulation is the cause of the observed effect. Some factors, such as students' gender, intelligence, socio-economic background, and ethnicity, simply cannot be manipulated. Some factors, such as the grades and career advice given to students, cannot be manipulated without ethical considerations of interfering with students' rights to an honest, good faith education. The manipulation of other factors, such as access to certain study aids or curricular tools, may be possible and ethical, but only if students agree to participation in that manipulation. It is very important when planning any study, particularly one where manipulation is made in the name of hypothesis-testing, to get approval from IRB. Practical and ethical realities supersede scientists' desire for control: the art of designing a good hypothesis-testing study is in grappling with this loss of control.

To improve a hypothesis-testing study design, the researcher must focus on both the causes and the effects. When focusing on the causes, the researcher must choose the best independent measures to manipulate, be able to demonstrate the strength and success of the manipulation, and identify and control for other factors which might yield alternative explanations. When focusing on the effects, the researcher must choose the best dependent measure, be able to demonstrate its validity and reliability, and show that enough data has been collected to treat the experimental results as definitive. Some advice on these topics is given below.

Focus on causes

Choose focused independent measures and manipulate them strongly: If Dr. Marcus is interested in improving his lecture style, he has to decide what aspect of that style to manipulate. If his hypothesis is that encouraging student participation during lectures will improve their learning, he must decide exactly how to manipulate that aspect, how strongly to manipulate it, and how best to make sure the manipulation worked. He may decide that for half his lectures he will ask an open-ended question and remain silent for at least 10 minutes to let the students try to generate answers and for the other half of his lectures he won't ask any questions. This is a very strong

manipulation and will probably yield some observable differences in student reaction, if not their learning. But he will still need to make sure that students participate: does the class remain silent during those 10 minutes? If so, the manipulation has probably failed. Who speaks? Perhaps students who don't speak up should be analyzed separately. The independent measure should strongly reflect the cause in the causal model: does Dr. Marcus expect each student's participation level to improve that student's learning, or that any student participation will benefit the whole class?

Control for confounds in the design, or in the analysis: Dr. Marcus may not be entirely free to choose which lectures to use for experimental or control conditions. Some topics may contain too much book-knowledge to lend themselves to open-ended questions. Some topics may contain too little factual content to warrant a purely didactic lecture. It may be that students do better on such non-fact-rich topics anyway. If there are too many non-fact-rich topics in his experimental condition, he would be wrong to ascribe higher student learning in the experimental condition to lecture style. This alternative explanation is a confound—it cannot be ruled out using the experimental design in question. To control for the confound, Dr. Marcus has two basic options: (1) make sure that topic difficulty is equally represented in the experimental and control conditions or (2) use a statistical method to handicap learning scores on easier topics when making the comparison between lecture styles.

The first option is superior if it is practicable. Dr. Marcus can categorize each lecture topic as book-knowledge-heavy or subjective and include an equal number of each in each condition. If he has a large enough set of lectures, he can make sure he randomly assigns each lecture to one condition or the other and force himself to apply the assigned lecture style to that topic. The random assignment is important and worthy of another cardinal rule:

Cardinal rule #6: Randomization ensures unbiased results.

By randomizing assignment of lecture topic to condition, Dr. Marcus is avoiding creating a confound like the one above, in which his preference to use a certain lecture

style for a certain topic confounds topic difficulty with lecture style. He may flip a coin, use a computer program, pull numbers out of a hat, or perform any other non-biased assignment method. Note that the randomization can not ensure that a confound will not occur: confounds also happen by accident. But the randomization will free Dr. Marcus of the fear that he has biased the results with his own experiences and expectations.

The second option is less powerful, requires some statistical *savoir faire* and requires some assumptions about the causal model being tested, but when confounds cannot be practically resolved, it is the only method. To use this method, Dr. Marcus can use student performance from previous years to estimate the difficulty of each lecture topic, giving each topic a numerical difficulty score. He might assign these scores subjectively after reviewing student performance over several years or he may use direct measures, such as average student test performance on previous final exam items covering each topic. Now he is free to assign lecture topics to experimental or control condition as non-randomly as he prefers. He can then use a statistical method called Analysis of Covariance (ANCOVA) to model student learning by topic difficulty and, separately, by lecture style. The ANCOVA doesn't resolve the confound—easier lectures topics are still overrepresented in the experimental condition—but it gives an indication whether student learning on experimental lecture-style topics is better than would be expected by the difficulty of those topics alone.

If Dr. Marcus has to use a lot of this sort of statistical control in the analysis of his results, he is not conducting a true experiment, but rather a quasi-experiment. Quasi-experiments have some of the weakness of exploratory research—causal assertions are tenuous. Unless Dr. Marcus actively assigns some easy topics to the didactic lecture style condition, he cannot definitively ascribe better student performance on those topics to his lecture style. ANCOVA can provide evidence, but it cannot replace experimental control. This rule is true in general:

> Cardinal rule #7: The better the experimental design, the simpler the statistical analysis and the stronger the causal assertion.

While statistical methods—ANCOVA, multiple regression, hierarchical linear modeling—allow the researcher to control for confounds during analysis, they cannot substitute for an experiment which controls them in the experimental design. Quasi-experiments are necessary when independent measures cannot be controlled for practical or ethical reasons, but should be avoided when possible.

Keep planned tests for interactions focused and simple: It common to include several independent measures in an experiment. Reality is complicated and single-cause models are not nearly as predictive (nor interesting) as multiple-cause models. Dr. Marcus may wonder if the males and females in his class react differently to the lecture style change. He may wonder if encouraging student participation would decrease learning for some topics and increase it for others. These questions raise hypotheses that can only be tested by manipulating or observing multiple independent measures: lecture style, topic, student gender.

If students in general learn more in student-participation lectures no matter their gender or the topic in question, that effect will show up as a main effect of lecture style. A main effect is the direct cause of one independent measure on the dependent measure. Main effects are easy to test, easy to graph, easy to explain, and easy to understand. Interactions occur when the effect of one cause is mediated by another. If females learn more on student-participation lectures and males learn more on didactic lectures, this pattern will show up as a two-way interaction between lecture style and gender.

Interactions can mask main effects. Figure 2 is a graph of a hypothetical two-way interaction between the amount of reading assigned to a student and that student's year in the program. In this interaction, more reading increases first year students' attitudes and decreases fourth year students' attitudes. If the researcher hadn't thought about considering the students' years, he would simply calculate the average of student attitudes with few assigned readings and with many assigned readings. As the dashed line in Figure 2 shows, he would find no main effect. If he happened to have only a few fourth year students in a large class of first years, he would likely find a positive main

effect. He might notice that the students who have worse attitudes with more assigned readings tend to be older and he might postulate the hypothesis of the interaction.

It is possible to find three-, four-, five-way interactions. Each independent measure a researcher adds to the experimental design can interact with each of the others and with each of the others' interactions. For instance, the interaction shown in Figure 2 might only be true of males while females' attitudes are not influenced by the amount of reading, no matter their year. That would be a three-way interaction between amount of reading, student year, and student gender. This three-way interaction might be true for mechanical engineers, but the gender effect might be the reverse for electrical engineers: a four-way interaction. These higher-level interactions are not only difficult to graph and describe, they are difficult to explain parsimoniously. When a researcher finds three-way or higher interactions, he should try to find different ways to explain the effect. Given the four-way interaction described above, he might ask himself: what do male mechanical engineers and female electrical engineers have in common such that they might react to reading assignments the same way? He should consider the question in general and for the specific students in the analysis in question. An alternative explanation could reduce the complexity greatly: perhaps students like to be assigned readings for topics they find confusing and hate to be assigned readings for topics they grasp intuitively. This hypothesis is testable using a main effect, or at most a two-way interaction, and further experiments can test for it. If it proves predictive of student attitudes, it is more useful than a model using the four-way interaction.

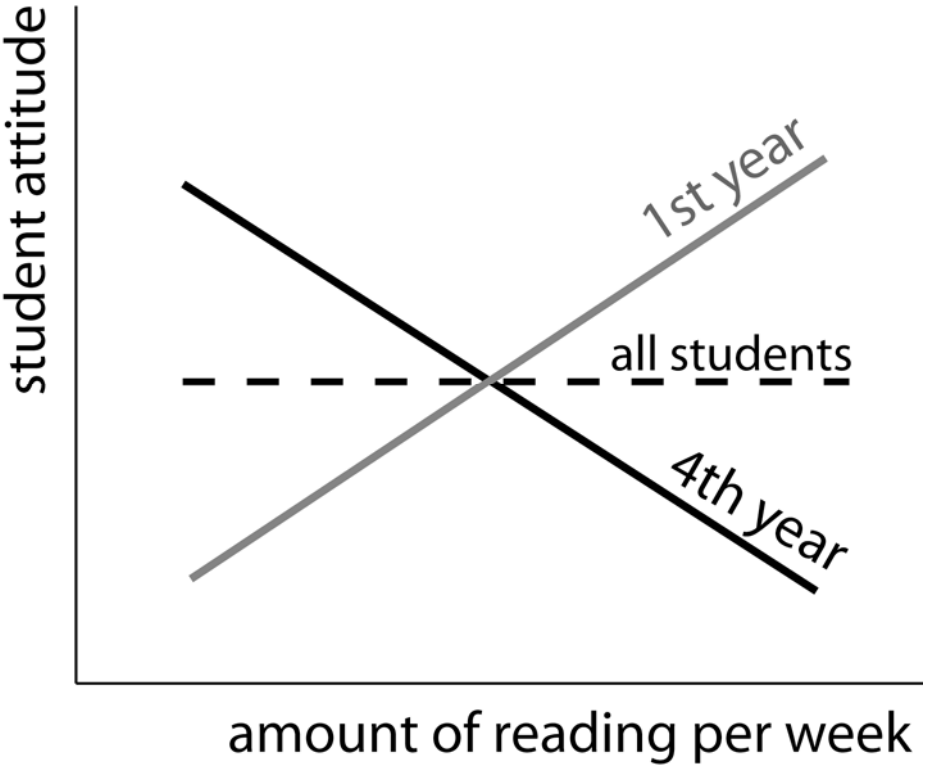

Figure 2: An interaction effect on student attitude between student year (1st year and 4th year) and the amount of required reading per week. The interaction effect exists despite the lack of main effects

Use the appropriate type of independent measure: There are three types of measures: continuous, ordinal, and categorical. Continuous measures are numerical and scalar, measuring an amount or degree of some concept. Height in inches, I. Q. score, and income are continuous measures and any individual at any given time has a measurable quantity on each of these scales. Ordinal measures are ordered, perhaps even numerical, but not scalar. Body mass index category (underweight, normal weight, overweight, obese), highest academic degree earned, and tax bracket are ordinal measures. These measures represent similar concepts to the continuous measures listed above, but the ordinal measures group individuals into classes, while the continuous measures do not. Categorical measures are neither ordered nor scalar.

Personality type, country of birth, and profession are categorical measures. Like ordinal measures, categorical measures group individuals into classes, but no internally consistent ordering of the classes is possible

Some factors can only be measured or manipulated by categorical measures. Gender, specialty, in- or out-of-state status, pass or failure of a particular class: each of these is inherently categorical. Many causal factors could be considered continuous or categorical, however, and the researcher needs to take care to use the most appropriate manipulation. Dr. Marcus has conceptualized lecture style as categorical: either didactic-only or using student-participation. He might consider treating the encouragement of student participation as a continuous measure and manipulating the amount of time he devotes to student-participation exercises, or the strength of conviction he uses to encourage participation and testing the impact of many quantities. Decisions to use continuous, ordinal, or categorical measures determine the types of statistics which may be used for analysis and therefore the types of conclusions which may be drawn from the results of those analyses.

Focus on effects

Choose an accurate and meaningful dependent measure: All inferential statistics require at least one dependent measure. That measure is a number that represents the construct affected by the manipulation. Dr. Marcus is hypothesizing that his lecture style impacts student learning, so his dependent measure must be some number that indicates how much a student has learned. He may use the percentage of final examination questions testing a particular topic that the student answers correctly. He may ask his teaching assistants to rate how much learning each student has demonstrated in each topic area. He may ask his students to rate how much they have learned in each topic area. Each of these measures is an attempt to measure student learning, but each carries a different shade of meaning.

Dr. Marcus has two concerns: (1) the dependent measure must be valid—it must represent how much the student has learned about a given topic– and (2) the dependent

measure must be reliable—if taken again, or in a different context, it would yield the same answer. In choosing a dependent measure, he assumes the construct of "learning" is some measurable thing (allowing a valid measure) that is fairly stable over time (allowing a reliable measure), both of which are major assumptions. There are several ways to estimate the validity and reliability of any given measure, which are beyond the scope of this document. Ultimately, Dr. Marcus' conclusions about the impact of his lecture style on student learning will rest on how well his dependent measure represents student learning. So, he should spend time convincing himself and the readers of his research results that his dependent measure is both valid and reliable.

No measure is entirely valid nor reliable. Even a construct like a person's height, which we can be intuitively sure is some actual, observable quantity, can never be measured perfectly. The person in question may slouch, or stand at an angle, or tilt his head. The measurer may start or end at slightly different places of his feet or head. The measuring tape or stick may expand or contract due to the temperature or humidity. Height itself is somewhat unstable, as individuals are slightly shorter in the evening after standing all day. As such, even when the construct in question is a tangible, physical property, several measures of that construct will yield different answers. The tendency of those measures to be approximately the same gives us an idea of the correct answer. The tendency of those measures to differ from each other gives us an idea of how much noise is in the measure.

If Dr. Marcus conducts his experiment using only two or three students, in only two or three lectures, he can have no idea whether the differences between them are due to the effectiveness of his experimental manipulation or due to the myriad possible irrelevant things that influence his dependent measure. Even if he has managed to remove every possible confound which might allow an alternative explanation for his results, he'll never be free of these random differences. Statisticians call differences "error" and all inferential statistics are ratios of observed differences to estimates of how much error is in the measures: this ratio is the effect size. If Dr. Marcus finds a dependent measure that is highly reliable, his measures will have little error and he'll be

able to find even small differences between his conditions. If he must use a low-reliability measure, he will be less sure that the differences he finds are not due to error.

Have adequate power: One of the most common questions a novice researcher asks is "how many subjects do I need?" A power analysis will answer this question. Power is a statistical term and is discussed below. Adequate power requires a certain number of subjects. After conducting a power analysis, you may find that you need more subjects than you could possibly run in a controlled experiment and should instead use a quasi-experimental or correlational design for which lots of data are more easily collected. Or, you may find that only a single classroom of students are necessary for adequate power, in which case a controlled experiment is entirely practical.

Power is a probability, ranging between zero and one. It is expressed as the opposite probability of the Greek letter beta, written $(1-\beta)$. It's the probability of an experiment finding a particular effect size using a particular number of subjects. One measure of effect size is Cohen's *d (the difference of two means, divided by the standard deviation of the sample)*. If the effect doesn't exist ($d = 0$), no number of subjects will manage to find it: $(1-\beta) = 0$. If the effect is huge ($d = 4$ or $d = -4$), it will be almost certainly be evident after observing a handful of subjects: $(1-\beta) > .99$. Therefore, in order to know how many subjects you'll need, you have to know how large the effect you're expecting will be. Since the experiment you're planning is (hopefully) asking a question no one has answered before, you'll have to guess how big that effect will be, and so your estimate of the number of subjects you'll need will be based on this guess. However, a review of the literature for similar past research can help justify that guess.

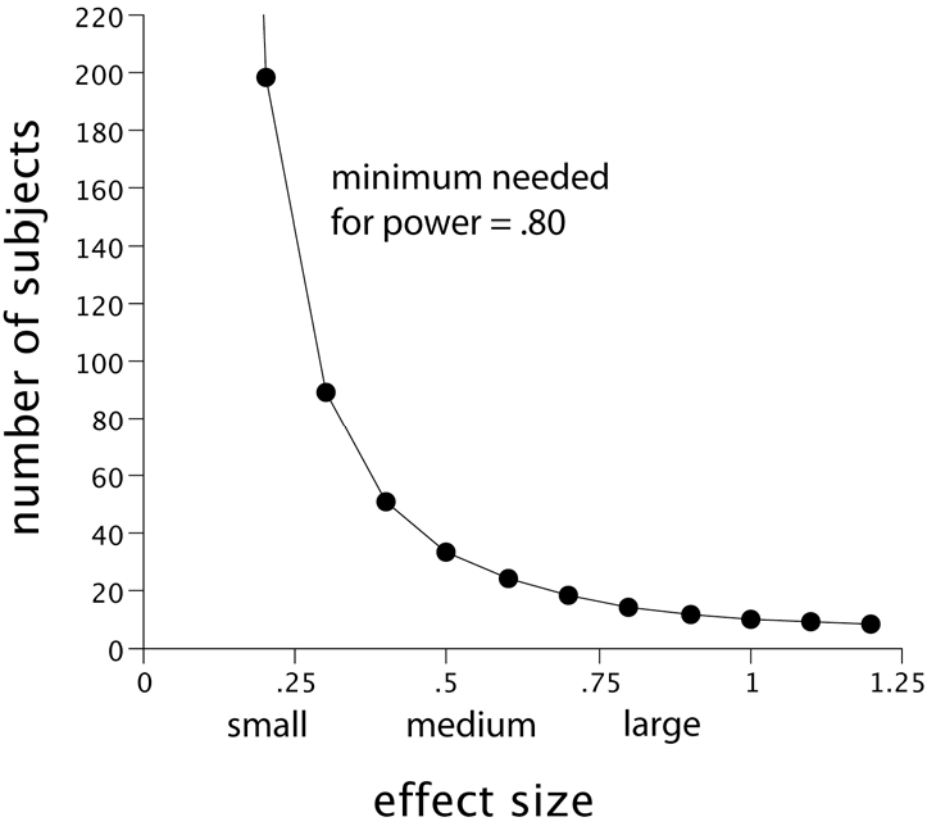

Figure 3: The minimum number of subjects needed for .80 power to detect an effect gets lower as the size of the effect gets larger. Detecting small effects is generally unfeasible for small-scale studies.

Typically, an experiment with $(1-\beta) = .80$ has adequate power. Since this means the experiment only has an 80% chance of finding the specified effect size statistically significant, a researcher may want more power than that. Figure **3.** shows how many subjects are needed to detect a single effect with $(1-\beta) = .80$ power. Cohen (19XX) described effect sizes typically seen in social sciences and defined the commonly used categories of small, medium, and large effect sizes. A small effect (*d*=.20) is very hard to detect: hundreds of participants are necessary to achieve adequate power. Medium (*d*=.50) and large (*d*=.80) effects are much more detectable, requiring no more than one moderately-sized class of students. Figure 3 is useful, but is a rough guide: the actual power will vary depending on the number of other effects being tested. Most statistical software packages can help determine the actual power of specific experimental designs.

Power analysis can also determine the smallest effect size that could be detected with $(1-\beta) = .80$ given a set sample size. This is a very common analysis when your data have already been collected or you already know the biggest sample size you can get (e.g., the size of your class). By telling you how large of an effect you would be able to characterize as statistically significant (i.e., not attributable to random variation), you are given information that may tell you whether your study or situation is adequate to give you confidence in answering your research question or hypothesis. Assume Dr. Marcus conducts his experiment and finds that students score better on topics taught using the open-ended question style, but that the effect is not statistically significant with a $p = .06$ (greater than $\alpha = .05$). He may conduct a power analysis after the fact and discover that if the effect size he's observed is true, 10 additional students would raise his experiment's power to .80. Next semester, he can raise his class limit by 10 students and re-run the experiment on the whole class to try to replicate the effect and achieve statistical significance.

Use within-subjects designs where appropriate: The reason we need many subjects to demonstrate any effect is because of the large amount of variety among people. A general statement such as "students learn better from lectures involving class-participation" cannot be true for everyone: certainly there are some learners for whom class participation has no effect, or even hurts. But the statement doesn't need to be true for all students in order to be broadly true. The statement "men are taller than women" is not true for every possible case. A random sample of subjects drawn from a women's basketball team and a men's racehorse jockey club will no doubt show the opposite effect. But the statement is true in general and is crucial knowledge for clothing designers and bicycle manufacturers, for example.

Similarly, individual differences constitute noise when trying to make general statements in social sciences. This stance may seem callous: how can an educator dismiss differences between students as "noise" or "error" when his job is to understand and adapt to such differences? This is a major point of criticism for researchers and theorists who prefer qualitative methods to the quantitative methods described in this document. These criticisms are well-taken and it may prove impossible to generate a

general, predictive educational theory—certainly such success has so far been elusive—but it may also prove possible. We will not know until we have either succeeded or given up. As Lady Macbeth mourned, "the attempt and not the deed confounds us."

A useful way of controlling for individual variation when estimating the differences between conditions is to use the same individuals. This design is called a "within-subjects" or "repeated-measures" design. Dr. Marcus' experiment is a within-subjects design since the same students will be tested in both lecture style conditions. The most powerful statistic he can compute in this design is a paired t-test. To compute this he will calculate each student's amount of learning (by whatever dependent measure he has decided on) for class-participation topics and for purely didactic topics. Each student will show some difference, since the odds of everyone getting the same score in both conditions by chance alone are very small. The paired t-test compares the average of these differences to zero and the result indicates how sure Dr. Marcus should be that students are learning more in one condition than the other.

Dr. Marcus might wonder whether the impact of student participation lectures would only work if the entire course was taught that way. In that case, he would have to use a between-subjects design. It might be possible to make the same students take the course twice—once with didactic-only lectures and once with student-participation lectures, then compare overall learning using the paired t-test described above. But (1) students' time is better spent taking courses they haven't yet had and (2) some students would have to take the didactic-only course first and others take the student participation course first to control for the possible confound of repetition. While the within-subjects design may be preferable, it may not be practical. In a between-subjects design, one set of students take the course when taught in didactic-only style and another set of students take the course when taught in student-participation style. So, the differences between the particular students in each class add more statistical error, which reduces power.

Use continuous dependent measures when possible: In general, continuous dependent measures make better dependent measures for they allow parametric

statistics. Ordinal and categorical dependent measures can only be analyzed by non-parametric statistics. Parametric statistics are very powerful, for they include differences between observations within an experimental condition into the estimate of the strength of the differences between experimental conditions. Non-parametric statistics cannot do that. For instance, if a researcher is curious about whether the use of an in-class audience response system improves student performance, he may decide to compare student class scores from a previous year (when the system was not used) to a current year (when it was). Assume, unbeknownst to the researcher, the audience response system only helps good students to perform better and has no effect on poorer students. If his class scores are continuous, say a percentage of test questions answered correctly, he may use a parametric statistic—an independent samples *t*-test—and find that overall scores did indeed rise, since students who would have achieved around 80% or so were able to achieve around 85%. If his class scores had been pass-fail, he would have had to use a non-parametric statistic—a chi-square test—and would have found no result, since the number of failing grades relative to passing grades would not have been any different.

Continuous measures can be sliced into any number of ordered categories for other types of analysis. If you collect study participants' adjusted gross incomes, a continuous measure, you can later determine their tax brackets, an ordinal measure. But if you collect only tax bracket, you cannot later turn this information into the finer, continuous measure of income.

Handling Data and Analysis

Managing data

Quantitative data is typically represented using spreadsheets, with one variable or measure per column and one subject per row. Most statistical analysis software programs use this format. It is a good idea to familiarize yourself with the statistical software and the specific analyses you will be using before you begin entering data into a computer. Some analyses are much easier to perform when the data are organized a

certain way. For instance, if Dr. Marcus' lecture style experiment described above uses a within-subjects design and so requires a paired *t*-test, he should have one student per line, each with two columns: (1) a measure of amount of learning on didactic lecture topics and (2) a measure of the amount of learning on student participation lecture topics. Any statistical package, when asked to perform a paired *t*-test, will ask which two variables are to be compared.

Many novice researchers are familiar with spreadsheet programs, such as Microsoft Excel©. Spreadsheet programs are useful for entering and organizing data into different row-column configurations. Excel© in particular is a powerful program with many subtle, scriptable features, many of which allow statistical analysis. However, it is always worth the monetary and time investment to buy and learn a dedicated statistical software package such as SPSS©, SYSTAT©, or JMP©. Talk to colleagues at your institution to find out what they use and what site licenses your institution offers to faculty and staff. Adopting the same software as colleagues with whom you may be collaborating and sharing data will save a lot of time and headache in converting file formats and coding conventions. If you lead a lab or department, talk to your students and staff to find what programs they tend to prefer. The people who do the work should choose the software.

Particularly computer-savvy users may opt to use a more generic mathematical software platform, such as SAS©, S-PLUS©, MATLAB©, or R (R having the advantage of being free and open-source). These programs offer a great deal of flexibility and allow the researcher to perform various exotic, customized, or subtle analyses unavailable on dedicated statistical software packages. These packages are not for the mathematically faint of heart, however, and require a lot more time and learning.

Below are some basic tips on managing data sets to make analysis easier:

Use subject codes and anonymous subject notes. In education research, one of the IRB's primary concern is with the anonymity of student data. Typically, an experiment or even an exploratory study does not require the researcher to know which students contributed which data points. It is usually enough to know that a student's basic demographic information (gender, age, year in program) and the values of that student on the independent and dependent measures relevant to the study. Once this

data is collected, throw away the student's identifying information (name, student ID number, social security number) and assign the student some meaningless code name or number. This practice will ensure that data will never be used against the students, which will make it easier to get permission from the IRB and remove the primary reason for data security.

That said, keep some notes on some aspects of data collection that might be needed later. For instance, if a student in one class is unusually prone to asking questions during class, Dr. Marcus might make a note that "SUB0020" (that student's meaningless code) is very talkative. So when he or another researcher finds that student's learning is unusually unaffected by the lecture style manipulation, an alternative explanation (that the student always participates) is available. Such notes are best kept in the statistical software, in a string variable called "notes" or "comments" so that anyone analyzing the data can see them easily.

Code missing data as missing. When a student doesn't answer a question or misses an assignment, use a coded value to represent that the data is missing. Software packages tend to use a period (".") to represent missing data. Some older conventions use impossible numbers ("999" or "-1" on scales that range from 1 to 10) to code for missing data, which requires explicitly telling the software to treat those values as missing. A novice researcher may be tempted to use zero, which can lead to misleading statistics.

Keep track of which subjects you exclude and why. It is incredibly common for subjects in social science to misconstrue instructions, behave erratically, or quit during the procedure. Students may transfer, become ill, come consistently late, miss exams, or other problems. If the researcher has good reason to believe that these problems have rendered the data collected from that subject unrepresentative of the phenomenon being studied, he may elect to exclude that subject from the analysis. This decision is best made before any analysis is conducted, since the temptation will always exist to rid the data set of "troublesome" data points which seem to be preventing effects from being significant. The easiest way to avoid that temptation is to vet each subject for appropriateness before analysis.

If you do exclude subjects, keep their data in the data file and keep notes on why they were excluded. Most statistical software allows the user to exclude subjects from analysis without deleting them. The researcher is responsible for reporting the number of excluded subjects and the reasons for the exclusion. Readers of the research will want to know how generalizable your results are.

Think about data security. Using anonymous subject codes makes your data files less sensitive, since no information from the file could be used to harm the study participants. If, for some reason, student identity is important to retain, the data set should be guarded carefully. Even if subjects are anonymous, you may not want the data to be too widely available for a variety of personal or professional reasons. Using password file encryption, keeping data files on secured computers, not available through networks, and storing disks in locked cabinets should become part of routine maintenance of a research lab.

Data corruption is more likely than theft. Especially in the case where several analysts are accessing the data, a researcher should always keep a single, primary copy that can be used to resolve conflicts if unauthorized changes are saved, or if other data corruption occurs. It is always a good idea to burn the full data set to a writable CD or DVD for safekeeping. If the data set is not too large, printing a hard copy of the data to paper should be considered. Analysis should be performed on copies of the original to avoid accidental corruption. If data becomes untrustworthy, results become suspect and the entire purpose of the study is nullified. Data is the *sine qua non* of science: treat it with care.

Choosing the appropriate statistics

The appropriate statistics for planned comparisons in planned studies are determined almost entirely by the type and number of dependent and independent variables. This section is a brief outline of some of the basic statistics appropriate for certain types of measures and designs. Depending on the specific nature of the data, the recommendations here may not be the best choices, but for the vast majority of research, these recommendations will be adequate.

The mathematical specifics of these tests is beyond the scope of this paper, though the researcher who uses them bears the responsibility of familiarizing him or herself with that math to facilitate appropriate interpretations of the results.

No matter the statistical tests used, it is always a good idea to graph the data:

Cardinal rule #8: Graph your data.

Inferential statistics can be entirely blind to relationships and trends that are perfectly obvious to the eye. One of the most common difficulties of using inferential statistics is their vulnerability to outliers. Outliers are data points which have values on one or more measures which are very different to most of the other data points in a sample. For instance, if two groups of students perform identically except for one student in the control group who missed the final and scored a zero, it is possible that a *t*-test would find a difference between the groups. Looking at the statistics alone, a researcher might conclude that the intervention successfully raised scores. But once the data are plotted, that researcher is likely to notice the outlying zero and take the peculiarities of that case into account, perhaps by excluding that subject and re-running the analysis. As noted above, this zero should have been treated as missing data and that subject perhaps excluded from the start.

Histograms are the best graphs for examining measures. Histograms are bar charts showing the distribution of a continuous variable: the number of subjects within each of a set of ranges of values of that variable. Educators often use histograms of student scores to “curve” a test. Dependent measures should yield a fairly normal distribution on a histogram. Parametric statistics assume that the continuous measures in question have normal distributions, so distributions that look extremely skewed or kurtotic might require some explanation or special treatment. It is a good idea to generate a histogram and some simple descriptive statistics (mean, median, standard deviation, skewness) for each continuous measure in the data set to make sure the values make intuitive sense and there are few, if any, outliers.

The next most important graphs are bar graphs and scatterplots, which can show the relationship between a continuous dependent measure and one independent, either categorical or continuous, respectively. On both types of graphs, the dependent measure is plotted on the vertical axis and the independent on the horizontal. In a bar graph, each condition of the independent measure is represented by a vertical bar, the height of which represents the mean of the values in that condition. In a scatterplot, each subject's value on the continuous independent measure determines the horizontal position of a data point and his or her value on the dependent measure determines its height. These graphs quickly and easily show the relationship between two measures. Most software programs generate them easily. It is worthwhile to generate at least one such graph for every statistical test you perform. Sometimes the specific bar heights, or the placement of individual data points on the scatterplot can be more informative than the statistic that attempts to summarize those patterns.

Once you have confirmed that your data set is correct, secure, and free of errant data caused by uncooperative subjects or uncontrollable circumstance, generate your summary statistics and basic plots. Once you have confirmed that your dependent measures are fairly normally-distributed and appear related in sensible ways to the independent measures in the data set, you may conduct your inferential statistics to test your hypotheses. Choosing the correct inferential statistic is tricky: there are myriad statistics, each with a slightly different meaning and purpose. Consultation with a statistician is always a good idea. However, we have provided a quick overview of the most basic, commonly-used statistics. As mentioned above, complex analyses help correct weak study designs. If practical or ethical circumstances constrain study design a great deal, those complex analyses may be necessary. But ultimately, a causal model is not well understood if its predictions cannot be tested using these basic statistical tests.

What are statistics for?

Statistics allow us to make general statements about data containing variance. As mentioned above, the statement "men are taller than women" is true in general, though not always specifically true. Given a large enough sample of men and women, drawn

fairly and randomly from a representative population of people, the truth of the statement will become clear. The statistical test of the statement on the sample will take into account the mean height of the men, the mean height of the women, the individual variation in height shown in the sample, and the number of people in the sample and generate two relevant pieces of information: the effect size (how much taller is one gender than the other relative to the typical individual variation in height) and the statistical significance of this observation (how confident should we be that this effect size is not an erroneous measurement of what is really an effect size of zero). The former is represented by the test statistic itself (*t* in t-tests, *F* in ANOVAs, *r* in regression, chi-square in chi-square tests, etc.), the latter is represented by a *p*-value.

Cardinal rule #9: Do not obsess about *p*-values. Concentrate on effect size.

A word about *p*-values: Every inferential statistic will generate a *p*-value. These *p*-values receive a great deal of attention since they determine the statistical significance of a result. However, the logic of statistical significance is slippery and commonly misunderstood. The lack of statistical significance (a "null result") may be a sign that there is no real-world causal effect between independent and dependent measures.. But studies with low power may fail to find a statistically significant effect even though there actually is a real-world causal relationship (this, by the way, is called a Type II error). On the other hand, statistical significance alone does not merit mention: any effect, no matter how small, can be measured with statistical significance.

For instance: the statement "men named Dave are taller than men named Mark" is not true in any real sense. However, if we managed to find every single man named Dave and every single man named Mark, measure their heights and compute the mean of each, those means would certainly not be exactly the same. They may be very close, but if measured finely enough, the mean height of one group will be higher than that of the other. At some point during data collection, that small, meaningless difference will become statistically significant. At that point the researcher can be somewhat certain that the real-world difference is not exactly zero. Having achieved that amount of

information, however, has no bearing on the relevance, importance, or general truth of the hypothesis in question. It is therefore, a fruitless exercise to keep collecting data until an effect becomes significant. It is much better to compute the power of your experiment beforehand and conduct your inferential statistic only after you have collected enough data to give you sufficient power.

The *p*-value is only a useful indicator of the reliability of an effect if the reader has a sense of the quality of variance of the measures in question. For instance, a report that men named Dave are on average .003 inches taller than men named Mark ($t(4,999) = 1.96$, $p < .05$) is likely uninteresting since it took the measurement 5,000 individuals (indicated by the 4,999 degrees of freedom of the *t*-statistic) to achieve a barely significant effect of a tiny (3 thousandths of an inch) difference in height. A reader who doesn't know that people tend to be around 70 inches tall won't know that .003 inches is a tiny difference. A reader who doesn't know that people are easy to find and vary in height a great deal won't know that 5,000 individuals is an unusually large sample size. As described below in the section on writing results sections for publication, it is always a good idea to report group means and standard deviations to help the reader understand the importance of a statistical effect.

A relatively new value for one-tailed statistics, *p-rep* is gaining popularity in the psychology literature. It is a non-linear transformation of the *p*-value and explicitly represents the expected likelihood that a result in the same direction would occur if the study were conducted again. For instance, if a researcher found that women enjoyed a given learning activity more than men ($t=2.18$, $p<.05$), she might report a *p-rep* of .85, indicating that if the study were rerun, there would be a 85% chance that women would enjoy the activity more than men. The intuitive interpretation of this statistic makes it less confusing than the raw *p*-value. However, this statistic is very new and very rare.

Parametric statistics

When a planned study uses one continuous dependent measure, parametric statistics may be used. These are the most powerful statistics, which is why, as mentioned above, continuous dependent measures are superior to categorical or ordinal ones. When Dr.

Marcus tests the impact of a guest lecture series on student motivation, he may choose to measure student motivation using a questionnaire that generates a score from 0 to 100 for each student. He routinely gives this survey, so he already has data from previous years. The survey score is a continuous dependent measure, so parametric statistics are possible. The specific analyses available are determined by the type and number of independent variables:

Independent samples *t*-test. Dr. Marcus has one independent measure (presence of that guest lecture series) that has exactly two condition (yes and no). He may compare student motivation scores from his most recent class, which incorporated the guest lecture series, to a previous class which did not. Since the two conditions contain different students, he is using a between-subjects design, so his test is an independent samples (not a paired) *t*-test.

Independent samples oneway ANOVA. Dr. Marcus has two years, each with a different guest lecturer. Because he has more than two conditions of the independent measure, he can run an Analysis of Variance (ANOVA). Since there is only one independent measure, the analysis is called an independent samples oneway ANOVA (or simply, oneway ANOVA). The oneway ANOVA is similar to running an independent samples t-test between every pair of conditions, but corrects for the possibility of finding results by accident due to multiple tests using the same data set. He may find that motivation was higher when there was a guest lecturer than when there wasn't. He may find that only one of the guest lecturers yielded higher motivations. A bar graph of motivation scores by condition will help Dr. Marcus summarize the results of this test.

Paired *t*-test. If Dr. Marcus measures student motivation about topics covered by the guest lecture series and topics not covered, he can conduct a paired *t*-test to compare each student's motivation between these conditions. As described above, this test is more powerful than the independent samples t-test since it uses a within-subjects design.

Repeated measures oneway ANOVA. If Dr. Marcus measures student motivation about six topics, each covered by the guest lecturer, to a different extent, he may perform a repeated measures oneway ANOVA. This test is more powerful than the

independent samples oneway ANOVA mentioned above because it uses a within-subjects design. It is also analogous to performing paired *t*-tests between each pair of conditions.

One important note when using repeated measures ANOVA with more than three conditions: make sure the assumption of sphericity is not violated before interpreting the results. The math behind this assumption is beyond the scope of this manuscript, so if the situation arises, we advise a statistical consultation. Most statistical packages will compute Mauchley's W, which tests for the violation of sphericity, automatically. If W is significantly high, consult a statistician about the appropriate correction to apply to the test.

ANOVA. Dr. Marcus hypothesizes an interaction: that males are more motivated by guest lectures than females. To test it, he has measured two classes (one with a guest lecturer and one without) and recorded student gender and motivation level. He must run an ANOVA. ANOVA is the appropriate analysis if the design is balanced—that is, each unique combination of values of the independent measures contains the same number of observations. Since neither of his independent measures is within-subjects, his ANOVA will contain only independent samples tests. The use of two independent measures makes it a two-way ANOVA and it will test for three effects: the main effect of gender, the main effect of guest lecturer, and the interaction of the two.

Often, a design will have both within- and between-subjects independent measures—this requires a mixed-model ANOVA. For example, a comparison of the effectiveness of a learning intervention on men to women may have a repeated measure (pre- and post-intervention) and an independent sample (gender). In order to test for the interaction of these two measures, the mathematical assumptions behind the analysis of within- and between-subjects models have to be reconciled.

More complicated ANOVAs are also possible, but it is important to remember that each additional independent measure can interact with each of the previous main effects and interactions. A three-way ANOVA tests for 7 effects (3 main effects, 3 two-way interactions, 1 three-way interaction), a four-way ANOVA tests for 15 effects (4 main effects, 6 two-way interactions, 4 three-way interactions, 1 four-way interaction), and so

on. Extraneous tests reduce the power of the study and—as noted above—complicated interactions seldom help explain phenomena.

Linear regression. Dr. Marcus thinks that a guest lecturer only improves motivation to the extent that he uses concrete examples. So for each guest lecturer, he records the number of concrete examples used and he measures student motivation for each guest lecturer's topic. He finds that lecturers use between zero and eight examples, with most using about three. He can conduct a linear regression analysis to see if more examples yield higher motivation. Linear regression does more than compare the mean motivation of zero-example topics to the mean motivation of one-, two-, and three-example topics, etc. It tests for the existence of a trend where each example adds some constant amount of motivation. A scatterplot best represents the results of a linear regression and may identify reasons why the regression finds or does not find a result. Regression is very powerful and rests firmly on the assumption that the continuous measures being used have normal distributions. Histograms and scatterplots will help the researcher understand what the results of a regression analysis mean.

Multiple regression. Dr. Marcus also thinks that the lecturer's age might influence motivation, perhaps interacting with concrete examples—perhaps older lecturer's examples impact motivation more than those of younger lecturers. If he has recorded the lecturers' ages as well as the number of concrete examples they used, he can conduct a multiple regression analysis, testing for the linear effects of age and number of examples and also their interaction. When a researcher has a specific hypothesis such as Dr. Marcus does in this case, he may conduct the multiple regression analysis to specifically test for the hypothesized interaction. However, conducting a multiple regression analysis with several independent variables with no a priori hypotheses will surely lead to bad results. Multiple regression is too powerful for this purpose and will almost always find significant results, even in perfectly random data. It is a technique best used conservatively.

Non-parametric statistics

Non-parametric statistics are needed when the dependent measure is not continuous. Sometimes it is a good idea to treat a continuous dependent measure as if it were

categorical and run non-parametric statistics. For instance, each student receives a percentage score on a test (a continuous measure), but these are ultimately converted into letter grades (an ordinal measure); an administrator interested in the impact of a new textbook on students' scores may not be interested in the change in percentage scores if it doesn't change the number of A's, B's, and C's. However, non-parametric statistics have less power and so are less likely to detect smaller effects. There are many non-parametric statistics available. Two of the most common, chi-square and Spearman's correlation, are described below.

Chi-square. Dr. Marcus is interested in students' attendance of guest lectures vs. his own lectures by male and female students. His dependent measure is inherently categorical: each student either attends a lecture or does not. He could compute the percentage of lectures of each type each student attends and treat that percentage as a continuous measure. He could also test for an effect non-parametrically using a 2 by 2 matrix of 4 numbers: the numbers of males and females who attended all the guest and his own lectures. The statistic to test for the difference is chi-square.

Chi-square requires more than four or five students in each cell of that matrix. If, for instance, only 2 males attended all the guest lectures, the chi-square statistic will be too unreliable. Chi-square is computed by comparing the observed numbers to a hypothetical set of numbers we'd expect if there were no effects. This is important to remember because Dr. Marcus isn't interested in whether there are more males than females in his class. Unless he tells his software to ignore that main effect, his chi-square statistic will test for that.

Spearman correlation. Dr. Schneider may be interested in the impact of a new admissions ranking process for applicants. She wants to compare the new system of ranking applicants to the one used in previous years by applying both methods to the current crop of applicants. She knows that rank is more like an ordinal measure than a continuous measure. Moreover it's not a parametric measure: it doesn't have a normal distribution. She wants to run a correlation to estimate the similarity of the two rankings, but knows that a Pearson's correlation is parametric, and therefore inappropriate. She needs to run a Spearman's correlation, which compares two sets of ranks.

Sharing your results with the world

Since theories contain many causal models and each causal model can generate many hypotheses and each hypothesis must be tested by multiple studies, science is inherently a collaborative project. No single theorist can make a significant contribution to her field without an army of similarly-minded researchers attempting to falsify and support various aspects of her theory. Every seminal theory or experiment in every scientific literature has become seminal only through repetitive testing and replication by a community of scientists. Sharing the results of a study is one of the most important aspects of research. The research publication—an internal report or a peer-reviewed journal article—is the ultimate and permanent form of a research study.

The goal of any research publication is twofold: 1) to allow any scientist to replicate the study exactly as it was first conducted and 2) to give that scientist an idea of what results to expect. Without these aspects, a research publication has little value to other scientists. Research publications in most fields, including education, tend to adhere to this basic form: introduction, methods, results, conclusions. The introduction briefly describes the theories, causal models, and hypotheses in question and summarizes other studies findings on the same topic. The conclusions briefly describe the researcher's interpretation of the results and her sense of the implications for the theories in question. Both of these sections chiefly contain subjective interpretations of widely-available resources and contain arguments which anyone can challenge or defend. Only the methods and results sections contain information known only to the researcher. It is the researcher's responsibility to accurately and completely describe her methods and results and only the researcher can vouch for that accuracy.

It is not uncommon to hear the lay press criticize the editorial and peer-review process of a journal when a result published in that journal later proves to be founded on falsified data or statistics. Such criticisms are entirely misplaced. The editorial process can only constrain the validity of the claims researchers can draw from results. Only the researcher can speak to the validity of the results. There is no glory to be gained from

publishing results based on falsified data or statistics. Replication will not support fictional results and empiricism will defeat lies. It is probably the most important cardinal rule:

Cardinal rule #10: Report results honestly and share datasets freely.

Researchers tend to permanently archive the datasets behind the results they published. Colleagues will ask to see the raw datasets in order to confirm surprising analysis results, try different analyses to test other hypotheses, or combine multiple data sets into one large one for exploratory secondary analyses or meta-analyses. Such analyses are important and such requests are common. Increasingly, it is becoming common for researchers to put data sets on Internet servers for public access and to give the Internet address in the manuscript. Be prepared to share your data. If data is collected by students or lab staff who answer to you, maintain the contact information for those individuals in case readers or editors question the veracity of the data. Do not put your name on publications reporting results based on data you do not trust entirely, or would not feel comfortable sharing with colleagues.

Before writing a manuscript, choose the journal you want to submit it to. Every journal has a different protocol for manuscript submission. Some have Web-based submission software, some allow submissions by email, some require paper documents be mailed to the editor. Each journal has a different page limit, word limit, manuscript format, section and sub-section outline, title-page requirements. Some require the authors to put their names on the title page and some forbid it. Read and adhere to the journal's submission guidelines when preparing your manuscript. A lack of adherence to the journal's published requirements is a sure-fire way to have your manuscript rejected.

Choose the journal carefully. It is scientifically unethical to have a manuscript under review at more than one journal at any given time. It is scientifically bad form to have multiple manuscripts based on the same analysis, or even the same data submitted to multiple journals simultaneously. Journal review processes take several months and revising the manuscript for a different journal can take just as long. So choosing a

journal that is unlikely to publish your work could unnecessarily delay its publication for a year or more. It is often a good idea to ask a journal editor to quickly scan a manuscript draft, or even just a brief description of the research methods and findings, to get feedback on its appropriateness for that journal. Such simple communication can save years of wasted time and resources and can help strengthen your relationships with the journal editors in your field, making subsequent publication processes smoother and easier.

Reporting Methods

A methods section is a recipe for replicating the study. It must describe the materials in enough detail that another researcher could find or create the same. It must describe the independent measures: what concepts were they measuring, how were manipulations achieved, what manipulation checks were used. It must describe the procedure for data collection: what were the dependent measures, how were they collected and quantified. It must describe the statistical analyses and the software used to conduct them.

In any study, it should be possible to write the entire methods section before the study is completed. This document serves as a research protocol and is good academic practice. Some journal editors are glad to look at research protocols to assess the suitability of the research for publication in that journal; acquiring those assessments and consequent feedback can be invaluable for avoiding wasting resources on unpublishable research. The research protocol also allows a team of research assistants who may be collecting data for the same project to standardize their methods and avoid subtle biases due to methodological differences.

A publication must describe the human subjects who participated—how they were recruited, how they were compensated, how many excused themselves or were excluded and why. In all social sciences, this information is crucial because it indicates the generalizability of the results. Generalizability is the breadth of the implications of a result. A result obtained from volunteer graduate-level students may or may not be replicable in a random sample of undergraduate students. A result found in an engineering education program may or may not be replicable in a business school. If a

result appears in many different types of samples, it is highly generalizable, and has deep implications for a theory. When a finding has low generalizability, the composition of the sample which exhibits the result can inform further research, such as the complicated interaction between number of assigned readings, gender, specialty and year in program described above: the more specifics given in the presentation of the sample, the more likely an astute reader will generate a more parsimonious alternative hypothesis.

Reporting Results

Report the effects you planned to test for and little else. Publishing unexpected results merely because they are statistically significant is counter-productive to science. This occurs more frequently in the literature than one might expect, primarily because statistical significance is typically necessary for acceptance for publication in a peer-reviewed journal. Countless experiments produce no significant results and are never published. This is called the “file drawer effect” and causes the literature to overstate the validity of many causal models. Since any result can be found to be statistically significant given a large enough sample (as explained above, in the section on statistical power), researchers are tempted to keep collecting data until their *p*-values are low enough to publish. Even worse, statistical significance will happen by chance alone if one runs enough inferential statistics on a dataset, so researchers are tempted to run many tests and not report the non-significant results. To the extent that these decisions are made after examining the data, they are *post hoc*, rendering those results exploratory. If you find a result you think is interesting and potentially valid, report it, but explicitly state it is a *post hoc* result and therefore exploratory.

Ideally, each reported result should include 1) the inferential statistic 2) the effect size (if the independent measure is categorical), and 3) means and standard deviations of groups (if the independent measure is ordinal or categorical) or slopes of trends (if the independent measure is continuous or ordinal). The inferential statistic informs the reader of the likelihood that the result differs from null due to chance alone, implying the likelihood that it is replicable. The effect size informs the reader of the strength of the

causal link between the independent and dependent measures relative to other causes and measurement error and therefore of its practical and theoretical importance. The third gives the reader meaningful values of the dependent measure, allowing a complete understanding of the effect.

State the each result in a single sentence and immediately follow each parenthetically by its supporting statistic, its degrees of freedom if appropriate, whether its *p*-value is below or above the appropriate level of significance, and its effect size (e.g., *d*). Effect size can be calculated from knowing the value and degrees of freedom of several statistics, so some journals might not want to publish the redundancy. However, it is good form to provide effect size wherever permitted to help future readers plan their own related studies. Parenthetically report means and standard deviations of each group in the sentence wherever possible.

Tables and figures are very useful for presenting the means and standard deviations of many groups and are crucial for presenting complex interactions. Journals tend to limit the number of figures and tables a manuscript may contain, so they are best used carefully. However, remember that a figure is an accessible visual representation of a great deal of information that may be difficult to explain in the text. A good figure will be useful for presenting your results to colleagues and for your colleagues to summarize your results in their own presentations. It is a good idea to make a figure representing the most important finding in your study, even if the result is easily explained in the text. The figure will highlight the result and more readers will remember it.

Publication

Ultimately, a research publication is only one puzzle piece that a careful reader must reconcile with the rest of the literature. The measures and methods used, the results expected, and the results found (planned and exploratory) define the contribution of the study to the larger theoretical picture. Only from these can we infer the actual causal structure of the constructs in the world and the most parsimonious theories for describing that structure. A manuscript with carefully written methods and results

sections will provide great value to the scientific community. Clarity and honesty of presentation are paramount.

Most journals will send a submitted manuscript to two or three of your colleagues, who will read it and rate the quality of the research presented and the theoretical conclusions made by the authors. Most journals will withhold the manuscript's authors' names from the reviewers to avoid biasing their ratings. The editor will usually share these comments with the authors. Commonly, the editor will request the authors amend their conclusions, add some introductory statements, run additional analyses, or even collect additional data and then resubmit the manuscript. As mentioned above, scientists are skeptical and the peer review process exists to hone the validity of theoretical statements in the literature. Reviewer comments, no matter their harshness or seeming irrelevance, are valuable information about how the community will react to the manuscript. If the editor gives a "revise and resubmit" judgment, it is best to quickly address all the reviewers' comments and resubmit.

The published article is the final form of your study. Make sure it is concise, accurate, and clear. The world is a messy place and transparent, methodical empiricism will help make sense of it. Published articles that make vague claims, summarize small or careless data sets, or present ambiguous results will merely be glossed over by an already overwhelmed community. Published articles that address important questions directly and simply and provide interesting and well-supported results will be remembered and cited by other researchers.

But no work ends with a study's publication. Every article raises enough alternative explanations and questions to motivate a dozen more studies. Read articles by your colleagues, challenge their assertions with a critical skepticism, and practice developing alternative explanations for published results and inventing hypotheses that can resolve conflicts between incompatible theories. Write methods sections that test these hypotheses and develop ways to collect, store, and analyze the data that can test those hypotheses. Every result you find will raise many alternative explanations, each of which is a challenge to your causal model. Each alternative explanation suggests another hypothesis, another research study that can further test the validity your causal

model. Each study is another possible publication and another possible line for your research program.

Appendix: The 10 Cardinal Rules

1) Exploratory studies can suggest, but cannot test theories.
2) Regard your results with skepticism. Type I error results from excess credulity.
3) Correlation does not indicate causation
4) Theories can never be proven true, they can only fail to be proven false.
5) No single study can prove a hypothesis false.
6) Randomization ensures unbiased results.
7) The better the experimental design, the simpler the statistical analysis and the stronger the causal assertion.
8) Graph your data.
9) Do not obsess about *p*-values. Concentrate on effect size.
10) Report results honestly and share datasets freely.