Experimental Design

Steven Bellman
University of South Australia
Steven.Bellman@marketingscience.info

Word count (not including abstract): 9,023

Abstract

Well-designed experiments provide the best possible data for showing that one variable *causes* another variable. Causative explanations are the building blocks of theory in communication research and science generally. Experimental design is what researchers do before carrying out an experiment. This chapter describes the differences between the two basic kinds of experimental design, between-groups designs and within-participants designs. It provides guidelines for choosing between these two types of designs, or for using both in a mixed design. To make a strong case for causation, a well-designed experiment has to rule out alternative plausible explanations for the experiment's results that threaten its internal or external validity. Finally, this entry discusses how new methods allow communication researchers to carry out experiments with high levels of control outside the unnatural environment of a lab. These discussions are illustrated with examples of real-life experiments conducted in a busy media research lab.

If you are reading this entry, it's probably safe to assume that you are open to idea of experimenting on people, with ethics review board permission of course. You agree that positive, quantitative, evidence-based communication research has something to offer. Experiments are a special kind of quantitative research, and experimental design is what researchers do before carrying out an experiment. Case studies can suggest theory, and surveys and analyses of secondary data can show how the variables in a theory correlate with each other. But well-designed experiments provide the best possible data for theoretical arguments about causation: that one variable *causes* another. Of course, science is an inductive process, so no experiment can actually prove causation, it can only disprove it, although it can take several experiments to convince the scientific community that causation has been disproved (Campbell, 1988; Okasha, 2002). Nevertheless, experiments provide the best possible evidence for these kinds of arguments, which are essential to the development and refinement of theory. And often experiments reveal unexplained findings that trigger new theory, to be tested in further experiments (Smith, 2000).

This entry describes the differences between the two basic kinds of experimental design, between-groups designs and within-participants designs. It provides guidelines for choosing between these types of designs, or for using both in a mixed design. It also describes the various threats to validity and sources of noise that a well-designed experiment seeks to minimize. Finally, it discusses the various options for carrying out experiments that are available nowadays, allowing communication researchers to conduct experiments with high levels of both internal and external validity. Experiments used to be confined to the lab, but with online questionnaires, Web cams, mobile phones, and smart watches, the potential for carrying out highly controlled experiments in the field has expanded. These discussions are illustrated with examples of real-life experiments conducted in a busy media research lab.

The Point of Experiments: Arguing for Causation

The whole point of carrying out experiments is to make the best possible argument for causation. One good experiment is worth more than a huge "big database" of field data. This is because data from the field may be correlated, but may not be causally related. Armstrong (2012) provides an example from World War 2. In 1944, a computer built in Harvard's air-conditioned gymnasium was used to identify the best mix of metals in the alloy used to make turbine engines. The new mix was predicted to last several hundred hours at high operating temperatures. In test experiments, the new alloy lasted for only two or three hours. Clearly, the experimental results were the ones to rely on. In addition to correlation, or what Cook and Campbell (1979) called **concomitant variation**, two more conditions are needed for a plausible argument that a causal relationship exists between two variables.

The second condition for causality is **temporal precedence**. The proposed cause should occur before its proposed effect. This can be very difficult to observe in real life, as often events in real life influence each other in both directions. For example, the TV ratings firm Nielsen (2013) announced the results of a study showing that TV programs with more Twitter messages also increased their audience ratings, most likely because the Twitter messages increased interest in viewing the program. But also, the bigger a program's audience, the more Twitter messages there would be about that program, because there would be more tweeters watching it. Disentangling these two-way influences in real-life data requires sophisticated econometric analysis. Experiments get around this problem because in an experiment, the independent variable, the X, is controlled by the experimenter. The experimenter knows when the possible causal influence of X began and ended in time, and can see whether the outcome variable, the Y, moved before or after X was present.

The third and most important of all the conditions for establishing causality is the **absence of any alternative plausible explanations** (APEs). This is the heart of the scientific method. A theory can't be proven, but in the absence of any alternative plausible explanations, it can be generally accepted. An APE that makes it hard to make a causal argument is called a threat to the experiment's internal validity. Different APEs threaten the experiment's external validity, that is, its claims that an observed causal relationship also occurs outside the specific conditions of the experiment. The aim of good experimental design is to eliminate or control for APEs. A key problem when experimenting with people is that the experience of being observed will put participants on guard, so they behave differently in the experiment than they do normally. This provides a threat to external validity. Participants will also try to guess the point of the experiment and may do their best to help it achieve a favorable outcome, threatening internal validity. Research assistants may also be overly helpful and guide participants to do what is expected of them, or "see" what they expect rather than what actually happened (Doyen, Klein, Pichon, & Cleeremans, 2012). This is the reason why drugs trials go to extensive lengths to blind patients and doctors from knowing whether the patient received the new drug or a placebo (the "double blind" experiment method). After discussing the two main types of experimental design, we will return to discussing threats to validity and their remedies.

Between-Groups and Within-Participants Designs

There are two basic choices when it comes to experimental design: (1) between groups, or (2) within participants. In a *between-groups experiment*, there is usually a control group, or comparison group, and at least one experimental or treatment group. The classic example is the randomized double-blind drug trial. Participants are randomly assigned to one of two groups. Both groups are given identical pills, but for the control group the pill is a placebo, while for the experimental group the pill is the new drug being tested. The researchers later measure a key outcome variable in both groups, such as "severity of symptoms" for a treatment, or more chillingly, "number of deaths" for a cure. The results from the two groups would be compared, using a statistical test. A statistical difference between the two groups would be evidence that administering the treatment caused a reduction in symptoms or mortality.

In a *within-participants experiment*, the same participant provides evidence about the control condition and the treatment condition. For example, a drug that treats symptoms could be tested by giving the same person a course of pills over several weeks, and randomly swapping the pills so that in some weeks they are placebo pills and other weeks they are treatment pills. (Obviously, mortality test experiments are impossible using a within-participants design.) Again, the researchers would compare the outcome measure from control weeks with the same measure from treatment weeks. If there is a statistical difference between these two types of weeks, then again there would be evidence that administering the drug reduces the symptoms of the disease.

The main reason for choosing a within-participants design is that fewer participants are needed before there are enough to carry out statistical tests with sufficient power to detect the hypothesized effect, if it exists. This is because in a within-participants experiment, each participant is a perfectly matched control: there are no differences between the control and treatment "groups" because they are the same person. At the extreme, within-participants experiments have been published that have had just one participant (e.g., Krugman, 1971). In a between-groups design, differences between individuals in these groups add noise that can be counteracted only by a large sample size.

A well-designed experiment minimizes the risks of making either a Type 1 or a Type II statistical error. A Type I error occurs when there really is no difference between the control and the treatment conditions (the null hypothesis is true), but by random chance the experimental data reveal a false-positive result suggesting there is a difference. Researchers reduce this risk by choosing an acceptably low level for the probability of making a Type I error, and this probability has the symbol α (alpha). In communication research, α is usually set at $p < .05$, which means that a false-positive result is likely in less than 5 in 100, or 1 in 20 experiments (Fisher, 1935). A Type II error occurs when there really is a difference, but the experiment does not have sufficient data points (power) to statistically reject the null hypothesis. It's important to choose a sampling design that minimizes the probability of a Type II error, symbolized by β (beta), as experiments with false-negative results waste researchers' time, and more importantly waste participants' time and scarce resources like funding. The power of a test is the probability of *not* making a Type II error (i.e., power = 1 – β). Traditionally, the level of β has been set much higher than α, typically $p < .20$ (i.e., power = 1 – .20 = 80%), as increasing participants and therefore power is costly (Cohen, 1988). A useful tool for experimental designers is G*Power, free software that calculates required sample size, given estimates of α, β, and the size of the hypothesized effect (Faul, Erdfelder, Lang, & Buchner, 2007). For $\alpha = .05$ and $\beta = .2$ (80% power), and differences ranging from small (a just-noticeable difference) to large in effect size, here is a table of the required sample sizes for a between-groups experiment and a within-participants experiment, both with two conditions, control versus treatment. In the within-participants design, each participant receives both conditions.

Effect Size	**Between-Groups**	**Within-Participants**
Small (*f* = 0.10)	*N* = 788 (394 per group)	*N* = 199
Medium (*f* = 0.25)	*N* = 128 (64 per group)	*N* = 34
Large (*f* = .40)	*N* = 52 (26 per group)	*N* = 15

On average, a between-groups experiment requires nearly four times as many participants as a within-participants experiment testing the same two conditions. So the question arises, why would anyone choose to carry out a between-groups experiment when within-participants experiments are so much more efficient in terms of participants, and therefore time and resources? The next section introduces some reasons why researchers are forced to use between-participants designs to obtain valid results.

When a Between-Groups Design is More Appropriate

In a within-participants design, the same person experiences multiple conditions. Outcome measures from these conditions should be valid providing none of the conditions interfere with each another. So the key question for communication researchers when choosing between a between-groups or a within-participants design is whether the conditions will interfere with each other if they are administered to the same person. More realistically, carry-over effects from one condition to the next are inescapable using a within-participants design. The question is whether it is possible to control for these interference effects by experimental design, or by using continuous measurement for statistical control (Smith, 2000). Cross-condition interference can have two sources: (1) bottom-up, or stimulus-driven, or (2) top-down, or participant-driven.

Stimulus-driven contamination is more likely as the intensity of the experience increases. An extreme example of a high-intensity experience is “binge viewing,” which is watching two or more episodes of a TV series in a single viewing session. When our lab designed an experiment to test the effects of binge viewing we decided not to make our participants sit through one 3-hour binge session and then do another 3-hour control session (watching 3 different shows) immediately afterwards. Instead we used a between-groups design and each group just watched one 3-hour session.

In a less extreme example, we based our decision to use a mixed within-between design on evidence from prior research about how quickly our manipulations would wear off (Bellman, Wooley, & Varan, 2016). We were manipulating the excitation-transfer effect of TV programs onto ads in the breaks. Research shows that this excitation-transfer effect lasts for over four minutes, which is longer than a typical ad break, but shorter than the time between ad breaks (Mattes & Cantor, 1982; Wang & Lang, 2012). We used a between-groups design to test the effects of different programs, rather than showing the same participant a mix of short program segments, because the excitation-transfer effect of one segment might have carried over across the ad break to interfere with the next segment. But since the excitation-transfer effect wears off between ad breaks, we could show each participant multiple ad breaks during each program, using a within-participants design to vary the ads in the breaks.

Participant-driven interference is a major problem for any experimenter using human participants, as humans are active thinkers (Campbell, 1988). Just like scientists themselves, human participants, and research assistants, propose and test their own hypotheses about what the experiment is about (Smith, 2000). The problem is even worse when the participants or research assistants are students, familiar with the theories the experiment is testing. While top-down interference is a problem for between-groups designs as well, the fact that within-participants designs show participants two or more conditions increases the chances that participants will see a relationship between the conditions and come up with a hypothesis for what the experiment is trying to achieve, even if it is the wrong hypothesis. This was another reason for using a between-groups design to test the effects of different programs in our excitation-transfer study (Bellman et al., 2016). Although a lab-environment signals the fact that some kind of experiment is being conducted, if the experimental procedure seems natural, then it will be harder to guess the hypothesis (Campbell, 1988). People are used to seeing a single program with different ads in ad breaks, so they would have had difficulty guessing the hypotheses that the program and ads were testing. On the other hand, a within-participants design would have been perfectly alright if we had been testing the excitation-transfer effects of magazine articles, as people are used to reading multiple articles in print magazines (Smith, 2000). A later section of this chapter describes how to measure the hypotheses that participants generate about an experiment. Normal practice is to delete any participants who guess the hypotheses correctly. More generally, a pilot test should be used to test whether more than one or two participants will guess the hypotheses, and so a between-groups design should be used.

The Importance of Randomization

When designing an experiment it is useful to remember that the whole purpose of the experiment is to obtain data that satisfy the assumptions of your statistical test. The most common statistical test, analysis of variance (ANOVA), assumes that variation in the observed data come from just two sources: (1) the manipulated differences between the conditions, and (2) normally distributed random error. The smaller the amount of random error, the easier it is to detect a significant difference between any of the conditions.

To make sure that these key assumptions are true—that apart from the manipulated conditions, all variance is random—it is important to control the manipulations so that each participant gets the same manipulation, and to randomize all other sources of variance.

There is no such thing as a “pure” manipulation (Campbell, 1988). Besides the intended manipulations, two conditions will differ in many different ways, such as the time of day, the lighting conditions, the temperature in the room, what happened to participant the day before, etc. All these un-manipulated differences provide alternative explanations for any observed differences and also increase the amount of random error in the differences between conditions.

Applying controls can help to minimize these sources of noise. For example, the interaction between research assistants and participants should be carefully scripted and rehearsed so that each participant gets the same explanation for what the experiment is about, and is equally relaxed and ready to experience the experimental conditions. Artificial lighting and temperature control should be used to eliminate differences on these variables. Other differences, such as time of day, can be controlled by sampling evenly across hours of the day.

In a between-groups design, the groups are expected to differ only according to the group manipulation and by random error. Random assignment of participants to groups ensures that each group differs only by random error, prior to experiencing the group manipulation. The many possible differences between participants in the different groups, such as the time of day they participated in their lab session, should cancel each other out over the course of the study, and look like random error. There are various ways of randomly assigning participants. You can use the rand() function in Excel, or toss a coin, or roll dice. Nevertheless, it is important to test later whether the random assignment procedure was successful. Measure some key variables, such as demographics, and any individual differences likely to affect performance on the dependent variable (e.g., familiarity, experience, mood), and test whether these differ across groups (Smith, 2000). Ideally, these are measured in a pretest, before the experiment, as the experiment could have an effect on these related variables. If there is a difference between supposedly random groups, for example in baseline skin conductance, then this un-manipulated group-difference is an APE, and should be controlled for as a covariate in an analysis of covariance (ANCOVA).

There are differences in opinion on whether ANCOVA should be used even when there are no differences between groups, but solely to reduce the amount of random error in the dependent variable. For example, skin conductance declines with age because older peoples' sweat glands produce less sweat (Neiss, Leigland, Carlson, & Janowsky, 2009), so controlling for age will reduce the error in measures of skin conductance from samples that vary widely in age. Taken to the limit, an experimenter might control for every measured variable that correlates with the dependent variable, and argue that the results are very convincing because all APEs have been controlled for (West, Cham, & Liu, 2014). But ANCOVA uses regression to control variance, and others are wary of the illusory results of regression analyses (Armstrong, 2012; Campbell, 1988). Experiments are used to find causal evidence untainted by regression, so researchers should try, using a well-designed and implemented experiment, to show that the results hold without using ANCOVA.

In a within-participants design, the participants themselves change very little over the course of the experiment, so usually there is no need to control for individual differences such as age. But within-participants designs raise another threat to the assumption of random error between conditions. Purely random error data points are independent, that is, they are not influenced by any other random data point. But when participants experience two or more conditions, there are potential order effects, not only on the conditions themselves, but on the random error associated with each condition. For example, an exciting first condition is likely to have an excitation-transfer effect on the second condition a participant experiences. One solution is to provide a rest period between conditions that allows participants to return to normal prior to each condition. But that would be unnatural in a study of the effects of television viewing. A more general solution is to counterbalance or randomize the order of conditions, so that these order effects are canceled out. A variety of statistical techniques can also be used to control for order effects on supposedly random error variance (West et al., 2014).

Counterbalancing means evenly rotating stimuli and participants so that you end up having equal numbers of observations from all possible orders-of-presentation. This is easy if you have just two conditions. For example, half the participants would experience the first presentation order (e.g., AB), and the other half would experience the reverse order (i.e., BA). Because presentation order in this case has only a small number of levels, it can be treated as another experimental manipulation. If the order manipulation has significant main or interaction effects with any other manipulation, it should be maintained in the design and reported in the results. The total number of presentation-order permutations can be calculated using the PERMUT function in Excel. When the number of permutations gets too large for adequate group sizes (e.g., 4 conditions means 24 permutation groups) then order should be treated as-if random or actually randomized.

Randomizing presentation order is easy when lab software is used. Alternatively, you can write your own program to randomize presentation order on the fly, or randomly assign participants to a large set of presentation-orders, randomly chosen from the many possible permutations. Random orderings will vary the cross-contamination effects, but they may not adequately control for serial-position effects, that is, whether the condition is experienced first, second, or third, etc. Across the randomly selected permutations, the same condition may appear in the same serial position several times. When serial position does not vary across conditions, then serial position is an un-manipulated difference between conditions that provides an APE for any difference between two conditions. Latin square designs rotate conditions across serial position to provide very strong control over serial position effects, and therefore eliminate serial position as a potential APE. But Latin square designs provide very poor control over transfer effects, as the consecutive ordering of conditions does not vary. For example, if you had three within-participant conditions, designated by the letters A, B, and C, then the Latin square design would be:

ABC
CAB
BCA

The first row is the presentation order seen by the first presentation-order group, the second row is the second presentation-order group's ordering, and so on. As you can see, each letter appears in each serial position only once. But in this design, B follows A in every presentation-order group except the last one, in which A appears at the end of the row and B appears at the beginning. The last row weakens the potential transfer effect of A on B, but not enough to eliminate it as an APE.

Graeco-Latin squares layer one Latin square over another to rotate presentation order for two variables simultaneously. This can be handy to ensure that each participant sees each combination of variable-levels just once, which limits hypothesis guessing. Continuing the above example, if the letters A, B, and C indicated low, medium, and high levels of one variable (e.g., ad-induced arousal), then the numbers 1, 2, and 3 could be used to indicate low, medium, and high levels of another variable (e.g., number of cuts between shots). The following Graeco-Latin square could be used:

A_3 B_2 C_1
C_2 A_1 B_3
B_1 C_3 A_2

As you can see, the Graeco-Latin square design rotates the two variables in opposite directions, so that each letter (arousal-level) is combined with each number (cutting-level) just once. However, as well as providing poor control over transfer effects, the Graeco-Latin square provides no control over serial position effects. Each letter-number combination appears in just one serial position in the experiment, which may provide a new APE for any observed effects. If it's at all possible, it's best to randomize presentation order for all the combinations of within-participants conditions.

Threats to Validity in Within-Participant Designs

The purpose of designing an experiment, as opposed to using any other form of research design, is to obtain statistical evidence for a "causal" effect. While we never actually "prove" causation, if we design an experiment carefully, we make it very hard for others to come up with APEs for our results. Cook and Campbell (1979) very usefully categorized the main species of these APEs, in the context of arguing whether it is possible for research made outside the lab (e.g., quasi-experiments) to also rule out these APEs and provide evidence for causation. Using a lab makes it easier to control your manipulations and minimize random variance, but it is still important to know the basic APEs and what you need to do to control for them.

APEs fall under three main headings: (1) *Statistical conclusion validity* threats compromise the ability to provide statistically significant evidence of causality, (2) *Internal validity* threats weaken the argument for causation, and (3) *External validity* threats weaken the argument that the experiment's findings generalize to real life. The first two types of APE, statistical conclusion validity and internal validity, matter more for experimental research, as the whole point of an experiment is to demonstrate a causal relationship between two variables. In physics, this often happens just once, inside a particle collider (Smith, 2000). For theory and practice, a plausible causal finding matters more than any number of correlations (Armstrong, 2012; Cook and Campbell, 1979). But experiments typically occur at the middle of the research cycle. At the beginning of the cycle, when theory is being tentatively proposed via exploratory case studies, or at the end of the cycle, when theory is being applied, it is more important to consider whether the theory might generalize across situations. An experiment, showing a causal relationship, ends one phase of exploration, and begins another, especially when the proposed causal explanation is not complete (Smith, 2000). All three types of validity threat apply to within-participants designs.

Statistical Conclusion Validity Threats. As was discussed above, there is no such thing as a "pure" manipulation (Campbell, 1988). The only thing that experimenters can know for sure is that they have found a statistically significant difference between conditions. Exactly what "caused" those differences is a matter of argument. But first, in order to find a statistically significant difference between conditions, the experimental design needs to minimize the amount of random error across conditions.

One easy way to reduce random error is to take advantage of the law of large numbers and make a large number of observations. Obviously, an experiment needs to sample an adequate number of participants in order to have the power to detect the hypothesized effect (as discussed above, you can determine sample size using G*Power or other such programs). But experimenters can also take advantage of the law of large numbers by showing each participant a large number of stimuli. Each participant will have a mean level of random error in their responses, and the confidence interval surrounding this mean will shrink as the number of observations increases. So if you are showing participants multiple stimuli, try to show as many as possible within the time you can reasonably expect someone to devote to participating in an experiment with equal ability, effort, and alertness, so that the mean error rate won't change over time. In a lab, that might be up to an hour. Online, it might be for only five minutes.

Sometimes, it isn't possible to put participants through more than two within-participants conditions. For example, novel experimental content will not be novel when it is shown a second time. Similarly, if measuring the effect reveals the purpose of the experiment, you can't put the participant back into an unaware state again. If you have only a couple of conditions, then it's especially important to have a dependent variable measure that is as free of random error as possible. In other words, it is important to use a highly reliable (free of random error) measure of each dependent variable.

Usually, the most reliable measures are the ones with the highest coefficient alphas in the published literature. There is not much a covariate can do to reduce the random error variance in a variable with an alpha of .95. But make sure you are not trading off reliability for construct validity. A measure can have high reliability (low random variance) but measure the wrong thing. Bergkvist and Rossiter (2007) validated many useful single-item measures of communication outcome variables, which work as well as validated multiple-item measures of the same constructs. Multiple measures sap participants' patience, so they can have lower construct validity as their reliability increases, because of participant-response biases (e.g., picking the middle response for every item). Podsakoff and colleagues (2012) describe ways of testing for and correcting for response biases when multiple items are used (e.g., in a very long survey).

The concept of using reliable measures to minimize random error and increase statistical conclusion validity also applies to the experimental manipulations. It is no use using highly reliable measures if the experimental conditions are so different across participants they are practically random noise. As much as possible, researchers should ensure that each condition is experienced identically by standardizing procedures, stimuli, and test environments. Statistical conclusion validity also requires that the assumptions of the statistical test used are satisfied. Above, we discussed the importance of randomizing order of presentation to satisfy the assumption that errors are unrelated across conditions. However, a variety of statistical techniques can be used (e.g., repeated measures ANOVA) when this assumption cannot be maintained (Smith, 2000).

Internal Validity Threats. Internal validity threats are the most important for within-participants designs, and these also fall under three main headings: (1) Over Time Threats, (2) Testing Threats, and (3) Sampling Threats.

Over Time Threats are sources of change over time that were not manipulated by the experiment. Maturation threats occur within the participant. For example, relaxation while watching a TV-program provides an APE for low measured arousal at the end of the program. History threats happen outside the participant. For example, a crash or explosion outside the lab provides an APE for high measured arousal at the end of the program. Over-time threats can be eliminated by randomizing or counterbalancing presentation order, and sampling over different days and times.

Testing Threats occur when the process of measuring the dependent variable affects the results of the experiment. Testing threats often affect social experiments outside the lab (Campbell, 1988), when the introduction of a new policy (e.g., more police) is accompanied by new ways of measuring results (e.g., crime statistics). But researchers also need to control for potential instrumentation effects inside labs. For example, skin conductance is affected by the temperature in the room, so it is important to make sure that lab temperature is constant across each participant's session. Similarly, differences in reaction time measures may be the result of participants getting faster at reacting rather than differences between conditions. Differences between measures can also be due to using different observers, or the same observer getting better or worse over time. Randomizing or counterbalancing presentation order safeguards against over-time testing threats. Another safeguard is to use multiple measures: if one is compromised, another may not be. You may be fortunate to have equivalent tests available for the dependent variable so that you can use different items for a pretest and a posttest. In practice, however, equivalent items are easier to find for math tests than for measures of communication effects.

Randomization and using multiple measures cannot control another type of testing threat, which occurs when measures interfere with each other. For example, pretesting for interest in buying the brand can clearly indicate that the main purpose of the experiment is to test for an increase in this measure. Participants might then try to be helpful, and respond too well, or react against this purpose and respond in the opposite direction, reducing their stated interest in buying the brand. If you need to use a pretest measure, make sure you avoid these hypothesis-guessing effects by disguising the key pretest measure in a maze of distracter items. For example, ask participants about their buying intentions for a number of brands, not just the test brands. The same logic applies to the order of measures in a posttest, when using a posttest-only design. The first measure may signal the hypothesis, and therefore interfere with all the subsequent measures. To combat this testing threat, it's best to randomize the order of questions in the posttest (easy using Qualtrics or other software), or ask the key dependent variable first, ensuring it is the least affected by hypothesis guessing (Rossiter & Percy, 1997).

In one of our experiments, we used a pretest-posttest design to measure interest in products online, to see whether this information could be used to target TV commercials (Bellman, Murphy, Treleaven-Hassard, O'Farrell, Qiu, & Varan, 2013). The pretest and the posttest were conducted in the same lab session, which was disguised as two separate experiments. Our participants normally did two experiments in one session, so this was not an unnatural experience, but to prevent test interference we ran the pretest and the posttest in different rooms, with different research assistants, and different colored paperwork. To further avoid any interaction between the pretest and the posttest, we used a very long filler task to separate the two phases of the experiment. This filler task was a 20-minute survey about the pretest Web site, which participants were told was a new "consumer reports" site, featuring content about different product categories. The 20-minute filler task also introduced a realistic delay between measuring product interest online and downloading and inserting targeted commercials via a digital TV set-top box. We measured (using log files, not survey questions) which products a participant looked at the most during the pretest and in the posttest showed TV commercials for these products during a TV program. Participants could skip commercials if they wanted to, using a remote control, and whether or not an ad was skipped was our dependent variable (again, not a survey question). We found that participants were less likely to skip commercials for products they had shown an interest in during the pretest, compared to commercials for products they had not shown an interest in (at least, for low-involvement products). Finally, to make sure that the two phases did not interfere with each other, we asked participants a staged series of debriefing questions, designed to elicit any successful hypothesis-guessing that would invalidate a participant's results (Aronson, Wilson, & Brewer, 1998). Fortunately, none of our participants guessed that the two "experiments" were related to each other.

Sampling threats describe APEs related to differences between conditions that might be due to characteristics of the participants used in the experiment. Sometimes participants are not randomly assigned to conditions but assigned because of their scores on a pretest. In other research into the effectiveness of targeted TV commercials, we used pretests to measure participants' stated interest in different product categories (Bellman & Varan, 2014). To minimize test interference, the pretest was conducted online, more than a month before the person turned up at the lab for their experimental session. But what happened in these experiments can be explained by a common APE that affects samples selected according to their pretest scores, rather than by random assignment. The participants who said they were highly interested in a product in the pretest, were often so interested that over the month between the pretest and their lab session they went out and bought the item. This meant that by the time of the experiment they were only as interested in the product as the average person (they had "regressed to the mean"). When participants were shown ads for products they had been very interested in a month ago, they were just as likely to skip these targeted ads as they were to skip untargeted ads. More generally, people fluctuate above and below their average score on any test, so people chosen because they are high or low on a pretest are likely to score closer to their mean on the posttest, and this could artificially give the impression that the experiment has moved them up or down compared to their pretest score.

Even when a random sample of participants is used, participant drop-out (mortality) can result in a non-random sample by the end of the experiment. Some random mortality is to be expected, but sometimes drop-out is systematic and provides an alternative explanation for its results. For example, in an eye-tracking study, people who wear glasses are more likely to lose calibration and drop out of the study. This would mean that the results of the experiment would be less likely to apply to them. Worse, the significant results might be entirely due to the characteristics of the people who completed the experiment (i.e., those who did not wear glasses), rather than the experimental manipulations. Mortality threats are just like random assignment failure, and so similarly it helps to collect information on several variables (e.g., demographics) to compare stayers with dropouts (Smith, 2000). Differences between these two types of participants can be controlled for using ANCOVA.

External Validity. The third and final family of APEs for within-participants designs are threats to external validity. The whole purpose of an experiment is to cleanly observe a significant causal relationship between two variables, X and Y, which exists in the real world. External validity threats argue that this significant causal relationship is somehow unique to the experiment, and does not apply generally. External validity APEs can relate to (1) *Who* took part in the experiment, (2) *What* they did, (3) *Where* they did it, or (4) *When* they did it.

An example of a credible *who threat* is the use of student samples in experiments. Students are very convenient for academics to recruit, but they tend to be younger than the general population. Young people typically have better memory (DuBow, 1995), a greater need for stimulation (Zuckerman, 1979), and larger emotional responses (Droulers, Lacoste-Badie, & Malek, 2015; Neiss et al., 2009). For this reason, causal findings based on student samples may not translate to the general population. A study by Nelson, Meyvis, and Galak (2009) illustrates the problem and its solution. Their first experiment, using a student sample, found that TV commercials make programs more enjoyable. But in a second experiment reported in the same article, they tested the external validity of this finding using older viewers. They found that, generally, and more intuitively, people don't like commercial interruptions. These days, with easy access to non-student samples via Amazon's Mechanical Turk and Qualtrics, it is practically inexcusable not to verify the external validity of student-sample findings.

"Who" threats can be countered by wider sampling and replication, and the same remedy applies to all the other threats to external validity. For example, using just one stimulus (e.g., one ad) to manipulate a condition raises a "what" threat. The results of the experiment might apply only to that specific stimulus. A general recommendation is to use at least three stimuli to represent any manipulation (Potter & Bolls, 2012). "Where" threats classically apply to experiments conducted in a lab. People know that labs are where experiments occur, and behave differently compared to how they behave normally (Campbell, 1988). Again, APEs relating to "where" threats can be countered by repeating the experiment in various locations, including outside the lab, for example, online (Johnson, 2001). "When" threats relate to the timing of the experiment. For example, researchers can find different results if they observe reactions to a live presidential debate, as opposed to a recorded debate shown after the election result is known (Wicks, 2007).

Threats to Validity in Between-Groups Designs

Earlier, I discussed reasons for choosing to use a between-groups design rather than a within-participants design. These reasons related to the intensity and carry-over effects of the manipulated conditions. Another reason for using a between-groups design is when it is not possible to control for all internal-validity threats using a within-participants design. For example, it might be suspected that a within-participants experiment is threatened by a testing-threat related to the use of a pretest. The best way to rule out pretest contamination as an APE is to use a between-groups design in which neither group receives the pretest.

To fully investigate whether pretest contamination is affecting results, it is necessary to use a Solomon 4-Group design (Cook & Campbell, 1979). This design has two experimental groups, and two control groups. One experimental group receives the pretest and the posttest, and so does one of the control groups. The other two groups are posttest-only groups. To rule out APEs relating the “what” and “where” of the experiment, it’s best to make all four groups highly comparable by making the control groups carry out a similar task (e.g., in a lab). The same reasoning applies to drugs trials where the placebo group takes an identical-looking sugar pill, to rule out the placebo effect as the explanation for the drug’s success. If the pretest-posttest control group and the posttest-only control group differ on their posttest results, this would be solid evidence of pretest-contamination. Other over-time threats can also be investigated by comparing the pretest scores from the pretest-posttest control group with the posttest from the posttest-only control group. Because of the expense of collecting data from four groups, most experimenters choose to avoid the problems that pretests can introduce and use a posttest-only design (Smith, 2000).

Using a between-groups design can help to rule out internal-validity threats to within-participants designs, but introduces another set of APEs unique to between-groups designs. Cook and Campbell (1979) identified three key threats to between-groups designs: (1) Selection, (2) Interactions between Selection and Within-Participants Threats, and (3) Social Threats.

Selection. Typically, experimenters use random assignment to allocate participants to the experimental or the control group. This ensures that both groups were equal before assignment. ANCOVA can be used to remedy any failures to randomize that might arise on measured variables. But outside the lab, when researchers analyze “natural” experiments, a selection-bias threat argues that the significant causal relationship is due to different types of people receiving the treatment and the control conditions. For example, the benefits of attending an expensive university are obscured by the fact that the people who attend these universities come from very wealthy social networks, which would probably have ensured success without a university education (Campbell, 1988).

Selection bias used to be a rare phenomenon in the lab. But above, I gave an example of regression-to-the-mean using the results of a lab study in which participants were assigned to conditions based on a pretest. And increasingly communication researchers are investigating the effects of interactivity, that is, the effect of choosing to interact or not. This choice gives the participant control over their own assignment, to either the treatment (interacted) or control (didn’t interact) conditions. When participants choose their own assignment, the experiment no longer features random assignment and it can no longer be argued that the treatment and control groups were definitely equal prior to assignment (Aronson et al., 1998). The reasons for interacting or not interacting, whatever they are, provide many APEs for the apparent effects of interacting versus not interacting.

One solution to this problem is to provide “choices” that no participant refuses (Aronson et al., 1998), but this can produce a very unnatural experience. A better solution to the problem of non-random assignment (e.g., to interaction vs. non-interaction) is to use econometric regression techniques or sample-matching procedures (Bellman & Varan, 2012; West et al., 2014). These methods were developed by natural-experiment researchers estimating, for example, the real value of a university education. Estimating each person’s propensity to choose the treatment (e.g., education, interaction) allows this propensity to be controlled for when estimating the effect of the treatment. For example, a TV viewer might send a tweet about a commercial for a variety of reasons, besides the creativity of the commercial (e.g., they may be heavy users of Twitter). In a study of "social TV", that is, sending messages about the TV content you're watching, we manipulated ad creativity, but our participants were free to self-select whether or not they sent messages about the ads (Bellman, Robinson, Wooley, & Varan, 2017). After controlling for self-selection using regression, we found that ad creativity significantly increased ad-related messaging.

Interactions with Within-Participants Threats. Even when participants are randomly allocated to groups, it can happen that each group experiences a different set of within-participants threats to internal validity. Again, this is more likely outside the lab, such as when different schools are randomly assigned to the control and treatment groups, and the experience of the experimental school differs from that of the control school, leading to differences in history, maturation, or testing. Inside the lab, it is best to ensure an equally wide-ranging sample of background conditions for each group. For example, the treatment and control groups should be alternated or run simultaneously in different rooms to avoid APEs relating to being tested at different times, or days of the week, or during certain external events.

Social Threats. The main disadvantage of using different groups is that the groups can hear about the different manipulation received by the other group, or communicate directly with each other about their manipulations. These social processes can lead to leakage of manipulations across group boundaries. For example, the control group might hear about the intervention being trialed by the treatment group (e.g., exercise more), and try out this intervention for themselves. The results would very likely show no difference between the treatment and the control groups, using a posttest-only design. The danger of social threats is one argument for using a pretest-posttest or even a Solomon 4-Group design.

Compared to using a pretest-posttest design, other solutions to social threats only tend to make matters worse. For example, if the control group hears that the treatment group are getting an intervention aimed at improving performance, they might engage in "compensatory rivalry," working harder than normal to make sure they don't look bad. Again, this might lead to no difference between the treatment and control groups, or the control group performing better. Alternatively, the control group might suffer "resentful demoralization" on hearing they have been deprived of the performance-enhancing treatment, and perform worse than normal, making the treatment appear effective when in fact it is not. To avoid compensatory rivalry or resentful demoralization, an experimenter might decide to use "compensatory equalization"; giving benefits to the control group to encourage them to behave normally. But rewarding the control group might also encourage them to work harder, and erase the difference between the treatment and the control.

Another reason that interactivity researchers should use regression to control for selection-threats is that social-threats are likely when interactivity is manipulated by forcing participants to interact (Bellman & Varan, 2012). People in the control condition might resent being shown the less-exciting content, or compensate for this by working harder.

A Brief Note about Ethics

Prior to the Helsinki Convention of 1964, experimental research on humans was associated with some of the most appalling abuses of human rights. Since that time, experimentation on humans has required ethics approval before the research can commence. Furthermore, journals will not publish research that has been conducted without ethical clearance.

The key document for researchers and participants in experimental research is the information sheet and consent form that allows people to give informed consent to participating in the experiment. This document should be short and clear to ensure that it is read. But difficulties arise when the information given to the participants potentially harms the value of the research. For example, social threats to validity can arise when the information sheet describes the treatment condition in more glowing terms than the control condition. Worse, the information sheet may convey the hypothesis being tested by the study, potentially invalidating the results of the experiment. The best remedy is to improve the design of the experiment so that the control and treatment conditions are practically identical, as in a placebo trial. For example, in our excitation-transfer study, we could legitimately describe the experiment as a study of people's responses to TV content from other countries, rather than give away our hypotheses about excitation-transfer differences across programs. Every group saw a program with ads from another English-speaking country, to increase unfamiliarity and attention to the content (Campbell & Keller, 2003).

Ethics Review Boards will allow researchers to use active deception, in the information sheet and other information provided to participants, if it can be argued that the benefits of the research outweigh the costs of consenting without full information. Communication research usually involves very minor harm to the person, so that the hidden aspects of what the participant consents to are usually no worse than the known aspects (e.g., no illegal behavior is required). In other words, the person would still consent to participation after being told what the experiment was really about. Ethical permission to use deception usually insists that participants are fully debriefed afterwards and told what the true purpose of the study was. A debriefing procedure can be useful for obtaining evidence about hypothesis guessing, as in our pretest-posttest study of Internet-targeted TV ads (Bellman et al., 2013). But others argue that if the harm is minor, debriefing can only hurt, because all it does is hurt the participant's pride (they are revealed to be a dupe) (Campbell, 1988). Using that argument, you may be able to convince your Ethics Board to allow a waiver of the need to debrief after deception. If you need to use debriefing, make sure that participants do not talk with other potential participants about what the true nature of the experiment is. This can be easier when using non-student samples than student samples.

New Developments in Experimental Design

I will conclude this entry with a brief presentation of some of the new developments that have made recent years an exciting time for researchers using experimental design.

Online Experiments. The Internet has made it easier for experimenters to control for threats to internal and external validity. In the lab, instead of paper questionnaires, sites like Qualtrics can be used to deliver online surveys during pretests and posttests, with randomized question order to avoid measurement interference. For example, in one of our studies, a Qualtrics survey was used to randomize order of presentation of the within-participants conditions (Bellman, Potter, Treleaven-Hassard, Robinson, & Varan, 2011). The most useful aspect of the Internet for experimental researchers, however, is the ability it offers for gathering non-student samples that are more representative of the general population, using, for example, Qualtrics or Amazon's MTurk. Online surveys can randomly assign participants to treatment and control conditions. For example, Goldfarb and Tucker (2011) used data from online experiments to show that targeting online banner ads increases their effectiveness by 65%. These experiments used pop-up surveys to quiz online consumers about their interest in buying a target brand hidden in a list of brands for various products. Half of the sample for each target brand had been randomly assigned to previously see either an ad for the brand, or a placebo ad for a non-profit. This design provides strong internally valid evidence for a causal effect of ad exposure on purchase intention, coupled with high external validity, because the data were obtained in the field, not in the lab.

Biometric and Neuropsychological Measures. Another exciting development in experimental research is the increasing affordability of biometrics and other mechanical measures that do not depend on participants' self-reports. Because these measures tap automatic, uncontrollable responses, it is impossible for the pretest (e.g., a baseline) to interfere with the posttest measure by affecting how participants think about the experiment. Results based on these measures provide strong evidence against APEs based on measurement interference effects. More importantly, they allow researchers to observe changes in multiple variables continuously through the content used in an experiment, something impossible to achieve using self-reports (Potter & Bolls, 2012). If conditions have carry-over effects on subsequent conditions, these effects can be unobtrusively observed, and if necessary, controlled for. For example, in our excitation-transfer study, we used computerized facial-expression coding to measure positive emotion (smiling) transferring from funny programs to funny and serious ads in the ad breaks (Bellman et al., 2016). Our results showed that positive emotion did transfer from the funny program to increase smiling during serious ads. But this program effect had disappeared by the onset of the second ad in the break. Funny programs also tended to be make funny ads slightly funnier, but the difference was not significant.

Field Experiments using Biometrics. A further development in experimental research is the combination of field experiments, using non-student samples outside the lab, with biometric measures that avoid hypothesis-guessing. This combination promises to deliver the most reliable evidence for how variables are causally related in natural behavior. The software we used to measure smiling in our excitation-transfer study (Bellman et al., 2016) was developed online using crowdsourced Web cam videos of people making facial expressions (McDuff, El Kaliouby, & Picard, 2012). Prior to our lab-study, this software was used, via Web cams, to measure smiling by members of the general public in response to TV commercials (Teixeira & Stipp, 2013). The results of this study showed that the best ads use an optimal amount of humor, not too much and not too little.

Although some forms of neuropsychological measurement, such as fMRI machines, are unlikely ever to be portable, validation studies are providing encouraging evidence for the use of portable "neuro" measures outside the lab. For example, the Emotiv headset, developed to measure brainwaves (EEG) to control video games, has been shown to measure processing-related peaks in EEG response (event-related potentials [ERPs]) almost as well as expensive medical-grade equipment (Badcock, Mousikou, Mahajan, de Lissa, Thie, & McArthur, 2013). Such headsets could be worn by people carrying out activities in their daily life, such as visiting a supermarket. An APE for any such study, however, is that people would be unable to forget that their brainwaves are being monitored, and that they look very different from the shoppers around them.

Smartphones, smart watches, and fitness monitors provide a more unobtrusive source of biometrics and neuro-metrics from the field. The Apple smart phone can monitor heart rate, potentially providing continuous measures of orienting (attention) responses to all the stimuli a person encounters during the day (Lang, 1994). Perhaps the most impressive use, so far, of mobile phones for a field experiment is the *Mobile Century* experiment (Herrera, Work, Herring, Ban, Jacobson, & Bayen, 2010). This study showed that if only 2–3% of drivers have mobile phones set up for GPS monitoring, it is possible to both preserve their privacy and provide accurate measurement of the current velocity of traffic on busy roads.

These new developments that make the current era of experimental research very exciting. Carefully designed experiments will continue to be important for isolating variables and providing the best possible evidence for causal relationships between variables. Causal relationships are the building blocks of new developments in communication research theory.

SEE ALSO: Communication Research Methods ; Communication Theory ; Ethics ; Human-Computer Interaction ; Mass Communication Theory ; Media Psychology ; Media Theory ; Models Of Communication ; PR Theory ; Quantitative Methods ; Science Reporting ; Survey Methods ; Television Theory ; Visual and Non-Verbal Communication

References

Armstrong, J. S. (2012). Illusions in regression analysis. *International Journal of Forecasting*, 28 (3), 689-694. doi:10.1016/j.ijforecast.2012.02.001

Badcock, N. A., Mousikou, P., Mahajan, Y., de Lissa, P., Thie, J., & McArthur, G. (2013). Validation of the Emotiv EPOC® EEG gaming system for measuring research quality auditory ERPs. *PeerJ*, 1, e38. doi: 10.7717/peerj.38

Bellman, S., & Varan, D. (2012). Modeling self-selection bias in interactive-communications research. *Communication Methods and Measures*, 6 (3), 163-189. doi: 10.1080/19312458.2012.703833

Bellman, S., & Varan, D. (2014). The dynamic of addressable TV ads: Tracking today's consumer. Presentation at the Advertising Research Foundation's *2014 Audience Measurement* conference, New York, June 8-10.

Bellman, S., Wooley, B., & Varan, D. (2016). Program–ad matching and television ad effectiveness: A reinquiry using facial tracking software. *Journal of Advertising*, 45(1), 72-77. doi: 10.1080/00913367.2015.1085816

Bellman, S., Robinson, J. A., Wooley, B., & Varan, D. (2017). The effects of social TV on television advertising effectiveness. *Journal of Marketing Communications*, 23(1), 73-91. doi: 10.1080/13527266.2014.921637

Bellman, S., Potter, R. F., Treleaven-Hassard, S., Robinson, J. A., & Varan, D. (2011). The effectiveness of branded mobile phone apps. *Journal of Interactive Marketing*, 25 (4), 191-200. doi: 10.1016/j.intmar.2011.06.001

Bellman, S., Murphy, J., Treleaven-Hassard, S., O'Farrell, J., Qiu, L., & Varan, D. (2013). Using Internet behavior to deliver relevant television commercials. *Journal of Interactive Marketing*, 27 (2), 130-140. doi: 10.1016/j.intmar.2012.12.001

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, 44 (2), 175-184. doi: 10.1509/jmkr.44.2.175

Campbell, M. C., & Keller, K. L. (2003). Brand familiarity and advertising repetition effects. *Journal of Consumer Research*, 30 (2), 292-304. doi: 10.1086/376800

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd. ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago, IL: Rand McNally College Publishing.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, 7 (1), e29081. doi:10.1371/journal.pone.0029081

Droulers, O., Lacoste-Badie, S., & Malek, F. (2015). Age-related differences in emotion regulation within the context of sad and happy TV programs. *Psychology & Marketing*, 32 (8), 795-807. doi: 10.1002/mar.20819

DuBow, J. S. (1995). Advertising recognition and recall by age—including teens. *Journal of Advertising Research*, 35 (5), 55-60.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191. doi: 10.3758/bf03193146

Fisher, R. A. (1935). *The Design of Experiments*. London: Oliver and Boyd.

Goldfarb, A., & Tucker, C. E. (2011). Privacy regulation and online advertising. *Management Science*, 57 (1), 57-71. doi: 10.1287/mnsc.1100.1246

Herrera, J. C., Work, D. B., Herring, R., Ban, X., Jacobson, Q., & Bayen, A. M. (2010). Evaluation of traffic data obtained via GPS-enabled mobile phones: The *Mobile Century* field experiment. *Transportation Research Part C: Emerging Technologies*, 18 (4), 568-583. doi: 10.1016/j.trc.2009.10.006.

Johnson, E. J. (2001). Digitizing consumer research. *Journal of Consumer Research*, 28 (2), 331-336. doi: 10.1086/322908

Krugman, H. E. (1971). Brain wave measures of media involvement. *Journal of Advertising Research*, 11 (1), 3-9. doi: 10.4135/9781452231501.n13

Lang, A. (1994). What can the heart tell us about thinking? In A. Lang (Ed.), *Measuring Psychological Responses to Media Messages* (pp. 99-111). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mattes, J., & Cantor, N. (1982). Enhancing responses to television advertisements via the transfer of residual arousal from prior programming. *Journal of Broadcasting*, 26 (2), 553-566. doi: 10.1080/08838158209364024

McDuff, D., El Kaliouby, R., & Picard, R. W. (2012). Crowdsourcing facial responses to online videos. *IEEE Transactions on Affective Computing*, 3 (4), 456-68. doi: 10.1109/t-affc.2012.19

Neiss, M. B., Leigland, L. A., Carlson, N. E., & Janowsky, J. S. (2009). Age differences in perception and awareness of emotion. *Neurobiology of Aging*, 30 (8), 1305-313. doi:10.1016/j.neurobiolaging.2007.11.007

Nelson, L. D., Meyvis, T., & Galak, J. (2009). Enhancing the television-viewing experience through commercial interruptions. *Journal of Consumer Research*, 36 (2), 160-172. doi: 10.1086/597030

Nielsen (2013). New Nielsen research indicates two-way causal influence between Twitter activity and TV viewership. Press Release. Available online at http://www.nielsen.com/us/en/press-room/2013/new-nielsen-research-indicates-two-way-causal-influence-between-.html

Okasha, Samir (2002). *Philosophy of Science: A Very Short Introduction*. Oxford/New York: Oxford University Press. doi: 10.1093/actrade/9780192802835.001.0001

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, 63, 539-569. doi: 10.1146/annurev-psych-120710-100452

Potter, R. F., & Bolls, P. D. (2012). *Psychophysiological measurement and meaning: Cognitive and emotional processing of media*. New York: Routledge. doi: 10.4324/9780203181027

Rossiter, J. R., & Percy, L. (1997). *Advertising communications and promotion management*, 2nd ed. New York: McGraw-Hill.

Smith, E. R. (2000). Research design. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 17-39). New York: Cambridge University Press. doi: 10.1017/cbo9780511996481.006

Teixeira, T., & Stipp, H. (2013). Optimizing the amount of entertainment in advertising: What's so funny about tracking reactions to humor? *Journal of Advertising Research*, 53(3), 286-296. doi: 10.2501/jar-53-3-286-296

Wang, Z., & Lang, A. (2012). Reconceptualizing excitation transfer as motivational activation changes and a test of the television program context effects. *Media Psychology*, 15(1), 68-92. doi: 10.1080/15213269.2011.649604

West, S. G., Cham, H., & Liu, Y. (2014). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology*, 2nd ed. (pp. 49-80). New York: Cambridge University Press. doi: 10.1017/CBO9780511996481.007

Wicks, R. H. (2007). Does presentation style of presidential debate influence young voters' perceptions of candidates? *American Behavioral Scientist*, 50 (9), 1247-1254. doi:10.1177/0002764207300054

Zuckerman, M. (1979). *Sensation seeking: Beyond the optimum level of arousal*. Hillsdale: Erlbaum.

Further Reading.

Aronson, E., Wilson, T. D., & Brewer, M. B. (1998). Experimentation in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 1, pp. 99-142). New York, NY: Oxford University Press.

Reeves, B., & Geiger, S. (1994). Designing experiments that assess psychological responses to media messages. In A. Lang (Ed.), *Measuring Psychological Responses to Media Messages* (pp. 165-180). Hillsdale, NJ: Lawrence Erlbaum Associates.

Campbell, Donald T. (1988). *Methodology and epistemology for social science*. Chicago: University of Chicago Press.

Cochran, W. G., & Cox, G. M. (1992). *Experimental designs*, 2nd ed. Hoboken, NJ: John Wiley & Sons.

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences*, 4th ed. Thousand Oaks, CA: SAGE Publications. doi: 10.4135/9781483384733