CHAPTER 11

Protocols for Experiments

CHAPTER PREVIEW

What's It All About?

Experiments are the method of choice when testing questions or hypotheses that involve causation—the effect of one variable on another. Experimental design attempts to isolate the relationship between the variable that is thought to be the cause and the variable that will show the effect so that only those two variables are in play.

What Are the Major Topics?

For experiments to make sense, there has to be a foundation of a causal relationship. In classical terms, if you touch a hot enough stove with an unprotected finger, you will get a burnt finger. In communication, causal relationships are considerably more mushy with both multiple causes and absent effects.

Well-designed experiments follow a deductive model in which theory forms the major premise, the experimental protocol forms the minor premise, and the hypothesis to be tested is the conclusion.

Without an adequate theoretical foundation, the analyst cannot anticipate the needs for control in the experimental design.

There are four kinds of variables that can be in use in an experimental design—the experimental or independent variable that is thought to be the agent of the effect, the criterion or dependent variable that is the effect to be measured, covariate variables that may affect the relationship between the independent and dependent variables, and control variables that serve to isolate contaminants.

An experimental treatment is a set of conditions determined by the analyst based on his or her examination of theory and previous research that will produce the desired effect.

An experimental protocol involves the management of variables, their presentation, the assignment of respondents, and the statistical procedures of analysis.

Statistical analysis should always begin with an examination of the characteristics of the data set called the descriptive statistics.

The ethics of statistical design include the requirements for a meaningful test, the care of respondents, and the appropriateness of the test's conclusions.

What Special Terms Are Used?

Central tendency
Composite variable
Cross-respondent effects
Interrespondent effects
Intrarespondent effects
Necessity clause
Pattern recognition
Pre-post design
Research hypothesis
Response demand
Simple means
Sufficiency clause
Surrogate measure
Telegraphing

INTRODUCTION

Within the scientific method, experiments are the gold standard for demonstrating causation. A perfectly designed protocol for an experiment tests potentially alternative routes to the same outcome, holding all other conditions constant. The alternative that initiates the projected outcome is determined to be the cause of that outcome. The argument follows Mill's classic canon for *A* being the cause of *B*: Whenever *A*, then *B* (the sufficiency clause); and no *A*, no *B* (the necessity clause). We demonstrate Mill's canons every time we flip a light switch: When the electricity flows (*A*), the lamp glows (*B*); no flow, no glow.

Mill's canons are wonderful standards for designing electrical circuits. They are much less useful for studying communication variables. The reasoning here requires me to locate you into some prior assumptions about why these limitations in classical causation occur. Let me begin with the least controversial. Let's presume that *A* is some message and *B* is some behavioral response. Few communication scholars would argue that *A* can exist as a consistent force or motive for action regardless of its context and conditions of presentation and reception, and few would argue that *B* can be an automatic and fully encoded action routine independent of its context and conditions of performance. The result is that both *A* and *B* are not elemental variables but are composite variables with many different manifestations.

These multiple manifestations of either *A* or *B* share at least something in common but over the range of manifestations may show more differences than commonalities. An *A* is never just an *A*. Nor is a *B* ever just a *B*. The result is that an *A* in an experimental protocol cannot be the *A* that appears on your living room television set. And similarly the *B* of the laboratory is not the *B* of the streets. Big problem; bigger still, we don't know what the differences might be or what differences those differences might make. We don't deal with Mill's certainties; we deal with the shadows of possibilities and implications. Some will immediately argue that it's not all that bad, but, yeah, it really is.

The controversial part of this problem comes when we add agency to the equation. Agency, remember, is some ability to do otherwise. In simple terms, this ability means that no cause occurs unless the individual allows it to occur, whatever the term *allows* entails. Experimentalists and all the other shades of determinists don't like and usually don't accept the concept of agency.

Agency is an axiomatic belief; it is irreducible to evidence. Hold it or don't, but interrogate the requirements of what you hold and be consistent in its application. I believe in a restricted form of agency that comes into play only with effort. It's a supple position.

What, then, is the role of experiments in a composite variable world? Experiments can clearly demonstrate the change in the probabilities of outcomes. I can, in fact, design an experiment in which *B* is a more likely outcome than Not *B* (almost Shakespearean). This is a contingency outcome. It occurs when the cause is necessary but not sufficient for the outcome or neither necessary nor sufficient for the outcome. In the first of these two conditions, the outcome *B* does not occur without *A*, but also occurs not always with *A* as some other condition is also required. In the second, *B* can occur without *A*, but occurs at a higher rate with *A*. Of these two conditions, the first gives us a lot more information. We know we can eliminate *B* by eliminating the appearance of *A*. And we know that to produce *B*, we need something to work in concert with *A*, although we may be uncertain as to what that something is.

The second condition is what we usually get. We can set up conditions in which *A* increases the likelihood of *B*, but *A* alone will not produce this increase, because there are other conditions in which the likelihood of *B* does not increase when *A* is present. Further, *A* can be absent and *B* will still appear. This contingency relationship typifies most effects experiments.

Consider the example of the public health announcements concerning the use of condoms for safe-sex practices. That condom use is the *B* of this example. The basic message of those announcements is "every partner; every time." That is the *A* of this example. The *B* of this example is widely practiced and for many is the standard of practice. Those people do not need *A* for *B*.

Clearly, if I set up an experiment varying the appearance of *A* (the message) to see the effect on *B* (condom use), I will have to collect indirect measures on that effect, most likely some paper-and-pencil measure on respondent intentions. In that experiment, *A* will most likely increase the probability of surrogate *B*. A well-designed message in a well-designed experiment will almost always produce the effect on a surrogate *B*. Does that mean that engaging the public service announcement will eliminate this unsafe-sex practice from among the announcement viewers? Eliminate all those "What the hell?" moments? I think not. But individuals might indeed be more likely to be prepared or at least to feel some concern (or—dare I say it?—guilt) about the risk.

What do we learn from an experiment like this one? First of all, if the research hypothesis fails and the message does not produce the expected change in the surrogate measure, the information gained is confounded in that we cannot distinguish between a failed message and an inadequate experiment. That said, if I were responsible for the distribution of the message, it wouldn't be going out.

If the research hypothesis is supported and the measures show gains in the expected direction, then I have reason to believe that the message will be effective for some people under some conditions (called the some-some conclusion). That some-of-the-people-some-of-the-time conclusion is the fundamental limit of information gained from composite variables (and their surrogates) within contingent outcomes. The remaining issue is "How much trust can one invest in the effectiveness of the message?" or, in other words, "How broad is the application of the some-some conclusion?" The answer to that question lies in the topic of this chapter—the quality of the protocol of the experiment.

COMPONENTS OF EXPERIMENTAL DESIGN

There are four overarching components of experimental design: causality, theory, control, and ecological validity.

Causality

Experimental design is built on a causal model. It is a test of the possibility that one condition or set of conditions is the result of another condition or set of conditions. The first requirement of a competent experimental design, then, is the reasonable basis to hold the possibility of a causal relationship between two variables or variable sets.

Cause in the social and discursive sciences is not the same sort of thing as cause in Newtonian science. That latter sort of cause works like the presence or absence of an electrical current on a functioning incandescent lightbulb. Flip the light switch on to connect the circuit, and the filament glows. In a working system, there is no equivocation. If the light does not go on, one changes the bulb and does not suspect that for this lightbulb the character of electricity is different or that the tungsten filament has decided not to glow.

In an experiment, then, what we would like to happen is that a consistent, perhaps slightly variable but nonetheless robust, contribution to the criterion variable be made by the experimental condition to each respondent's score. We do not have that sort of certainty in our field of study. In our field of study and with a very strong relationship, it is quite likely that one flips the message switch on 100 audience circuits and 20 of the lightbulbs glow more brightly, 75 register no measurable change, and 5 return even less light. Fair example? I compared the means of a random set of numbers from 1 to 7 with that same set where the first 20 numbers of 5 or less had 2 units added to them and the last 5 numbers of 2 or more had 1 unit subtracted from them. The respective means were 3.96 and 4.31. The t value was 4.083; $p < 0.000$. That result would have brought on dancing in the hallway.

This Monte Carlo example emulates a pre-post experimental design and demonstrates the some-some conclusion of composite contingencies. Of the respondents (circuits in the example), 75% were unchanged by the message (light switch); 20% changed in the direction hypothesized; and 5% changed in the opposite direction. The outcome, however, was significant and in the direction predicted. The conclusion would have read something like "The evidence showed that the message created more positive attitudes toward. . . ." or some similar statement based on the global effect on the means. What actually happened was that *some* respondents (the 20 and the 5) under *some* circumstances (the experimental conditions) exhibited *some* change (both positive and negative).

Is this outcome typical? I believe I have some—albeit slight—basis to claim that it is typical. I do not have access to the raw data of others, and our theory development is so preliminary and data collection so limited that few researchers approach an analysis of the contrary cases in their respondent set. I can tell you that those contrary cases exist in every data set I have collected, and my theory in use has never been strong enough to explain the some-some effect. In short, I cannot tell you why some change (in any direction) and some don't.

Like others, I use a Las Vegas economy; I can make my contribution on the 20% who support my hypothesis, the 5% who run opposite do little damage, and the other 75% provide the mass I need for statistical power and acceptable numbers. Recognize that I am being crass here, but realistic as well. (I also sleep well at night.) This is the nature of cause in a world of composite contingencies.

The epistemological requirements for causality establish the ontological characteristics of all elements in an experimental protocol. On the face of it, one could run experiments using interpretive methods. An ethnographic technique called Garfinkeling (named after Harold Garfinkel, 1967) makes use of breaching experiments where social conventions are deliberately broached. The data collection is by field notes. There is almost no contemporary use of this approach in communication studies. I found but two entries in our core set of references in which breaching experiments or Garfinkeling appeared. Both were convention papers that used the term in the literature review. Clearly, breaching experiments are not much in use, and actually are not experiments as understood in this chapter. There are clear epistemological reasons why they are not that involve the researcher-respondent relationship, the standards of ecological validity, the concept of variables, the relationship among variables, the researcher as instrument, the use of field notes as a criterion of difference, and the like.

We deal with many of these issues in the chapter on ethnography. For our purpose here, we will talk about the required assumptions that apply to the nature of the variables in use. You will recall that metric research is based in part on the principle of atomism, which, as a philosophical concept, holds that the world is made up of independent elements that act upon one another in a more or less consistent fashion. There is a transcendent order that is the sum of its parts and that can be revealed through empirical analysis. In media research, this principle translates into a number of practical axioms that in turn direct experimental design (messages are independent of one another; messages are independent of the audience; audience members are independent of one another; the message is independent of its technology, etc.). These axioms allow us to design experiments that study message effects across audiences and technologies, as we will see in the examples presented in subsequent sections.

At the same time, these axioms are quite a challenge both to our experimental designs and to the conclusions we draw from them because they assume an ontological character for our variables that we (or at least a good portion of us[1]) are pretty sure is not the case. More and more media scholars hold, for example, that a message is not the same as the material text and that the message of a material text is created in the interaction among the text, the audience, the technology, and the provenance of action. We will see the tension in the opposing experimental requirements and philosophical understandings of how communication works in all of the discussions that follow, as we have seen it here in our discussion of causation.

Theory

In the best practices of research, an experiment is the test of a hypothesis that has been drawn up to test a theoretical proposition. As we saw in Chapter 6, theory provides the major and minor premises, and the hypothesis is the conclusion that is necessary if the major and minor premises are logically true. An excellent example of this process is provided in a study by Smith and Boster (2009). This study begins with the premise that individuals attend to mediated messages in the copresence and under the extended influence of others. That presence and influence creates some part of the context in which messages are interpreted. This concept has a long history in social psychology (e.g., Sherif, 1936). In that history, however, the meaning of the message was fixed and knowable (at least to the researchers); it was the individual perceptions of the message that changed.

Smith and Boster (2009) do not reference that literature, but rather take a more cultural studies approach that holds the meaning of any text to be fundamentally uncertain or unfinished (*ambiguous* would be their term) until it is actualized in some interpretation. The research problem that is the driving energy behind this experiment is whether the context of reception creates different perceptions of the same text or actualizes different texts. A "different perceptions" perspective poses the research question as one of accuracy of interpretation; a "different texts" perspective poses the question as one of identifying the message being processed. The experiment reported in Smith and Boster adds support to the latter.

Smith and Boster provide a methodical argument from theory to hypotheses that is unusual to see in the literature, but they have an advantage of working in an area where the theory has reached a fairly high level of development (even if they think it is wrong). In the much more common case, the analyst starts with a set of empirical studies that are more or less tangential to the problem at hand and creates a mash-up of theoretical propositions to justify a set of research questions and subsequent hypotheses. (No criticism of the analyst is meant here, just a recognition of the state of our theory development.)

Haumer and Donsbach (2009) were also concerned with contextual influence on judgments, but, in their case, not on texts but on the image of integrity, personal qualities, leadership, and competence projected by a political figure. They investigated the effects of nonverbal reactions to shots of the audience, the talk show host, and the nonverbal behaviors of the political figure in an experiment simulating a talk show interview of a political candidate.

Their study is a mixture of theory-driven research and opportunistic empiricism. We are not given, for example, a solid theoretical foundation for the choice of the four image

[1] Even "hardened" effects scholars like James Potter (Potter & Tomasello, 2003) have begun to incorporate interpretation in their research.

measures, but they are certainly reasonable. There is also no theoretical formulation of why a reaction shot, per se, should affect judgments on, say, leadership, but, then again, why not? The major problem, however, is the absence of any theory on the quality and influencing force of nonverbal behaviors and reaction shots and of any theory on the narrative structure of the talk show as text.

The findings from Haumer and Donsbach (2009) are mixed and, as is the typical result in experiments without strong theoretical foundations, subject to multiple interpretations. They, subsequently, bring us no closer to an understanding of how all this stuff works but do provide one more contribution to the theoretical bricolage on social influence.

Finally, theory is nearly absent in most proprietary research, which typically focuses on narrow empirical questions. Because this research is not ordinarily entered into the public domain, let me use one of my own examples. Researchers at the University of Utah and at Utah Valley University have conducted studies looking at the relative efficacy of multiple-screen and different-format computer displays. They compared one-screen, two-screen, three-screen, and wide-format screens across typical office editing tasks (spreadsheets, Word documents, and PowerPoint presentations). Respondents were faster and more accurate in multiscreen displays and gained advantages in wider-screen formats.

These studies are typical of industry-driven experiments. They offer no contribution to a body of theory but do aid in the decision making of IT officers and in justifying requests to budget managers. They even provide an aphorism to guide these decisions: “The real estate of the desktop should match the footprint of the work.” The findings, however, are completely dependent on “how things are right now.” Changes in the technology of displays, in how operating systems handle displays, in how applications display their information, and even in the cultural definition of an accounting sheet or a page of text will render their findings obsolete.[2] This circumstance, by the way, is also common in any research that chases the development curve of technology.

Managerial practices have become increasingly data-based, substituting atheoretical empiricism for managerial expertise, authority, or intuition. That may be a good thing (particularly for consultants) except that the substitution is too often characterized by a blind faith in the truth-telling qualities of empirical data. In our monitor studies, we used criterion tasks that were common in office work but that would also benefit from multiple or larger displays (and reported the relationship). The translation we hear back from managers is that multiple screens increase productivity. The fact of the matter is that they don’t but they can. The work has to be appropriate, and the worker has to use them appropriately. The same is true with experiments.

Control

The purpose of experimental control is to establish conditions such that the hypothesis provides a complete explanation for the outcome. If the hypothesis fails to be supported,

[2] That said, if you go into the business of research or scholarship, given present conditions, you should use multiple monitors with at least one in portrait orientation. If you use them to play more solitaire, your productivity will not increase, however.

it can be declared falsified. If it is supported, it can be certified as supported. (Note the absence of true-or-false designations.) Well-designed experiments do not readily admit alternative explanations for either success or failure.

Control, consequently, is a part of every aspect of protocol design. We want the specifics of the test to line up with the generalized knowledge claims we want to advance. This goal means that (a) testing conditions apply to actual conditions, (b) treatments break across intrinsic differences and eliminate extrinsic ones, (c) participating respondents correspond to the designated population, and (d) measurements have a secure connection from operational definition to variables, to constructs, to the theoretical concepts that are the basis of the initiating premises.

Proper experimental protocols are designed to control the influence of unmeasured variables and to allow the full expression of the measured ones. Controls are put into place either to eliminate the possibility of an effect by an untested variable or to permit the extraction of its effect through analysis. It is a harsh criterion because it demands both perfect anticipation and flawless execution. Time and again, the actual experimental protocol achieves something substantially less.

Experimental studies are meant to be works in progress with an aim toward the steady improvement both of the theory the experiments are testing and of the design of the experiments itself. Most of the experimental work in mediated communication has been concerned with the social implications of the findings rather than the steady improvement of the theory or experimental design. The actual result has been an amazing proliferation of theory and an equally amazing repetition of the same experimental design.

The same would hold true for much of the “Six Sigma” movement (no more than four errors in a million events) that is part of the data-based management initiative. If academic work trumpets overreaching social disaster, business research focuses on narrow empirical issues of the here and now (even as the here and now is so quickly there and gone) using prepackaged designs (the expertise is in the box) that provide little basis for the careful analysis of needed controls. The outcomes do solve the problem of what to decide today, given whatever assumptions are in play, but offer little insight into the quality of the decision or of what to decide tomorrow.

The consequence for both spheres of experimental work is that we have a lot of results but little understanding. The opposing tension, of course, is the multiple instances of the Edison myth where advances and fortunes are made by persistence, atheoretical empiricism, or the serendipity of messing around. “Demonstrate utility, and theory will follow” is the apparent dictum. Most of our actual experimental work falls somewhere between the best practices of deductive logic and the worst practices of plugging variables into design software.

Ecological Validity

We have probably all read or seen the news reports of some new food additive (usually a sweetener) causing cancer in rats when ingested at rates 800 times the normal amount. The story will usually go to underscore the problem in applying the results across species and at those dosage rates. The National Cancer Institute concludes simply that the studies "have not provided clear evidence of an association with cancer in humans."[3] The issue for reader, reporter, and scientist alike is the ecological validity of the experiments.

Ecological validity refers to the transferability of the results found under the conditions of the experiment to the ordinary conditions under which the constructs under study might present themselves. Unfortunately, the issue of transferability is fundamentally irresolvable. The conundrum resides in the relationship between normality and control. The closer one moves toward controlling all of the possible nonexperimental influences on the effect, the farther one moves from the normal conditions of that influence. And, similarly, the closer one moves toward normal conditions, the less control over those nonexperimental influences one can exercise. The simple story is that the better the experiment, the less probable its ecological validity, given the composite variables and semiotically dynamic conditions of communication studies.

Life is hard, but not unlivable. Good design works to hit the sweet spot that balances the issues of control and ecological validity. Bringing a mixed-sex group of 18- to 24-year-old college students into a classroom, showing them sexually charged video materials, and then comparing their scores on a sexual behavior measure with the scores from a control group that read a chapter on relationships probably does not achieve that balance (see Taylor, 2005). But every decision will be case-by-case as the analyst deals with the conflicting critiques that can be anticipated while trying to satisfy the motivating belief that there is something there to be revealed.

One of the major problems we have in achieving that balance is that the ecology of mediated communication varies widely across age, gender, socioeconomic class, and other demographic variables, and it has been in a state of rapid change for the past two decades. It would take a particularly dedicated researcher to be knowledgeable, for example, about the current media ecology of, say, junior high–aged respondents, the over-70s, or any group substantially distant from him- or herself. Even within those groups there are substantial differences. My friends and colleagues—all in the same age cohort—who do not have a time-shifting DVR, watch television in substantially different ways than I do.

In a similar vein, there is the issue of the differences between researcher and respondent literacies and sensibilities. I have already confessed to you that I am not a gamer. Now, I will reveal that I rarely carry a cell phone. These are deliberate choices that I have made that suit me, but they do close me out from the skills and understandings that gamers and texters possess. Quite frankly, the biggest threat to ecological validity is researcher ignorance and the passive acceptance of this ignorance by reviewers and readers.

The first step in working toward an acceptable level of ecological validity is to recognize that in working with composite variables, we cannot rule out the possibility that everything can make a difference. The thought that one can design a study that will control all conditions does not seem to be realistic. Given that assumption, the analyst will have to carefully specify the scope of the study by answering questions as to the conditions to which the study can generalize. What are the conditions of reception and common usage? What are

[3] http://www.cancer.gov/cancertopics/factsheet/Risk/artificial-sweeteners. Accessed April 28, 2011.

the characteristics of the audience practitioners?[4] How does the content vary in activation across those practitioners? What are the ordinary action routines of the class of individuals envisioned? What cultural, social, or societal challenges or privileges are in play? What is the typical technological environment, usage skill level, literacy? What are the interactive elements, shared experience, collaboration, intertextuality, social and cultural extensions? And, perhaps most important, does the analyst have a firm empirical basis for providing the answers?

To provide that basis, the analyst should consider starting the design process with an extended effort at observation of "behavior in the wild"; plan interviews with the target group(s); let focus groups reflect on the proposed design; and, finally, run a test group with full debriefing of the participants individually and together. This degree of careful work, I am sorry to say, is far more than what we typically see in the literature. I think we would be the better for it, nonetheless.

CREATING THE PROTOCOL

Creating experimental protocols is a lot like building custom furniture or sewing designer clothes. We all use a set of common tools and design elements, but the finished work is unique. The common tools in experimental design are the variables, the treatments or contrasts, the testing conditions, the selection and assignment of respondents, and the analysis—usually statistical. The experimental variables are the ones the analyst thinks have some causative effect, the criterion measure or measures are those that are suspected of being affected by the experimental variables, the treatments are the contrasting experimental-control comparisons that provide the test of the hypothesis, the testing conditions are the circumstances under which the respondents participate in the study, the respondents in experiments are almost always recruited rather than randomly selected but are then randomly assigned to one of the protocol groups, and the statistical analysis is usually analysis of variance (ANOVA), with one criterion measure; multiple analysis of variance (MANOVA), with two or more criterion measures; or analysis of covariance (ANCOVA), when one can test the effect of mitigating variables on the causal relationship or their (ANOVA/MANOVA/ANCOVA) general regression equivalents (whew). Let's walk our way through this list.

Variables

We start with variables in this discussion because the impetus for an experiment usually comes from an analysis of theoretical concepts or, perhaps more likely, the variable elements in an observed problem. In experimental design, we talk about variables that serve functional roles in the design and variables that serve as evidence for a claim (see below).

[4] Even our ordinary language for describing the relationship between media and user is stuck in early-20th-century conceptualizations of the audience as a passive receiver of media content. More and more, the end user is also the final moment of content production. One does not simply "receive" it anymore but also cocreates it.

There are four functional types of variables that can be brought into play in experimental design: experimental variables (the agent of the effect), the criterion measure (the measured effect), covariates (variables that presumably moderate the relationship between the experimental and the criterion), and control variables (potential contaminants of the criterion). Of these, the first two appear in every experimental design.

It is sometimes difficult to identify what variable serves what purpose in a complex design, as any variable could presumably serve any of the four purposes. Here are some rules of thumb to help in identifying a variable's role in a design. The experimental variable is the one that is manipulated by the analyst. If there is no manipulation, then the protocol is likely not a true experiment, but rather a correlational analysis. Sorting respondents by some preexisting condition (like education) is generally not considered a manipulation. Criterion variables are the ones that are tested for differences across the treatment conditions. There may be more than one criterion in a study, but mostly each criterion has its own analytical frame. (We don't see much MANOVA, but we do see replications of the same design using different criterion measures.)

Covariates and controls can look much alike. We have different kinds of interests in these two, however. We hold a theoretical interest in the consequences of covariate variables. We hope to demonstrate that *X* kinds of people or *Y* kinds of texts have consequences on the relationship between the experimental and criterion variables. We hold a cautionary interest in control variables. Our concern is that *Z* kinds of conditions (people, texts, circumstances, etc.) might contaminate the criterion such that we get a false reading of the relationship between the experimental and criterion variables. Analysts are an opportunistic lot. Disappointing covariates can become controls, and interesting controls can become covariates as the results begin to appear. (Squishy ethics? You bet.)

In media studies, there are six common evidentiary classes of variables: message (text), mode (medium or technology), audience (characteristics), reception (including issues of interpretation), interaction (Web 2.0),[5] and outcomes (cognitions, behaviors, etc.). Specific variables from these six classes can be used in any of the four locations (experimental, criterion, covariate, control) of the experiment, although clearly there are affinities between class and experimental location (e.g., outcomes are often the criterion).

How variables are used in what locations depends on the problem under investigation and on the theory in use. The analyst might be interested in the relationship between certain kinds of messages and the decision behaviors of the audience. In an effects model, the design would likely locate the message as the experimental variable, designating it as the agent of a decision, but a uses and gratifications model would likely locate the decision behaviors as the demand factor (agent) for certain kinds of messages, and a social action model would look at audience lifestyles as the experimental factor with both message selection and decisions as the criterion measures or in some combination of covariate-criterion configuration (e.g., certain lifestyles lead to different decision-making patterns given the availability of different technologies).

[5] Consider the common practice of real-time polling, for example, in which part of the media text is created by the audience texting responses to an on-air poll.

Let's look at a couple of examples from the literature: McKinney and Rill (2009),[6] working from the principle that the simple act of engaging campaign communication raises the level of civic engagement, set up a quasi–field experiment to test the effect of exposure to two kinds of debates during the 2008 primary and presidential campaigns. They used two outcome measures, an 8-item political cynicism scale and a 5-item political knowledge confidence scale in a preexposure-postexposure design.

Their experimental variables were exposure to a real-time debate and to the type of debate to which respondents were exposed. The exposure variable had only one value in that everyone saw a debate, but the nonexposed condition was inferred from the preexposure test scores. This design adequately isolates the effect of exposure, but is weak in supporting the importance of the content. What if control group members had watched a lively discussion of the importance of civic engagement? Would their political cynicism scores decline and their knowledge confidence scores increase as well?

The study also used a covariate variable of sorts. Respondents self-identified as committed to or leaning toward one party or another. They were then assigned to watch either the Democratic candidate debate or the Republican candidate debate according to the party of their self-identification. Here, too, one might wonder if there would have been more information gained concerning the initiating question if the groups had been split with half of each group randomly assigned to watch the opposite party debate. If participation is the basis of the effect, it might not matter if one is for or against the content. In that sort of design, party identity becomes a true (potential) covariate.

Grabe, Kamhawi, and Yegiyan (2009) designed a study to examine memory for news stories presented on television, in the newspaper, and on a news website by respondents with different levels of education and over different time periods. In this study, the criterion was a complex of three memory measures—encoding, storage, and retrieval—based on some form of written recall (recognition, cued, or free). The experimental variable was the presentation of topically equivalent messages across the three media. The covariates were education (no more than high school vs. postgraduate degree) and time of recall (immediate and two-day delay).

The researchers also added some control variables. To control for the possibility that there might be a content-medium interaction, they replicated the comparison over four different news stories. They also instituted a control for the order of the media the participant engaged, randomly assigning respondents to one of the possible sequences. And, finally, they collected information on the respondent's interest level for each topic in case interest affected recall. Although not fully reported, the authors apparently found no topic, sequence, or interest effects. Note that the authors had no expectation of a topic, sequence, or interest effect. They were checking for the possible contamination of the criterion by one or more of these effects. I would have preferred a full report of the finding, nonetheless.

Measurement Controls

Up to this point, we've been engaging the variables at the construct level. The empirical evidence for the operation of these variables will, of course, come from the measurement

[6] Read the full results available online at participating libraries everywhere.

process. We discussed the technical aspects of measurement in Chapter 5. Here we want to look at how the choices of measurement affect the design and possible outcome. Just to remind ourselves: The measurements we take are the variables we study. The whole enterprise hangs on the quality of those measurements, which in turn is dependent on the measurement process. Measurement controls, then, intercept that process in order to suppress the appearance of extraneous effects. Those effects would include fatigue, boredom, frustration—even sabotage, practice, telegraphing, pattern recognition, response demand,[7] the testing environment both physical and social, skill and literacy requirements of the measurement, and motivation of the respondent.

Further, measurement almost always involves some performance of its own that in turn demands its own skills and competence of enactment. Elements within the measurement set can interact with one another, setting up expectations for "what the experiment is actually trying to do" or for "what the right or desired answers are." As with the development of survey instrumentation, the best insurance is to carefully pretest the entire protocol, with special attention to the measurement process.

All of these effects are in play to some extent in every measurement situation. They are part of the noise in the system—the error component of measurement. What we control for is the systematic appearance of these effects. If nearly everyone hits a wall of fatigue at approximately the same place in the measurement process, then the measurements that follow are systematically affected by that fatigue. Further, if this effect is associated with one treatment group and not another, then the results of the comparison will be false—perhaps not fatally so, but false nonetheless.

In addition to these extrinsic and spurious effects, there are the intrinsic effects that are part of the measurement device itself. These are the standard test issues of item discrimination, reliability of the instrument, and the ever-elusive questions of validity. Generally speaking, these issues are dealt with prior to the experiment. But it probably doesn't hurt to remind ourselves of them here.

In even a cursory exploration of the literature, the first thing we notice is the heavy dependence on paper-and-pencil (P&P) measurements. Historically, this dependence developed because of the message effects orientation of mass communications and its confluence with the development of cognitivism, but it continues out of some practical considerations of the difficulties of measuring the actual behavior for which these P&P measures stand as surrogates. Nonetheless, they remain as Plato's shadows on the cave wall, mere indicants of possible material performance. And it is, of course, material performance that matters. (I consider the fact that we have not been able to advance our typical measurement devices in the past century our greatest failure.)

Paper-and-pencil measures do, however, allow us to conduct experiments over topics that would otherwise be difficult or ethically impossible to conduct. One cannot, for example, ethically conduct an experiment in which the criterion measure is actual physical harm inflicted upon another; even inducing a firm belief that one is harming another is

[7] These three effects have to do with test construction where a prior response telegraphs the answer to a following item, questions on different dimensions are in a recognizable pattern, or the items reveal the interests of the research in such a way as to "ask" for certain answers.

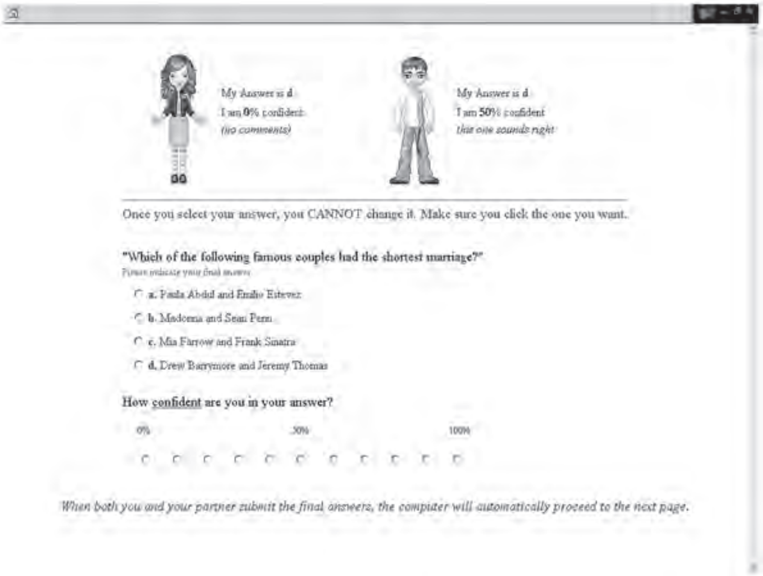
considered ethically suspect, as the response to the Yale studies showed.[8] And we cannot physically document that a condom was used following exposure to a message on safe-sex practices. (Despite their necessity, I would not trust the paper-and-pencil measures in either case, because violent and sexual behaviors are not under the governance of messages or even under the reliable control of language-based cognitive processes.)

The point is to use paper-and-pencil measures where we must but also to examine our protocols closely for measurement processes that allow the actual performance to be the criterion. For example, in Kelly Schmitt and Daniel Anderson's (2002) study comparing young children's behavioral learning from direct observation and from watching the behavior on television, the criterion measure was the actual performance of the activities to be learned (placing or finding a toy in another room).

Lee (2005) demonstrates that you don't have to lose any of the convenience of P&P measures with his study of gender roles in which he used a computer adaptation of a *Jeopardy* game (see Figure 11.1). This sort of computerized data collection has the added advantage of controlling for respondent and clerical input errors.

Whatever the instrumentation, measurement is not a neutral process. It often calls on the respondent to adopt an "as if" state of mind to render judgments as if the object or person were there; as if a single, global rating could stand for all the variations of conditions that an object or a person might encounter; as if the reality of the testing situation could

Figure 11.1 Computer Adaptation of *Jeopardy*



My Answer is d
I am 0% confident
(no comments)

My Answer is d
I am 50% confident
this one sounds right

Once you select your answer, you CANNOT change it. Make sure you click the one you want.

"Which of the following famous couples had the shortest marriage?"
Please indicate your final answer

○ a. Paula Abdul and Emilio Estevez
○ b. Madonna and Sean Penn
○ c. Mia Farrow and Frank Sinatra
○ d. Drew Barrymore and Jeremy Thomas

How confident are you in your answer?

0% 50% 100%

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

When both you and your partner submit the final answers, the computer will automatically proceed to the next page.

[8] http://www.garfield.library.upenn.edu/classics1981/A1981LC33300001.pdf. Accessed April 29, 2011.

be ignored. Consider how measurement participates in our understanding of the focal object. Presume that you are a researcher studying the online presentation of a research class. You are interested in getting some baseline information concerning the respondents' prior experience with online courses. You ask the "How many?" and "How often?" questions and then ask, "Overall, has your experience with online courses been very positive, positive, somewhat positive, neutral, somewhat negative, negative, or very negative?" There is a likelihood that you are causing the respondent to evaluate online courses in a way he or she has not done before. "Huh," the respondent thinks, "I never realized that I really like online courses. This really changes everything." And so it does.

Treatments and Equivalence

A treatment in an experiment is any set of researcher-determined conditions that is measured. Every experiment compares one condition with another. One cannot have an experiment without some manipulated comparison, which dismisses all those correlational studies from the realm of experimental evidence—parsing variance is not the same thing as comparing different conditions. In the next section, we will discuss the controls the comparison has to exercise; in this section, we are talking about the equivalency the comparison has to achieve. The basic form of the comparison is to set up identical conditions, both of which allow for the appearance of or change in the criterion event. In one set of these conditions, the experimenter introduces the variable thought to act on the appearance of or change in the criterion. Simple to write; hard to read; very difficult to achieve.

The design problem is complicated by the composite nature of our variables (here we go again). For example, let's say the analyst wanted to compare the effect of engaging a message, being exposed to that message, and the absence of that message on the perceived likelihood of some subsequent action, say wearing sunscreen while skiing (water or snow—your choice—and OK, snowboarding or skateboarding too). This question is a standard message effects question with the added dimension of an engagement-exposure comparison. The question of engagement versus exposure deals with the common criticism of message effects in that the laboratory setting encourages a heightened level of attention that does not occur in normal viewing.

In the engaged conditions, respondents are told that they will be eligible for a $20 drawing if they are able to correctly answer 5 questions about the message in order to simulate a prior interest in the topic and heightened attention during its presentation. In the exposure condition, respondents are periodically given a simple arithmetic problem to solve on an "at-hand" computer during the presentation of the message. This device allows for continuous aural attention but diverts visual attention to simulate a typical multitask viewing condition. In the absence-of-the-message conditions, one set of respondents (the engaged comparison) is told of the drawing but then not given the message (the screen remains blank because of technical problems), and another set (the exposed comparison) is given the periodic arithmetic problems but again technical problems with the television set prevent the showing of the message.

The design, then, has four independent groups with random assignment of respondents to group membership. It would be seen as a 2 × 2 design with *engaged-exposed* as one dimension and *message-no message* as the other. We could draw it up as Figure 11.2.

Figure 11.2 2 × 2 Design

	Message	No Message
Engaged		
Exposed		

The question we are engaging is whether these are equivalent conditions varying only across the engaged-exposed difference and the message-no message difference. Notice that we are not asking if they are the same. We know they are not the same. Equivalence, on the other hand, occurs within some framework. The framework for us is always the criterion measure. Our framework, then, is the reported likelihood of wearing sunscreen while skiing or boarding. Is there anything in the design that suggests that something other than the experimental variables will affect the outcome on this measure?

Beats me. I'd be a little concerned about the effect of frustration on the engaged-no message group, but that could work to motivate existing strategies for test success (what, you've never been unprepared for a test and still survived?) or to sit back and give up. So, maybe that's a wash. (Seizing a teachable moment here: In what order would the analyst administer the criterion measure and the content test? What would be the rationale for the order selected?) So this is it? This is the level of technical accuracy I can achieve in the design process? Initially, yes, technically we would say that it achieves face validity on equivalence or, in less technical parlance, "It looks good." The analyst doesn't have to settle for just face validity, and if one's job, degree, tenure, or funding were riding on the outcomes, it would certainly be worth the effort to pretest the design, using a small sample.

The results from the test run can be used to conduct a power analysis to determine the number of respondents needed to distinguish the effects. It can also be used to explore effective covariates (third variables that modify the relationship between the cause and the effect) and to improve on both the theoretical contribution and the sophistication of the hypotheses. Unfortunately, even a small sample of 10 per group results in a requirement of 40 respondents in this fully randomized design.

Comparisons and Contrasts

The design of the *contrasts*, which is the term used for the comparisons across groups we intend to make, provides the first level of control by comparing results across groups that are presumably under the same influences with the exception of the experimental variables. In our engaged-exposed example, the contrast between the engaged message-presented group and the engaged no-message group provides a control for the effect of the message; the contrast between the engaged and exposed message-presented groups controls for the engagement motivation; the contrast between the exposed message group and the exposed no-message group controls for the effect of the message when no engagement motivation is present; the final contrast between the two no-message groups would allow the analyst to explore the incitement effect (possibly confounded with a frustration effect) of the monetary reward.

What might be added to this set of controls would be a typical message-presentation group (no motivation or distraction) with a typical no-message control group. In this addition, the analyst would be anticipating concerns that the original design would not permit the "clean" extraction of either the engagement or the distraction conditions. The addition would, however, raise the costs of the study by nearly one third.

In considering treatments, we generally talk about experimental and control conditions. An experimental condition is one in which the variable under study is present at some value. A single variable can produce more than one treatment. Weaver and Wilson (2009), for example, posed three levels for their experimental variable, quality of depicted violence—none, sanitized, and graphic—to determine the consequences for three criterion measures, enjoyment of the program, emotional reactions to the program, and judgments on the content (violence, graphic quality, level of action). The variations were produced by editing a set of five programs to remove all depictions of violence (action without violence conditions) and retain only scenes with no blood or gore (action and sanitized violence), and by no editing of violent scenes (action and graphic violence).[9]

The five programs are also a variable in this study, a control variable in this case. They serve as a control for the possibility that the effect of the graphic quality of violence will be changed by the story line or characters in a given program. The combination of three levels of violence and five programs actually produces 15 different treatment conditions. Respondents were randomly assigned to one of these conditions.[10]

[9] Action is preserved in each of these conditions because of Zillmann's (1991) arguments about the effect of arousal in television viewing.

[10] Weaver and Wilson also tested the effects of sensation seeking, trait aggression, and prior exposure to violent programs. These are not treatments, because respondents come to the study with these conditions; they are not varied by the researchers, only measured.

The development of treatments is a substantive part of experimental design. Weaver and Wilson used selections from *24*, *The Sopranos, The Shield, Oz*, and *Kingpin*, which all follow the same basic story line of tarnished good versus burnished evil presented in gritty, screen noir. This choice is a deliberate one to contain the effect of the program on the criteria, which at this stage of investigation is perfectly appropriate. Thoughtless choices can set one up for failure, and as long as one's conclusions recognize the scope chosen, a reduction of risk to the hypothesis can still produce a fair test. On the other hand, a narrow set of conditions cannot be matched to a broad set of conclusions.

Respondents

To meet the requirements of a fully randomized design of either the Weaver and Wilson example or our sunscreen example, respondents have to be randomly entered into one of the independent groups. We know that true random samples of human respondents are very difficult to obtain. The usual solution is random assignment. Respondents are recruited from a common pool (say, the research requirement of an introductory communication course) and are then randomly assigned to one of the treatment or control groups. In both of our examples, a given respondent would participate in only one of the possible treatment or control groups. Random assignment of this sort is a technique that eliminates cross-group, intrarespondent effects as respondents appear only once. It also limits the influence of interrespondent differences by distributing those differences more or less equally across the treatment or control groups. Usually, this assignment is done by randomly selecting a starting group for the first respondent and then following a standard assignment sequence for all subsequent respondents.

Random assignment, of course, does not guarantee equal distribution of unknown covariates, but it does provide the strongest argument against the presence of some systematic effect. I consider it stronger than balancing treatment groups across selected demographic variables such as sex (not gender in this case) and age. Balance in the absence of evidence of some interaction is only for show.

Random assignment can offer no control over the differences the respondent pool may show in relation to the general population. Major areas of academic study are chosen by students for some reason, and fewer than half of U.S. citizens attend college, so differences clearly exist both between different groups of college students and between college students and the general population. I have no evidence, however, that those differences affect the outcome of the experiment. (If I did, I would also know the outcome of the experiment.) We saw in Chapter 8 that a random sample from the general population is mostly out of reach except for the well-funded few. Typically, the analyst would have no access to a general population pool, which renders the question (but not the criticism) moot at any rate.

Other experimental designs allow for other methods of controlling respondent effects. Repeated measures designs where respondents participate in more than one treatment group eliminate interrespondent effects but raise the likelihood of intrarespondent effects (covariates, practice effects, boredom, sabotage, etc.). Our sunscreen example has no control for topic. Perhaps some substantial portion of the pool disconnects from skiing of any sort. We could add two other safe-practices topics (automobile seat belts and bicycle helmets; we already did condoms). Each respondent would be randomly assigned to one of the message, no-message, engaged, or exposed treatment groups (one of the six combinations, that is) as before but would now complete the tasks across the three topics. Cross-respondent effects would be controlled over topics, but an order effect would be introduced. That effect could be controlled by randomly selecting the starting topic and then following a standard sequence so that each topic has an equal number of respondents who took it in the first, second, and third order position.

As we can read, respondent controls get complex quite quickly. Table 11.1 offers some order. We ought to offer a word on the "college sample." The purpose of conducting experiments is to create a generalizable claim about some part of the world in which we live. The more narrow the variables, measurements, and treatments, the less generalizable the information. This relationship also holds true for the issues that surround our respondents. Academic research has long been subject to the complaint that it relies upon a respondent pool that is demonstrably different from other subgroups of the general population, and yet offers conclusions that entail all of us. The simple answer to the complaint is "Yes, that is true—caveat lector." Any attempt to justify a college student sample as representative of the general population that does not use an authentic general population sample for comparison is just smoke and mirrors. The current college graduation rate in the United States is 25.6%, ranging from 45.2% in Washington, DC, to 17.0% in West Virginia.[11] Is Washington, DC, different from West Virginia? You draw your own conclusions.

Table 11.1 Respondent Designs and Effects Controls

Respondent Design	*Controls for:*	*Has No Effect On:*	*Additional Control*
Random sample	Pool effects	• Sampling error • Unequal distribution of unknown covariates	Random assignment to treatment groups
Random assignment from respondent pool	• Cross-group intrarespondent effects • Reduces probability of interrespondent effects	• Pool effects • Chance group inequalities	None
Repeated measures (with random assignment)	Cross-group interrespondent effects	• Sampling error or pool effects • Order effects • Chance group inequalities	Random assignment of order

[11] http://www.census.gov/acs/www/Products/Ranking/2003/R02T040.htm Accessed January 6, 2010.

Well, so what? The end user of the research has to exercise care in determining the amount of risk that is involved in applying a conclusion drawn from a limited sample to his or her particular situation. But that does not mean that academics should stop doing research or that all academic research is suspect. Total ignorance is not better than some information. You do have to think about it, however. Is an 18- to 24-year-old age group, regardless of education level, the same as a 45- to 64-year-old age group? I have never known these two groups to coidentify. Further, only 7% of the males in the younger group have a bachelor's degree, but slightly over 30% of the older-group males do. The rate for females is 11% and 27%, respectively.[12] What do these different educational attainment levels over sex and the reversal over age tell you? They tell me that the 18- to 24-year-old cohort is not the same as other age cohorts and not even the same through time. If my job depended on either extrapolating information from a college sample or collecting a sample of 45- to 64-year-olds, I'd get the new sample.

Finally, most experimental protocols present the need for methods in managing respondents from first contact to completed data set. These are not trivial issues; not only does the experiment depend on the respondents being there in a timely fashion, but also getting them there can be a major expense, being sure that they respect the work can determine outcomes, and their negative word of mouth can bring recruitment to an abrupt end. The entire process should be scripted or flowcharted. Entries should include timing of the elements in recruitment; method and message of first contact; scheduling and reminders; the informed consent process; dealing with people who are early or late; the step-by-step movement of individuals through the protocol, including the timing and location of these steps; debriefing; payment (money, course credit, eternal gratitude); and dismissal. There is a great deal of preplanning as well as disaster preparation that is needed.

Analysis

Nearly all analysis of experiments involves metric measurements and, therefore, statistical analysis. Remember that statistical analysis is both a way of displaying information about the characteristics of a data set and a way of providing a public decision making process for determining how we are to act toward some finding. We should find both analytical frameworks in any experimental study. The responsibility of the analyst is to provide sufficient descriptive information that the inferential tests can be placed in proper perspective and to conduct a sufficient number of the appropriate inferential tests such that the reader can be confident in the sources of the data events.

Descriptives

You know from the chapter on statistical analysis that I am a great fan of descriptive statistics. Let's take a specific example by using the report by Weaver and Wilson (2009) to show you why. They reported on the relative enjoyment of television programs edited

[12] Surprised? Check it out at http://factfinder.census.gov/servlet/DTTable?_bm = y&context = dt&ds_name = ACS_2007_3YR_G00_&CONTEXT = dt &mt_name = ACS_2007_3YR_G2000_C15001&tree_id = 3307&redoLog = true&_caller = geoselect&-geo_id = 01000U.S.&-search_results = ALL&-format = &_lang = en. Accessed April 28, 2011.

to display no violence, sanitized violence, and graphic violence. Let me just take one of their major results:

> An ANOVA on the two-item enjoyment measure revealed a significant main effect for treatment condition, $F(2, 8) = 10.24$, $p < .01$, $\eta^2 = .03$. Both the graphic version ($M = 2.29$a, 95% CI = 2.21, 2.38) and the sanitized version ($M = 2.29$a, 95% CI = 2.20, 2.38) were enjoyed significantly *less* than the no-violence version was ($M = 2.54$b, 95% CI = 2.45, 2.63). Thus, to answer research question 1, violence had a negative effect on enjoyment when action was controlled.
> (p. 454)

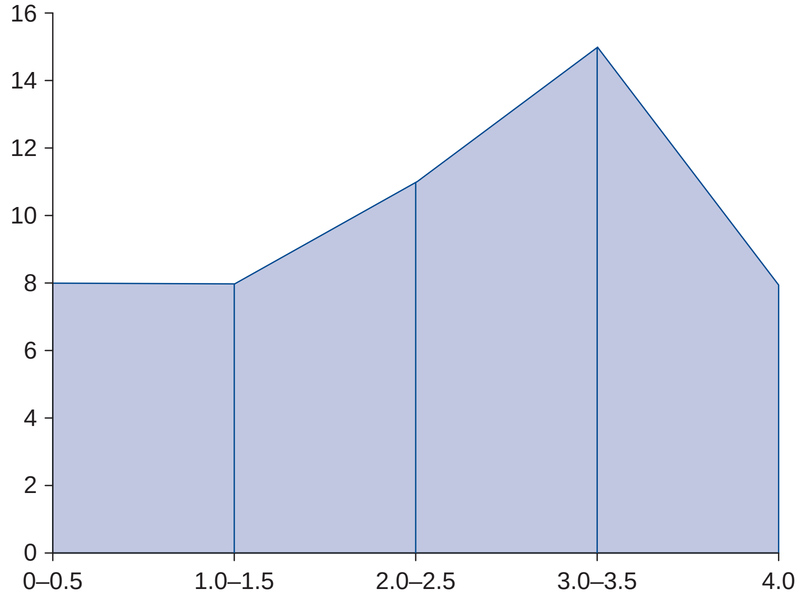
To begin our critical analysis of these three sentences, we would probably note that they are nearly unreadable with a great deal of information compacted into codes and abbreviations. Nonetheless, they follow the standard pattern of presentation. The first sentence reports the test of significance and tells us that under conventional rules we are justified in considering the difference among the levels of depicted violence across the criterion of enjoyment to be of import. The next sentence reports the descriptive data. It provides the three mean scores (2.29, 2.29, 2.54) and a measure of dispersion. The measure of dispersion is not the standard deviation or the standard error of the mean but rather a confidence interval (95% CI) that is based on an error term that comes out of the ANOVA. The 95% confidence interval is the range in which 95% of the scores can be expected to fall. The endpoints of 2.21 and 2.38 are reported for the first mean. What counts here is that the high end of the confidence interval is smaller than the mean for the no-violence condition, indicating that we are looking at two different sets of values. The last sentence is the implication of the result or the finding.

What work does the reader have to do to get an understanding of what is being said and what is very much missing? The first thing we have to do is to understand the criterion measure. On pages 451–452 we are told that the criterion measure is composed of two questions: "How much did you enjoy this program?" and "How entertaining was this program?" The two items were averaged on a 5-point scale that was anchored at 0 instead of the more conventional value of 1. That indicates that the expected mean would be 2.00 and not 3.00 as might be expected, a point easy to forget while reading.

Much further down (p. 456), we discover that the reported means are summed across the five programs. Consequently, the average reported is an average of an average of an average. The law of central tendency tells us that each of these consolidations will reduce the dispersion of the scores. The error term reported (about a .048 standard error of the mean) is very small but tells us little about the dispersion of the actual data.

In order to discover what a data set for a particular program might look like, I created a dummy set of data of 50 scores based on the average of a two-item 5-point scale with an averaged mean of 2.29 and a correlation of .82 between the two items to duplicate those two findings of Weaver and Wilson. I started with a set of random numbers from 0 to 4 in two columns of 50 each. I shifted a few numbers up in the first column to get a mean of 2.28. I then adjusted the numbers in the second column to get a correlation of .82. Figure 11.3 shows the distribution of scores.

Figure 11.3 Distribution of Scores From Simulation of Weaver and Wilson Data



Remember that these are manufactured data, not data from Weaver and Wilson. The first thing we see is that the numbers are skewed left (longer negative tail) and that the mode is between 3 and 4. Even though this mean score is for the lesser-liked depictions, most people could like it "a lot." The standard deviation for this data set was 1.33.[13]

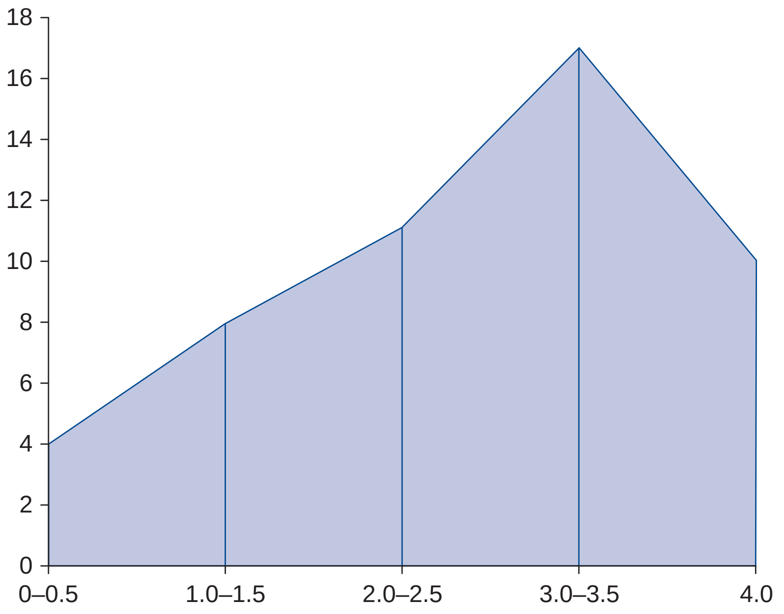
To continue this analysis, I adjusted the values of the two columns to get an averaged mean of 2.54 (the equivalent of the "no violence" condition) while maintaining the correlation of .82. Figure 11.4 shows that distribution.

The slope of skew gets sharper as I decrease the low-end values and increase the number of high-end values. The standard deviation of this data set is 1.22. It drops because the data have to become more compact as the respondents like the program more (ceiling effect), thereby increasing the likelihood of significant differences.

If Weaver and Wilson had given us this level of information, would we have reached a different conclusion than they did? I think we would have from the more relevant measures of dispersion but particularly from the histograms. The quoted data report a difference between the no-violence and two depicted-violence program sets as .25 or one fourth of one step on their enjoyment scale. Their conclusion was that violence had a generalized negative effect on enjoyment. That conclusion implies that there was a

[13] The value is similar to the standard deviations reported on p. 453. The standard deviation was for the first column 1.29 and for the second 1.49. The effect of averaging is to approach the lower value.

Figure 11.4 Adjusted Distribution of Scores From Simulation of Weaver and Wilson Data



lowered probability of low scores and/or a higher probability of high scores producing a global effect. In fact, no score can shift by one fourth, so the actual effect has to occur in relatively few data points. In the manufactured data set, it took about eight changes to adjust the mean and another four to recapture the correlation. More than three times the number of values remained the same as had to be changed. The Monte Carlo findings suggest that the entire effect may be the result of a few respondents holding particular attitudes about portrayed violence. We are back to the some-some conclusion with no evidence of a global effect.

If the means are significantly different, does it matter why they are significantly different? This Monte Carlo work gives us a resounding “Yes!” for the answer. But no reader can discover the actual facts unless the analyst provides the basis for them. Unfortunately, editors of carbon-based journals have a limited resource of space. They are often misguidedly ruthless in cutting the most valuable information and focusing on the inferential statistics instead. My recommendation, therefore, is to write a complete report, if for no one else than yourself, in which you provide both the inferential tests of the hypotheses and the complete set of descriptives. It is important to write it up, not just look at the results. Demonstrating the results and their implications in writing forces the writer to pay attention to the details so often missed but that lead to greater insight and, I think, more realistic conclusions.

Inferential Statistics

When the term *experimental design* is used, it most often refers to the design of the statistical analysis, not to the design of the protocol (which includes the statistical analysis). The complexities of statistical design have to do with the isolation of the treatment effect(s) and the calculation of the proper error terms and their conditional corrections. For nearly all of us who use statistics, these issues have been resolved in whichever statistical software package we use. Statistical design in a practical sense mostly means fitting our respondents, the criterion measure, and the experimental and control variables into one of these preexisting models.

Respondents in these models are categorized in one of two ways: in independent groups or as participating in repeated measures. An independent groups design such as that used by Weaver and Wilson employs a separate group of respondents for every treatment. Independent groups are called for when there is the possibility of contamination of responses by some previous activity in the experiment. Weaver and Wilson may have been concerned that watching several programs of that genre would lead to emotional or enjoyment fatigue for the later programs viewed.

Nonetheless, Weaver and Wilson could have respondents view one graphic, one sanitized, and one no-violence segment, randomizing the order and using different programs for each edit type. That design would employ a repeated measures factor. The supposed advantage of repeated measures is that they control for differences across subject groups that might contaminate the treatment effect. The repeated measures design suggests that the same set of respondent values is in play during each presentation type because the same respondents are involved. The choice between independent groups and repeated measures should be all about controlling the most likely sources of contamination. But, repeated measure designs are also ordinarily less costly to run as they require fewer respondents.

The criterion measure is the variable that is measured across all respondents regardless of conditions. The criterion in the Weaver and Wilson example we used was the two-item enjoyment-entertainment scale. Everyone completed those two items, no matter which of the edited versions each respondent saw. It is the criterion measure because the differences that occur across it are the basis of the claims of an effect by an experimental treatment and/or control variable(s).

Experimental variables are the ones controlled by the experimenter and are the center of the hypotheses to be tested. Control and covariate variables are typically characteristics, conditions, or states that the respondents walk into the experiment with. The terminology can be a bit confusing as we also talk about control groups or control conditions, which are actually one of the contrasts in a treatment set. A control group is often one where the manipulation of some treatment is absent; its value, however, is still manipulated by the experimenter (just set to zero).

Control and covariate variables can be exactly the same variable, and both function in the same way in an ANOVA design, but they differ in the role they play in the research argument. Analysts are not very consistent in the use of the terms *control* and *covariate*, but generally we use the term *control variable* when we expect the effect to be the same across the different values of the variable but of different magnitudes or at different starting places. The variable serves to "control" the variability of the scores (which would reduce power) by separating out the variance due to differences across the control variable. The control variable does not have (at least initially) theoretic implications.

We use the term *covariate* when the variable is expected to play some part in the hypothesis. In these cases, we have some theoretical basis for our expectation of difference. In the safe-sex message study, sensation seeking was theoretically linked to the likelihood of unsafe sexual behavior and to receptivity to more sensational messages. The more sensational message was hypothesized to be more effective for those with higher levels of sensation seeking.

Both control and covariate variables can be introduced to the design in a serendipitous, "just in case" sort of thinking. Weaver and Wilson add an inventory of shows watched "to control for prior exposure to violent media." They offer no justification for doing so or any prediction as to effect, but they also don't need to, unless, of course, it turns out it makes a difference. In that case, they would be caught in post hoc, catch-up reasoning. It was not significant. They also collect a fairly large number of other demographic and psychographic variables that were not reported in the study, but were undoubtedly examined.

Weaver and Wilson used sex of the respondent as a fully preplanned covariate. They hypothesized that males and females would enjoy violent and nonviolent content differently based on some preliminary findings from previous research. It too was not significant. I suspect that here was a case where gender and not sex is the operating variable. A measure based on masculinity or femininity rather than physiology might have been more effective.

An Invented Example

Let's see how it all gets put together by going back to our engaged-exposed example where we were considering the effect of sunscreen-use messages on those participating in outdoor activities with high sun exposure such as those based on snow, water, and other full-sun environments and/or at higher altitudes with less atmospheric protection.[14]

We would want to identify our respondents based on their participation level using some measure that would divide the group into high, moderate, and low participation segments. The criterion measure would be a measure of effectiveness, perhaps a combination of a P&P intent-to-use measure with the actual purchase of a sunscreen product from a kiosk outside the experimental setting. We would want to develop three to five public service announcements (PSAs), all based on the same content elements but with different production values. (Why?) We could also use different types of messages—fear or a threat, humor, social responsibility—but each would have to have multiple examples. And, we would need four treatment conditions—engaged, exposed, and two no-message conditions for every message variation we used. The study gets very big, very fast.

We are now faced with some decisions. The first has to do with how to use our respondents. Can we control for intergroup differences and gain some efficiency in cost and time by using repeated measures? Perhaps respondents could randomly cycle through engaged, exposed, and no-message treatments. That approach would cut the number of respondents

[14] I really wanted to write something here about the difference between the skilled grace of skiers and the knuckle-dragging character of boarders, but my reviewers wouldn't let me.

for each control group in half, but, unfortunately, the subterfuge of a technical failure might influence the two other treatment performances when the no-message condition is encountered first in the sequence. The danger of contamination seems too high to me.

I would recommend using different types of messages as there is a strong theoretical base to support it. It would give the study more appeal to editors and potential sponsors. Again, however, I don't think that respondents can evaluate more than one form because of the attenuation effect of message repetition.

Where we can use repeated measures is with the criterion variables. We have three criteria: content acquisition, intention to use, and a likely-to-purchase measure collected at the sunscreen kiosk. Everyone would complete all three criterion measures, but independent groups would be used for every other treatment, control, or covariate condition.

Let's say we decide on three message types, each with three examples. We then have three respondent groups sorted by level of participation, by three message types, by three examples, and by two engaged conditions (message and no message) and two exposed conditions (message and no message) with the three criterion measures (content acquisition, intent to use, and subsequent purchase likelihood). Consequently, we have three instances (one for each criterion) of a 3 × 3 × 3 × 2 × 2 for a total of 108 groups. If we put 15 respondents in each group, we would need 1,620 individuals, each of whom would have to sort nicely into one of the 108 participation groups. Good thing we have a big grant.

This is an enormously complex design,[15] but we can begin to reduce the complexity if we combine the three examples of each message type. In order to do that, we have to show that the effects of message type are stable across the three examples. Unfortunately, that puts us on the "wrong" side of the way inferential tests work. You may remember that tests are typically designed to make it difficult to show differences in order to protect from Type I error (accepting the research hypothesis when it is false). Consequently, the ordinary tests are set up to show no difference, which is what we would want in this case. It follows the old rule of thumb: "Don't ask questions you might not like the answer to." Even though this is a standard way of dealing with issues like these, we have to do better. The least we can do is to set our alpha level at something like .10 or .20. Much better would be to establish the specific conditions under which we declare the set of individual PSAs within each type to be consistent prior to testing to reduce our dependence on "wrong-sided" statistical testing.

The practical pressure to use the easiest test for "sameness" is quite strong, because the cost of finding difference is quite high. Weaver and Wilson used five different exemplars in each of their three violence edits. They found no significant interactions over the message factor (p. 456). But the significance level used was appropriate to the control of Type I error and not the Type II error they would be in danger of committing. They provide no graphs and indeed report no means. It is a typical approach. As readers, we can usually depend on the combined effort of analysts, reviewers, and editors to ensure that the right inferential tests are run, but only if the right questions are asked and asked in the right way.

The task of any protocol design is to answer the questions posed by the problem at hand and to generate additional information so that new problem statements can be formed. The task of any analysis is to definitively extract the answers and to mine the data for all that might

[15] And, frankly, very slow reading. The best way to work through this section is to run the Monte Carlo experiment yourself.

be of value. To demonstrate how one goes about these tasks, I built a data set for one portion of the engaged-exposed experiment, setting terms for what I would expect to be the intent outcomes for the engaged conditions. These terms can be considered hypotheses, which themselves would ordinarily be built on the careful review of previous research. I was working from what I know about the literature, personal observation in working with outdoor participants, and the goals of this section—a mixture of literature, plausible guesses, and pedagogical intent.

The requirements for the data set were as follows:

1. High participation levels would be associated with low intent for sunscreen use, and low participation levels would be associated with high intent.
2. All message types would show higher intent to use scores than the no-message condition.
3. Humor would be the most effective message type, personal safety the least, and fear intermediate.

In order to meet these requirements, I changed the permissible range for the random numbers to be generated. For example, in the no-message data, I truncated the upper values to ensure a lower mean score. Even with this manipulation, the data set could still show plenty of surprises. Let me remind you (as I remind myself) that none of our data or findings is real. The purpose is to show the techniques of analysis.

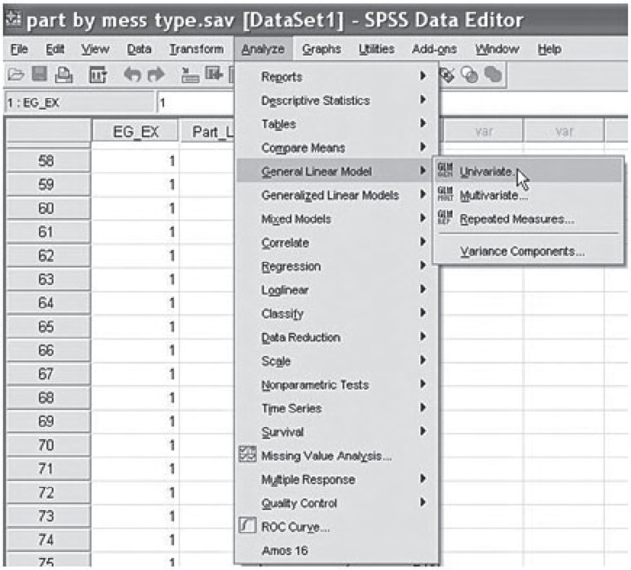
The data set produces a 3 (levels of participation) by 4 (message types—fear, humor, personal safety, none). The requirements do not specify an interaction between participation levels and message type, although the method of data generation allows one to appear. I use SPSS as my statistical application. As in all SPSS analyses, each row in a data set represents one respondent. Nominal values identify the participation group (1, 2, or 3) and message type the respondent was assigned. Table 11.2 shows the transition point from participation group 1 to participation group 2 with the message type resetting to the first type. With this data set, I would go to the “Analyze” menu and would select a univariate analysis under the General Linear Model (Figure 11.5). That selection generates a sorting table that allows me to enter the data into the analysis, according to its participation by message type source (Figure 11.6).

Intent is entered as the dependent variable as it is our criterion measure. Participation level and message type are entered as fixed factors because they have been manipulated

Table 11.2 Data Set Showing Transition Point

Part_Level	*Messge_Type*	*Intent*
1	4	2.00
1	4	5.00
2	1	3.00
2	1	6.00

Figure 11.5 Selecting Univariate Analysis From SPSS Menu

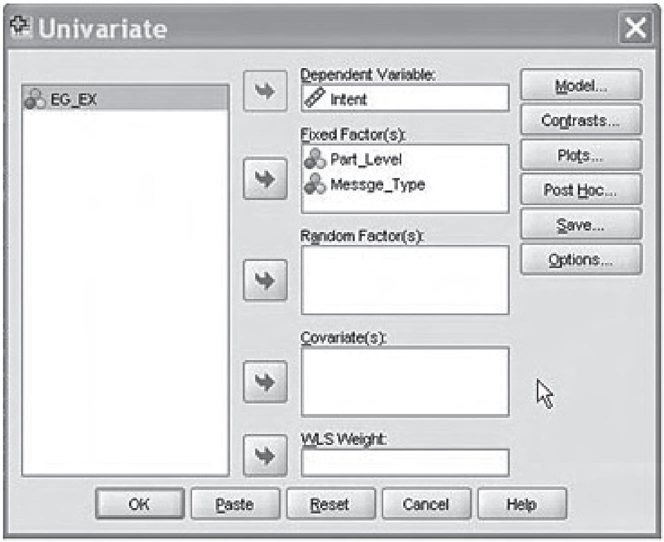


part by mess type.sav [DataSet1] - SPSS Data Editor

1 : EG_EX 1

	EG_EX	Part_L
58	1	
59	1	
60	1	
61	1	
62	1	
63	1	
64	1	
65	1	
66	1	
67	1	
68	1	
69	1	
70	1	
71	1	
72	1	
73	1	
74	1	
75	1	

Figure 11.6 SPSS Dialogue Box for a Univariate Analysis



Univariate

EG_EX

Dependent Variable: Intent

Fixed Factor(s): Part_Level, Messge_Type

Random Factor(s):

Covariate(s):

WLS Weight:

by the analyst (random factors are not manipulated). The table allows a number of options, which are beyond our scope here except to say that I always collect descriptive statistics and plot the simplest (lowest-level) means.

After I have confirmed that the data were entered correctly (using the methods described in Chapter 9), I then go to the inferential tests to check significance. The arguments that I can create for the results are different depending on how those tests turn out. If nothing is significant (and I still believe in the hypotheses), then the lesson to be learned is about the design of the protocol that led to the failure. The requirements for this data set establish two main effects hypotheses—there will be differences across participation levels and message types—and specify certain differences among levels and types. Consequently, if the results show a significant level effect and a significant message effect, and no interaction, I have the first level of support for the hypotheses and would move to an investigation of the main effect means to see if the specifics hold up as well.

In our case we have a significant interaction that indicates that the effect of level varies across different message types and that the effect of message types varies across participation levels. It is not legitimate (but also not unheard of) to discuss main effect means given this outcome. In short, my hypotheses failed, but in an exciting and informative way. These results suggest that I will be able to refine, not simply support, existing theory. Table 11.3 shows what the ANOVA table would look like. The far-right column indicates the significance level (some value beyond .000).

At this point, I go to the plots of the simple means as presented in Figure 11.7. The lines are participation levels, and the points are message types. The interaction is shown at the points where the lines cross or substantially deflect from parallel.

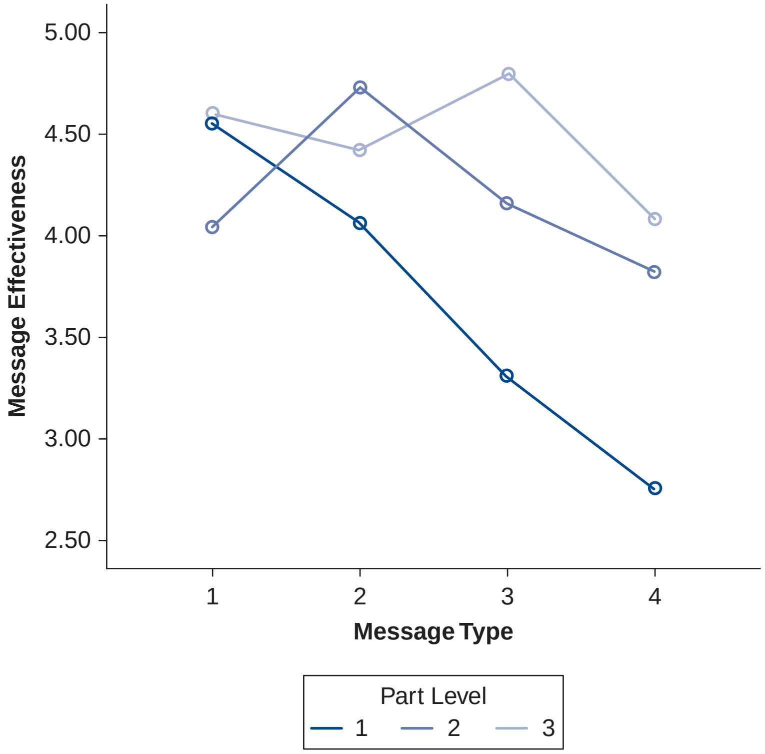
Table 11.3 ANOVA Results for Message Type by Participation Level Over Message Effectiveness

Dependent Variable: Intent					
Source	*Type III Sum of Squares*	*df*	*Mean Square*	*F*	*Significance*
Corrected model	178.420	11	16.220	7.903	.000
Intercept	9134.891	1	9134.891	4450.744	.000
Part_Level	59.959	2	29.980	14.607	.000
Messge_Type	64.850	3	21.617	10.532	.000
Part_Level by Messge_Type Interaction	53.611	6	8.935	4.353	.000
Error	1083.699	528	2.052		
Total	10397.000	540			
Corrected total	1269.109	539			

The analyst is charged with explaining this complexity with an appropriate set of mean scores and inferential tests. The initial explanation has to account for each point and each line on the plot. For example, early questions might include "Is the control group significantly lower than any message group for each of the participation levels?" "How does each message type vary significantly across participation levels?" But there are also mysteries to be sorted out such as "Why does the intermediate group not act like an intermediate group?" "Why are fear appeals so effective and humor appeals relatively ineffective for high-participation respondents?" (Remember these data and findings appear more realistic than they are.)

The analyst has to throw off the limitations of the hypotheses to conduct this exploration. The hypotheses have failed. The effort, now, is to develop better hypotheses, not to defend the failed ones as is ordinarily done in the literature. In the end, the analyst will have something to say about every meaningful combination of lines and points. How much of it ends up in a particular report will depend on the purposes and audiences for the report. Regardless, the analyst has to be secure in the totality of the analysis—that every valuable question has been asked and explored—even if uncertain as to the answers.

Figure 11.7 Message Type by Participation Level Over Message Effectiveness



THE ETHICS OF EXPERIMENTS

The ethical principles governing the design of experiments are threefold: The design must allow the research hypothesis to both meaningfully fail and adequately succeed; treatment and controls must not pose nonconsensual or inappropriate risk to respondents; and the design must engage the contexts of its conclusions.

The research hypothesis can meaningfully fail when the treatments and controls create a fair test. Weak or meaningless controls give the appearance of a test without actually challenging the treatment. Or a control might produce its own effect as the potential frustration induced by the no-message condition in our example above.

In most cases, communication experiments involve little more than everyday risks, but there are some interesting issues. What about exposure to content that is presumed to have negative consequences for the viewer? Researchers have testified that every exposure to violent content has a negative effect in the same vein as secondhand smoke. How do they justify another experiment? Physiological measures such as eyemarker cameras can cause damage in combination with underlying conditions that the researcher may not be qualified to evaluate. Deception that is not adequately debriefed may leave the respondents with a false assumption about a class of people or about themselves.

Finally, experiments must achieve a minimal level of ecological validity to permit the transfer of findings from the laboratory to naturally occurring contexts. The production of the evidence for that level of ecological validity has to be part of the overall design. Speculation that it might transfer is no conclusion. We can all speculate without running an experiment.

CONCLUSION

The design of experimental protocols represents some of the most demanding work in research. The researcher has to create the conditions under which we can have confidence not only in the testing of a causal relationship, but in that the relationship will generalize to the conditions found normally in society. It is work that requires a great deal of technical skill in measurement and metric analysis as well as creative solutions to the problems of ecological validity. Ofttimes they involve managing large numbers of people in protocols that allow respondents to return good information, that make efficient use of participants' time, and that in themselves do no harm.

The work starts with a problem that somehow implicates a causal relationship between an agent or agents and the context in which the agent is located that might include moderating conditions, preexisting states, and/or concurrent cognitive and actional requirements. The protocol sorts all these out into the components of theoretical concepts and constructs, variables and their measurement, treatments, controls, respondent requirements and assignments, statistical analysis, and, to close the circle, the implications for theory.

MOVING ON

This concludes the center of our engagement with metric methods. The next two chapters begin the transition to interpretive methods by taking up the analysis of content and texts.

REFLECTIONS

What Are Some Points to Remember?

- Experiments are based on a deductive approach based on theory and previous research. The stronger that theory and previous research, the better the experimental design.
- The experimental protocol has to isolate the relationship between the independent and dependent variables. The more complete that isolation, the more secure the conclusion. The greater the isolation, however, the greater the likelihood that the conclusions will fail to show ecological validity and will not translate to actual conditions.
- Good statistical analysis is not simply a report of the tests of significance. It involves the use of all available statistical tools to achieve a thorough understanding of how the treatment works.

Why Does It Matter?

Experimental designs are intended to draw instrumental conclusions. Instrumental conclusions are those that are intended to provide guidance for or to direct the actions of significant social actors such as parents, teachers, and policy makers.

What Else Could We Talk About?

A survey of high school students from 20 Cleveland-area high schools reported in late 2010 that hyper-texting—sending more than 120 text messages per day—was associated with increased likelihoods of smoking, drinking, and sexual activity. Hyper-texting was also associated with increased likelihoods of being female, a member of a minority, of lower socioeconomic status, and in a single-parent household with a missing father. The lead researcher concluded in a Case Western Reserve University School of Medicine press release, "The startling results of this study suggest that when left unchecked texting and other widely popular methods of staying connected can have dangerous health effects on teenagers."[16] This study poses significant issues both to a methodological critic (surveys posing as experimental evidence) and to a social advocate (ethnic, racial, class, and gender typifications).

What Else Might Be Interesting to Read?

Slater, M. D. (1991). Use of message stimuli in mass communication experiments: A methodological assessment and discussion. *Journalism Quarterly, 68,* 412–421.

[16] http://case.edu/medicus/breakingnews/scottfrankhypertextingandteenrisks.html. Accessed December 12, 2010.