DISCUSSION 1 & 2

Prior to beginning work on this discussion, please read the required articles by Skidmore (2008) and Henrich, Heine, & Norenzayan (2010). Carefully review the PSY635 Week Two Discussion Scenario. Apply the scientific method to the information included within the scenario and develop a null and a research hypothesis based on it. Using the hypotheses you have developed, compare the characteristics of the different experimental research designs discussed in the Skidmore (2008) article and choose the one that is most appropriate to adequately test your hypotheses. Identify potential internal threats to validity and explain how you might mitigate these threats. Apply ethical principles to the proposed research and describe the implications of this type of research in terms of the population(s) and cultural consideration(s) represented in the sample(s) within the scenario. Provide a substantive response. Please do not include the question prompt in your initial post.

-attached

PLEASE SEPERATE THE TWO ASSIGNMENTS

PSY635 Week Two Scenario

Three instructors teach the same online course and have devised an experimental intervention to improve student motivation to actively participate in discussions. The course is a core requirement for all psychology students, and students are assigned to particular sections at random rather than by instructor choice.

The average class size for this particular course is 45 students. To get a large enough sample for adequate analysis, the instructors have decided to include two sections for each instructor in the experiment. The first section will serve as the control group (no experimental intervention), and the second section will receive the intervention. Anonymous data about the dependent variable will be pooled for the three sections comprising the control group and the three sections that receive the intervention.

The independent variable is the intervention, which may be an incentive such as digital badges or an instructional intervention involving changing the instructions for the guided response. The dependent variable will be the number of response (not initial) posts per student that exceed two lines of text. The researchers have decided to use the Week Four discussion for data collection, reasoning that it may take some time for the intervention to become effective.

Source 1

Running Head: EXPERIMENTAL DESIGN

Experimental Design and Some Threats to Experimental Validity: A Primer

Susan Skidmore

Texas A&M University

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, Louisiana, February 6, 2008.

Abstract

Experimental designs are distinguished as *the* best method to respond to questions involving causality. The purpose of the present paper is to explicate the logic of experimental design and why it is so vital to questions that demand causal conclusions. In addition, types of internal and external validity threats are discussed. To emphasize the current interest in experimental designs, Evidence-Based Practices (EBP) in medicine, psychology and education are highlighted. Finally, cautionary statements regarding experimental designs are elucidated with examples from the literature.

The *No Child Left Behind Act* (*NCLB*) demands "scientifically based research" as the basis for awarding many grants in education (2001). Specifically, the 107th Congress (2001) delineated scientifically-based research as that which "is evaluated using experimental or quasi-experimental designs". Recognizing the increased interest and demand for scientifically-based research in education policy and practice, the National Research Council released the publication, *Scientific Research in Education* (Shavelson & Towne, 2002) a year after the implementation of *NCLB*. Almost $5 billion have been channeled to programs that provide scientifically-based evidence of effective instruction, such as the *Reading First Program* (U. S. Department of Education, 2007). With multiple methods available to education researchers, why does the U. S. government show partiality to one particular method? The purpose of the present paper is to explicate the logic of experimental design and why it is so vital to questions that demand causal conclusions. In addition, types of internal and external validity threats are discussed. To emphasize the current interest in experimental designs, Evidence-Based Practices (EBP) in medicine, psychology and education are highlighted. Finally, cautionary statements regarding experimental designs are elucidated with examples from the literature.

Experimental Design

An experiment is "that portion of research in which variables are manipulated and their effects upon other variables observed" (Campbell & Stanley, 1963, p. 171). Or stated another way, experiments are concerned with an independent variable (IV) that causes or predicts the outcome of the

dependent variable (DV). Ideally, all other variables are eliminated, controlled or distributed in such a way that a conclusion that the IV caused the DV is validly justified.

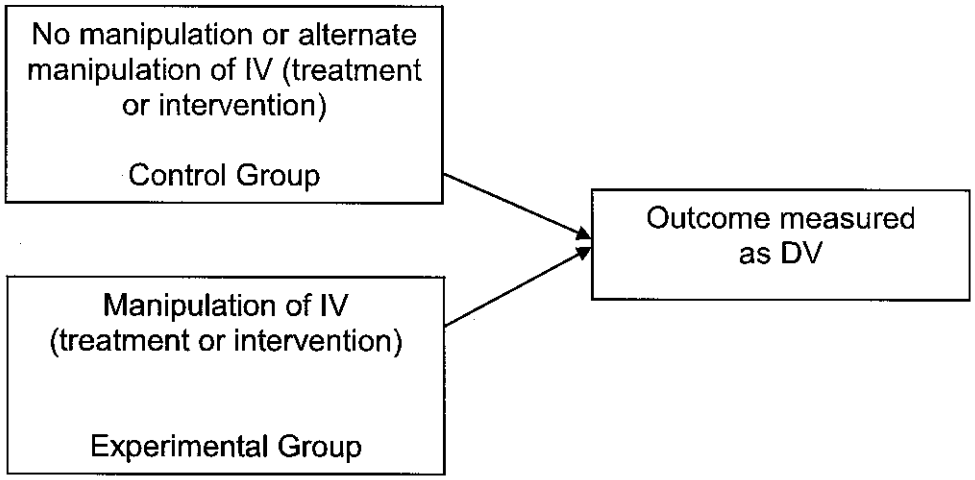


Figure 1. Diagram of an experiment.

In Figure 1 above you can see that there are two groups. One group receives some sort of manipulation that is thought (theoretically or from previous research) to have an impact on the DV. This is known as the experimental group because participants in this group receive some type of treatment that is presumed to impact the DV. The other group, which does not receive a treatment or instead receives some type of alternative treatment, provides the result of what would have happened without experimental intervention (manipulation of the IV).

So how do you determine whether participants will be in the control group or the experimental group? The answer to this question is one of the characteristics that underlie the strength of true experimental designs. True experiments must have three essential characteristics: random assignment to

groups, an intervention given to at least one group and an alternate or no intervention for at least one other group, and a comparison of group performances on some post-intervention measurement (Gall, Gall, & Borg, 2005).

Participants in a true experimental design are randomly allocated to either the control group or the experimental group. A caution is necessary here. Random assignment is not equivalent to random sampling. Random sampling determines who will be in the study, while random assignment determines in which groups participants will be. Random assignment makes "samples randomly similar *to each other*, whereas random sampling makes a sample similar to *a population*" (Shadish, Cook, & Campbell, 2002, p. 248, emphasis in original). Nonetheless, random assignment is extremely important. By randomly assigning participants (or groups of participants) to either the experimental or control group, each participant (or groups of participants) is *as likely* to be assigned to one group as to the other (Gall et al., 2005). In other words, by giving each participant an equal probability of being a member of each group, random assignment equates the groups on all other factors, *except* for the intervention that is being implemented, thereby ensuring that the experiment will produce "unbiased estimates of the average treatment effect" (Rosenbaum, 1995, p. 37). To be clear, the term "unbiased estimates" describes the fact that any observed effect differences between the study results and the "true" population are *due to chance* (Shadish et al., 2002).

This equality of groups assertion is based on the construction of infinite number of random assignments of participants (or groups of participants) to treatment groups in the study and *not* to the single random assignment in the particular study (Shadish et al., 2002). Thankfully, researchers do not have to conduct an infinite number of random assignments in an infinite number of studies for this assumption to hold. The equality of groups' assumption is supported in studies with large sample sizes, but not in studies with very small sample sizes. This is true due to the *law of large numbers*. As Boger (2005) explained, "If larger and larger samples are successively drawn from a population and a running average calculated after each sample has been drawn, the sequence of averages will converge to the mean, μ, of the population" (p. 175). If the reader is interested in exploring this concept further, the reader is directed to George Boger's article that details how to create a spreadsheet simulation of the law of large numbers. In addition, a medical example of this is found in *Observational Studies* (Rosenbaum, 1995, pp. 13-15).

To consider the case of small sample size, let us suppose that I have a sample of 10 graduate students that I am going to randomly assign to one of two treatment groups. The experimental group will have regularly scheduled graduate advisor meetings to monitor students' educational progress. The control group will not have regularly scheduled graduate advisor meetings. Just to see what happens, I choose to do several iterations of this random assignment process. Of course, I discover that the identity of the members in the groups across iterations is wildly different.

Recognizing that most people are outliers on at least some variables (Thompson, 2006), there may be some observed differences that are due simply to the variable characteristics of the members of the treatment groups. For example, let's say that six of the ten graduate students are chronic procrastinators, and might benefit greatly from regular scheduled visits with a graduate advisor, while four of the ten graduate students are intrinsically motivated and tend to experience increased anxiety with frequent graduate advisor inquiries. If the random assignment process distributes these six procrastinator graduate students equally among the two groups, a bias due to this characteristic will not evidence itself in the results. If instead, due to chance all four intrinsically motivated students end up in the experimental group, the results of the study may not be the same had the groups been more evenly distributed. Ridiculously small sample sizes, therefore would result in more pronounced differences between the groups that are not due to treatment effects, but instead are due to the variable characteristics of the members in the groups.

If instead I have a sample of 10,000 graduate students that that I am going to randomly assign to one of two treatment groups, the law of large numbers works for me. As explained by Thompson et al. (2005), "The beauty of true experiments is that the law of large numbers creates preintervention group equivalencies on all variables, even variables that we do not realize are essential to control" (p. 183). While there is still not identical membership across treatment groups, and I still expect that the observed differences between the control group and the experimental group are going to be due to any possible treatment effects

and to the error associated with the random assignment process, the expectation of equality of groups is nevertheless reasonably approximated. In other words, I expect the ratio of procrastinators to intrinsically motivated students to be approximately the same across the two treatment groups. In fact, I expect proportions of variables I am not even aware of to be the same, on average, across treatment groups!

The larger sample size has greatly decreased the error due to chance associated with the random assignment process. As you can see in Figure 2, even if both of the sample studies produce identical treatment effects, the results are not equally valid. The majority of the effect observed in the small sample size study is actually due to error associated with the random assignment process and not a result of the treatment. This effect due to error is greatly reduced in the large sample size study.

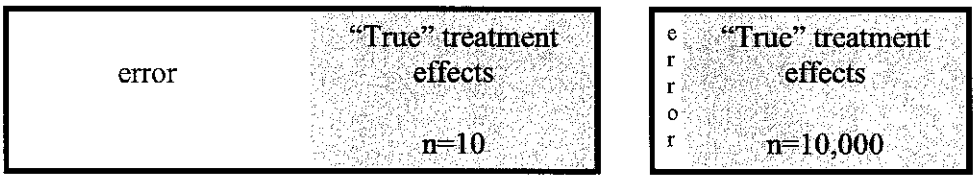


Figure 2. Observed treatment effects in two studies with different sample sizes. The white area represents the amount of the observed effect due to the error associated with the random assignment process. The grey area represents the "true" treatment effect.

Three Experimental Designs

When well-conducted, a randomized experiment is considered the "gold standard" in causal research (Campbell, 1957; Campbell & Stanley, 1963;

Sackett, Strauss, Richardson, Rosenberg, & Haynes, 2000; Thompson, 2006). In fact, "No other type of quantitative research (descriptive, correlational, or causal-comparative) is as powerful in demonstrating the existence of cause-and-effect relationships among variables as experimental research" (Gall et al., 2005, p. 249). There are three designs that meet the characteristics of true experimental designs, first described by Campbell (1957) and revisited in several research design texts. While other designs have the potential to produce causal effects (see Odom et al., 2005; Rosenbaum, 1995; Thompson et al., 2005) only the three classic true experimental designs are discussed in the present paper. For a more extensive description of other experimental designs, the reader is directed to research design works such as Campbell (1957); Campbell and Stanley (1963); Creswell (2003); Gall et al. (2005); Shadish et al. (2002); and Thompson (2006).

The first true experimental design is known as the Pretest-Posttest Control-Group Design. This research design meets the characteristics of a true experiment because participants are randomly assigned (denoted by an R) to either the experimental or control group. There is an intervention or treatment (denoted by an X) given to one group, the experimental group, and no intervention (or alternate intervention) given to the other group, the control group. Finally, there is some form of post-intervention measurement (denoted by an O). This is also known as a posttest, because this measurement occurs after the intervention. In addition, in this particular design, there is also a pretest, denoted by an O prior to the intervention. The pretest allows the researcher to test for

equality of groups on the variable of interest prior to the intervention. These designs are "read" left to right to correspond to the passage of time (i.e., what happens first, second).

Experimental Group	R	O	X	O
Control Group	R	O		O

The second true experiment is the Posttest-Only Control Group Design. This design varies from the first in that it controls for possible confounding effects of a pretest because it does not use a pre-intervention measurement. All three characteristics of a true experimental design are present as in the previous design: random assignment, intervention implemented with experimental group only, and post-intervention measurement.

Experimental Group	R	X	O
Control Group	R		O

The third and final design is the Solomon Four-Group Design. This design is the strongest of the three. It not only corrects for the possible confounding effects of a pretest, but allows you to compare these results, to an experimental and control group that did receive a pretest. The major drawback to this design compared to the others is the obvious increase in sample size needed to meet the needs of four treatment groups as opposed to two treatment groups.

Experimental Group (with pre-test)	R	O	X	O
Control Group (with pre-test)	R	O		O
Experimental Group (without pre-test)	R		X	O
Control Group (without pre-test)	R			O

In addition to detailing these designs in their seminal work, Campbell and Stanley (1963) firmly established their explicit commitment to experiments "as

the only means for settling disputes regarding educational practice, as the only way of verifying educational improvements, and as the only way of establishing a cumulative tradition in which improvements can be introduced without the danger of a faddish discard of old wisdom in favor of inferior novelties"(Campbell & Stanley, 1963, p. 172).

Validity Threats

Even when these designs are used, there are differences in how rigidly they are followed as well as to what extent the researcher addresses the multiple threats to validity (see Figure 3 below). Threats to validity are important not only to research designer but also to consumers of research. An informed consumer of research wants to rule out all competing hypothesis and be firmly convinced that the evidence supports the claim that the IV caused the DV. To merit this conclusion, an evaluation of the study is necessary to determine whether threats to experimental validity were recognized *and* mitigated.

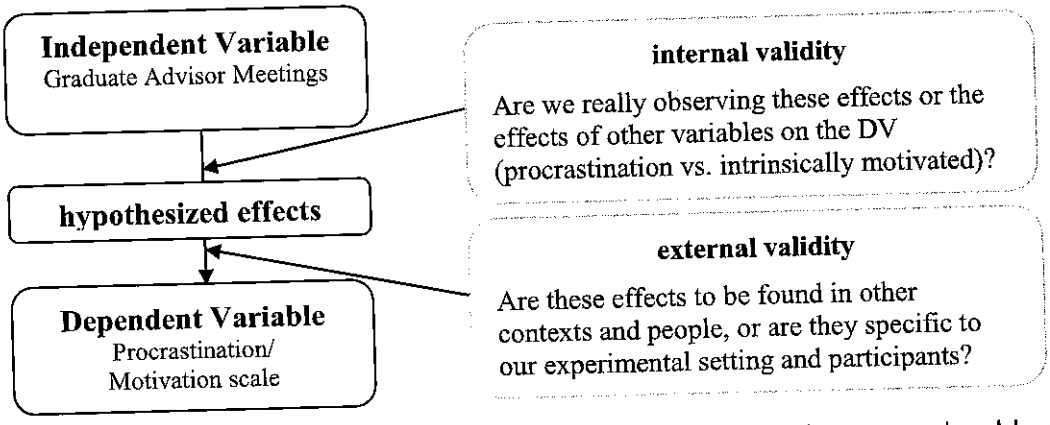


Figure 3. Example of a research experiment and the questions you should ask yourself about internal and external validity. Adapted from (Sani & Todman, 2006).

Internal Validity

Creswell defines internal validity threats as those "experimental procedures, treatments, or experiences of the participants that threaten the researchers' ability to draw correct inferences from the data in an experiment" (2003, p. 171). In their classic text, Campbell and Stanley (1963) identified eight threats to internal validity. In a more recent text, Shadish, Cook and Campbell (2002) addressed nine threats to validity which are described below. For an extensive list of threats to internal and external validity, the reader is directed to Onwuegbuzie's work that cogently expresses the need to evaluate "*all* quantitative research studies" (2000, p. 7), not just experimental design studies, for threats to internal and external validity.

1. Ambiguous temporal precedence: uncertainty about which occurred first (IV or DV) which would lead to questions about which variable is the cause and which is the effect.

2. Selection bias: a systematic bias resulting in non-random selection of participants to groups. By definition random assignment prevents selection bias, if and only if the law of large numbers can be invoked.

3. History: an event that may occur between measurements that is not part of the intervention that could impact the posttest measurement. For example, let us return to the ten fictional graduate students described previously in the study. Let's say they were all living in the same dorm and the fire alarm kept going off the night before they were to take the motivation/ procrastination measurement instrument. Due to lack of sleep, participants may perform differently on the

motivation/ procrastination scale than they would have had they gotten enough sleep.

4. Maturation: an observed change that is naturally occurring (such as aging, fatigue, hair length, number of graduate hours completed) that may be confused with the intervention effects but is really a function of the passage of time.

5. Statistical regression: the phenomenon that occurs when participant selection is based on extreme scores whereby the scores become less extreme, which may appear to be the intervention effect. If in our study of graduate students we purposively select students based on pretest scores of extreme procrastination, the extreme procrastinator graduate students will on the posttest not be as extreme in their procrastination tendencies.

Regression toward the mean was first documented by Sir Francis Galton in the late 1800s. Galton (1886) measured the heights of fathers and sons at a World Exposition. Galton found that very tall fathers tended to have sons who were not quite as tall, and that very short fathers tended to have sons who were not quite as short. Clearly, this phenomenon is not a function of the exercise of will (i.e., fathers did not say to their wives, "Let's make a shorter son" or "Let's make a taller son")!

6. Experimental mortality or attrition: a concern about a differential loss of participants, or of different types of participants from the experimental or control group that may produce an effect that appears to be due to the intervention. For example, if half of the students in the experimental group drop out of the study, but none of the control group members drop, we would likely question the results.

Were those students that left somehow different from the ones that remained? If so, would that difference have produced differential results than the ones we observed with the remaining participants?

7. Testing: the concern that a testing event will impact scores of a subsequent testing event. For example, if we give the graduate students the procrastination/ motivation scale prior to any graduate advisor meetings (the intervention), and then after the intervention we give them the procrastination/ motivation scale again, we may observe difference in the pre- and posttest that are due partly to familiarity with the test or the influence of the testing itself.

8. Instrumentation: the change in either the measurement instrument itself or the manner in which the instrument is implemented or scored that may cause changes that appear to be due to the intervention, or the failure to detect changes that actually did occur. For example, if between the first and second time that the procrastination/motivation test is given, the developers of the exam decide to remove ten of the questions, we do not know if the exclusion of those questions is responsible for differential scores or if the differences are due to treatment effects.

9. Additive and interactive effect of threats to internal validity: the concern that the impact of the threats may be additive or that presence of one threat may impact another. A selection-history additive effect occurs when nonequivalent groups are selected. For example, groups may be selected from two different locations, such as, rural and urban areas. The participants in the groups are nonequivalent by selection *and* they also have unique local histories. The

resulting net bias is dependent on both the direction and magnitude of each individual bias and how the biases combine. Selection-maturation, and selection-instrumentation are other versions of this type of effect.

External Validity

External validity threats are threats of "incorrect inferences from the sample data to other persons, other settings, and past or future situations" (Creswell, 2003, p. 171). Researchers must always remember the context from which their sample comes from, and take caution not to overgeneralize beyond that.

Campbell and Stanley (1963) included four threats to external validity. Shadish (2002) listed five external validity threats, as detailed below.

1. Interaction of the causal relationship with participants: an effect with certain kinds of participants that may not be present (or present to the same extent) with other kinds of participants. For example, reduction of salt intake in hypertensive patients is more beneficial to certain populations than others (American Heart Association Nutrition Committee, 2006).
2. Interaction of the causal relationship over treatment variations: the permanence of the causal relationship is dependent on fidelity to the specific treatment, thus possibly producing differential effects when treatments are varied. If a particular instructional intervention includes 5 components, the causal relationship may not hold if only 2 or 3 of the components are utilized.
3. Interaction of the causal relationships with outcomes: an effect that is present with one type of outcome measurement that may not be present (or present to

the same extent) if other outcome measurements were used. For example, if a person scores highest on a test for physical strength they may not necessarily score highest on a flexibility test.

4. Interactions of the causal relationship with settings: an effect that is present in a particular setting may not be present (or present to the same extent) in a different setting. For example, a particular after school character development program involving community project work may not work equally well in rural versus urban areas.

5. Context-dependent mediation: an explanatory mediator of a causal relationship in one context may not have the same impact in another context. For example, a study might find that a reduction in federal funding has no impact on student achievement because schools were able to turn to education foundation grants to provide them with additional resources. In another school district where schools did not have access to education foundation resources, the same causal mechanism may not be available.

In addition to internal and external validity threats, there are other threats that we need to be aware of in the design and evaluation of studies. Interested readers may refer to such texts as *Experimental and Quasi-Experimental Designs* (Shadish et al., 2002) or *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (Creswell, 2003) for information about statistical conclusion validity and construct validity concerns.

EBP in Medicine, Psychology and Education

While the origins of EBP may date back to the origin of scientific reasoning, the Evidence-Based Medicine Working Group (EBMWG) brought the discussion of EBP to the forefront of medicine (1992). In 1996, Evidence-Based Medicine (EBM) was defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research" (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996, p. 71). While EBP has many supporters in medicine, EBP has caused some concerns among practitioners. Researchers have addressed concerns regarding the perception of EBM as a top down approach that results in ivory tower researchers dictating how practitioners should practice (Sackett et al., 1996) or similarly that evidence from randomized controlled trials may be valued more highly than practitioner expertise (Kübler, 2000).

Yet, it is difficult to deny that there is great support for EBP considering the number of periodicals that have emerged since the years after EBMWG convened. A keyword search for "evidence-based" returns 100 serials on *WorldCAT*. A keyword search for "evidence-based" returns 96 serials in *Ulrich's Periodical Directory*. At least 32 active periodicals, either in print form, electronic form, or both contain "evidence-based" within the title of the periodical. At least 26 of these periodicals are available electronically. See Table 1.

From the titles you can see that the majority of these periodicals are from a health-related field. It is important to note that while EBP do not only include randomized, experimental trials, the purpose of the table is to demonstrate the popularity of EBP that began in the mid 1990s and continues today.

Table 1

"Evidence-Based" periodicals

Start Year	Title of Periodical
1994	**Bandolier: Evidence-Based Healthcare**
1995	**Evidence-Based Medicine**
1996	**Focus on Alternative and Complementary Therapies: An Evidence-Based Approach**
1997	**Evidence-Based Cardiovascular Medicine**
1997	Evidence-Based Medicine in Practice
1997 (1998)	**Evidence-Based Mental Health**
1997 (1998)	**Evidence-Based Nursing**
1997	**Evidence-Based Obstetrics and Gynecology**
1998	**EBN Online**
1998	**Evidence-Based Dentistry**
1998	Evidence-Based Practice
1998	**Evidence-Based Practice: Patient Oriented Evidence That Matters**
1999	Evidence-Based Dental Practice
1999 (2002)	**Trends in Evidence-Based Neuropsychiatry: T.E.N.**

Table 1 (continued).

Start Year	Title of Periodical
2000	**Evidence-Based Gastroenterology**
2000	**Evidence-Based Oncology**
2000	Trauma Reports: Evidence-Based Medicine for the ED
2001	**Journal of Evidence-Based Dental Practice**
2003	**Evidence-Based Integrative Medicine**
2003	Evidence-Based Midwifery
2003	**Evidence-Based Preventive Medicine**
2003	**Evidence-Based Surgery**
2003 (2005)	**International Journal of Evidence-Based Healthcare**
2004	**Evidence-Based Complementary and Alternative Medicine: eCAM**
2004	**Journal of Evidence-Based Social Work**
2004	**Worldviews on Evidence-Based Nursing**
2005	**Advances in Psychotherapy: Evidence-Based Practice**
2005	**Evidence-Based Ophthalmology**
2005	Journal of Evidence-Based Practices for Schools
2006	**Evidence-Based Child Health**
2006	**Evidence-Based Library and Information Practice**
2007	**Evidence-Based Communication Assessment and Intervention**

Periodicals available electronically are shown in bold.
Parenthetical dates indicate different start year date in *WorldCAT*.

The popularity of EBP is evident in psychology as well. The American Psychological Association's Presidential Task Force on Evidence-Based Practice specifically defined Evidence-Based Practice in Psychology (EBPP) as "the integration of the best available research with clinical expertise in the context of patient characteristics, culture, and preferences" (2006, p. 273). In addition to advocating evidence-based practices, this task force also established the two necessary components for evaluation of psychological interventions: treatment efficacy and clinical utility. Treatment efficacy specifically addresses questions such as how well a particular treatment works. This type of question lends itself to experimental investigation to draw valid causal conclusions about the effect of a particular intervention (or lack thereof) on a particular disorder (American Psychological Association, 2002). Chambless and Hollon (1998), in their review of psychological treatment literature, provide a description of variables of interest when evaluating treatment efficacy in research studies. The Task Force acknowledged that while there are other methods that may lead to causal conclusions "randomized controlled experiments represent a more stringent way to evaluate treatment efficacy because they are the most effective way to rule out threats to internal validity in a single experiment" (American Psychological Association, 2002, p. 1054).

The appeals for evidence continue also in the field of education. Grover J. (Russ) Whitehurst, who directs the Education Department's Institute of Education Sciences, defined Evidence-Based Education (EBE) as "the integration of professional wisdom with the best available empirical evidence in making

decisions about how to deliver instruction" (in Towne, 2005, p. 41). Whitehurst (2002b) explained that without empirical evidence education is at the mercy of the latest educational craze. In addition, asserted that cumulative knowledge cannot be generated without empirical evidence. To assist education practitioners in the identification of EBP, a practical guide has been provided (see Coalition for Evidence-Based Policy, 2003).

Table 2

Definitions of EBP in medicine, psychology and education

Field	Definition
Evidence-based medicine (EBM)	"the conscientious, explicit, and judicious use of current best evidence in **making decisions** about the care of individual patients. The practice of evidence based medicine means **integrating** individual clinical **expertise** with the **best available** external clinical **evidence** from systematic research" (Sackett et al., p. 71).
Evidence-based practices in psychology (EBPP)	"the **integration** of the **best available research** with clinical **expertise** in the context of patient characteristics, culture, and preferences" (American Psychological Association, 2006, p. 273).
Evidence-based education (EBE)	"the **integration** of professional **wisdom** with the **best available** empirical **evidence** in **making decisions** about how to deliver instruction" (Whitehurst, 2002b, Slide 3).

Medicine, psychology, and education all have seemed to have jumped on the evidence wagon. Their definitions share the common themes of integration of

expertise with the best available evidence (see Table 2 above). We *cannot* ignore this need to balance practitioner expertise with empirical evidence whether in the field of medicine, psychology or education. As Kübler (2000) cautions:

> Undoubtedly evidence based medicine is the gold standard for modern medicine. The results, however, should be applied in patient care with careful reflection. Otherwise evidence based medicine may acquire the same status for the doctor as a lamp post for a drunk: it gives more support than enlightenment. (p. 135)

Frequency of Experiments in Different Disciplines

One final caution is offered. It is imperative that consumers and producers of research critically evaluate research. In addition to threats to validity, we must keep in mind that experiments are conducted by people. People are fallible. We are prone to make mistakes, both consciously and unconsciously. An example of this is a graph that appears to be from the same data, yet describes different results. What is critical about these graphs is that depending on which one you look at, education ranks third, fourth, or first in cumulative total number of reports of trials identified from the *Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR)* (Petrosino, Boruch, Rounding, McDonald, & Chalmers, 2000).

One graph depicts education behind criminology and psychology, but ahead of social policy (Boruch, Moya, & Snyder, 2002, p. 63). The authors describe the graph as follows:

> Figure 3-4 shows the increase in the number of articles on randomized and possibly randomized experiments that have appeared in about 100 peer-reviewed journals and in other places since 1950. The figure is based on the Campbell Collaboration Social, Psychological, Educational, and Criminological Trials Registry (C2-SPECTR) that is being developed in a continuing effort to identify all RFTs. (p. 62).

The authors correctly cite Petrosino et al. (2000) as the source of the graph.

In another graph, which cites Boruch et al. (2002), education is now in *last place* behind criminology, psychology and social policy respectively (Whitehurst, 2002b). The following description was offered in Whitehurst's (2002b) presentation:

> This chart indicates the total number of articles about randomized field trials in other areas of social science research (criminology, social policy and psychology) has steadily grown over the last 40 years; however, the number related to education research has trailed behind. (Table Description, Slide 22)

In a very similar presentation by Whitehurst (2002a), a more extensive description of the same graph is provided:

> While the total number of articles about randomized field trials in other areas of social science research has steadily grown, the number in education research has trailed behind. The graph on this slide measures the growth of randomized field trials from 1950 to the present in the areas of criminology, social policy, psychology, and education. It shows that the

> most rapid growth has been in criminology, followed by comparable rates of growth in social policy and psychology, with education having the least amount of growth. Source for the graph: Robert Boruch, Dorothy de Moya, and Brooke Snyder, 2001. (slide 21)

The correct year for the citation is actually 2002.

Finally, in still another version of the graph, education is leading the pack followed by psychology, social and criminology (Petrosino, Boruch, Soydan, Duggan, & Sanchez-Meca, 2001). The following description is offered:

> To facilitate the work of reviewers, the *Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register* (*C2-SPECTR*) is in development. As Figure 2 shows, preliminary work toward *C2-SPECTR* has already identified more than 10000 citations to randomized or possibly randomized trials. (p. 28)

Petrosino et al. (2001) cite Petrosino et al. (2000), the same reference cited in Boruch et al. (2002). The only difference is that incorrect page numbers are given here. Instead of correctly identifying the pages as 206-219, Petrosino et al. (2001) identify pages 293-307.

Aside from the citations errors, one would hope that clarity about the results of the graph would be found in the original citation. Is education fourth, third or first in cumulative number of reports of randomized trials? The original citation, Petrosino et al. (2000) *does* match the results of the graph in Petrosino et al. (2001), but *not* the results of the graphs in Whitehurst (2002b) or Boruch et al. (2002).The original source offers the following description for the chart:

> C2-SPECTR thus currently contains a total of 10,449 records. Figure 1 shows cumulative totals of reports of trials published between 1950 and 1998, subdivided on the basis of the 'high level' codes which were assigned to indicate the sphere(s) of intervention. (p. 211)

See Figure 4 for a visual explanation.

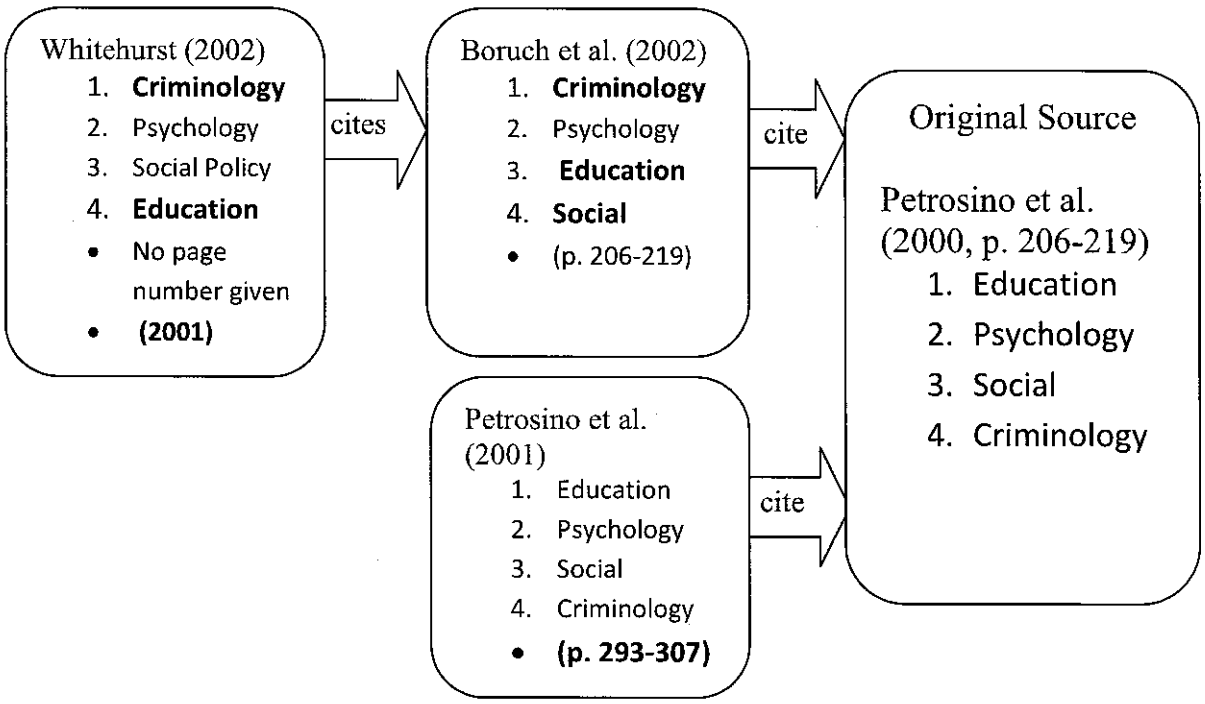


Figure 4. Diagram of citation errors. Deviations from original source are shown in bold.

Examining the graphs, it is easy to see how these changes could have been made inadvertently. Nonetheless, one has to consider the impact that these errors may have had. Whitehurst's presentation was disseminated in "a series of four regional meetings as part of its work to ensure the effective implementation of the *No Child Left Behind* (NCLB) Act" (U. S. Department of Education, 2002). In addition, the *Web of Science* shows that this presentation was cited at least 6

times, *Evidence Matters* (Mosteller & Boruch, 2002) was cited at least 35 times, and the original source (Petrosino et al., 2000) was cited 12 times with the correct page number *and* 6 times with the incorrect page number which is given in Petrosino et al. (2001).

Whitehurst's (2002b) presentation was described in a report by WestEd titled *Scientific Research and Evidence-Based Practice* (Hood, 2003). Hood gives the following description of the graph in Whitehurst's presentation:

> 22. Education Lags Behind Chart Description: This chart indicates the total number of articles about randomized field trials in other areas of social science research (criminology, social policy and psychology) has steadily grown over the last 40 years; however, the number related to education research has trailed behind. [By approximately 1996, the cumulative number of articles about definite and possible randomized field trials in criminology is approaching 6,000; the numbers in social policy and psychology exceed 2,000; while the number for education is less than 1,000.] (p.22)

In addition, Whitehurst's presentation is identified as one of the Editor's Picks under *Proven Methods*: *Doing What Works* within the *NCLB* page on the U.S. Department of Education's Website (see http://www.ed.gov/nclb/methods/whatworks/edpicks.jhtml). Colorado's Department of Education has apparently incorporated Whitehurst's graph into their *Fast Facts: Evidence-Based Practice* (2005). Perhaps because of Whitehurst's position as the Director of the Institute of Education Sciences, or

perhaps because of the wide dissemination of this presentation, citations alone are not enough to measure the impact that his presentation has had.

Errors in scholarly reports are not new. Thompson (1988, 1994) examined methodological mistakes in dissertations. Doctoral students and the prevalence of documentation errors are discussed in a recent article where the authors give several sources that address documentation errors in the literature such as "citation errors (for example, non-compliance to the prescribed editorial style), reference omissions, reference falsification, inconsistent references, inaccurate quotations, misspelled names, incorrect page numbers, and even fraudulent research" (Waytowich, Onwuegbuzie, & Jiao, 2006, p. 196). Mistakes will always be present; it is up to the research community, and informed consumers to make wise decisions regarding the worth of studies. There is no substitute for good judgment.

Summary

While true experiments do have the potential to provide the best possible causal evidence, it is imperative to keep in mind the threats that may undermine confidence in the findings, from internal and external validity threats, to simple human errors. In the wise words of Sackett and colleagues, the purpose of this type of research is to *inform*, but not to replace individual practitioner's knowledge (Sackett et al., 1996). This implies judgment on the part of the reader.

References

American Heart Association Nutrition Committee. (2006). AHA Scientific statement: Diet and lifestyle recommendations revision 2006. *Circulation, 114*, 82-96.

American Psychological Association. (2002). Criteria for evaluating treatment guidelines. *American Psychologist, 57*, 1052-1059.

American Psychological Association. (2006). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285.

Boger, G. (2005). Spreadsheet simulation of the law of large numbers. *Mathematics and Computer Education, 39,* 175-182.

Boruch, R., Moya, D. D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 50-79). Washington, DC: Brookings Institution Press.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297-312.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Chicago: Rand McNally.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*, 7-18.

Coalition for Evidence-Based Policy. (2003). Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide.

Colorado Department of Education. (2005). Fast facts: Evidence-based practice. Retrieved Jan 3, from http://www.cde.state.co.us/cdesped/download/pdf/ff-EvidenceBasedPractice_Intro.pdf

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Evidence-Based Medicine Working Group. (1992). Evidence-based medicine: A new approach to teaching the practice of medicine. *Journal of the American Medical Association, 268*, 2420-2425.

Gall, J. P., Gall, M. D., & Borg, W. R. (2005). *Applying educational research: A practical guide* (5th ed.). Boston: Pearson Education, Inc.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute, 15*, 246-263.

Hood, P. D. (2003). Scientific research and evidence-based practice. Retrieved January 2, 2008, from http://www.wested.org/online_pubs/scientrific.research.pdf

Kübler, W. (2000). Treatment of cardiac diseases: Evidence based or experience based medicine? *84*, 134-136.

Mosteller, F., & Boruch, R. (Eds.). (2002). *Evidence matters: Randomized trials in education research.* Washington, DC: Brookings Institution Press.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidenced-based practices. *Exceptional Children, 71*, 137-148.

Onwuegbuzie, A. J. (2000, November). *Expanding the framework of internal and external validity in quantitative research.* Paper presented at the annual meeting of the Association for the Advancement of Educational Research, Ponte Vedra, FL. (ERIC Document Reproduction Service No. ED 448 205).

Petrosino, A., Boruch, R. F., Soydan, H., Duggan, L., & Sanchez-Meca, J. (2001). Meeting the challenges of evidence-based policy: The campbell collaboration *The ANNALS of the American Academy of Political and Social Science, 578*, 14-34.

Petrosino, A. J., Boruch, R. F., Rounding, C., McDonald, S., & Chalmers, I. (2000). The campbell collaboration social, psychological, educational and criminological trials register (C-2 SPECTR) to facilitate the preparation and maintenance of systematic reviews of social and educational interventions. *Evaluation and Research in Education, 14*, 206-219.

Rosenbaum, P. R. (1995). *Randomized experiments*. New York: Springer-Verlag.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal, 312*(7023), 71-72.

Sackett, D. L., Strauss, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (Eds.). (2000). *Evidence-based medicine: How to practice and teach EBM.* New York: Churchill Livingstone.

Sani, F., & Todman, J. (2006). *Experimental design and statistics for psychology: A first course*. Malden, MA: Blackwell Publishing.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education.* Washington, DC: National Academy Press.

Thompson, B. (1988, November). *Common methodology mistakes in dissertations: Improving dissertation quality.* Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595).

Thompson, B. (1994, April). *Common methodology mistakes in dissertations, revisited.* Paper presented at the Annual Meeting of the American Educational Research Association New Orleans, LA. (ERIC Document Reproduction Service No. ED 368 771).

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach.* New York: Guilford.

Thompson, B., Diamond, K. E., McWilliam, R., Snyder, P., & Snyder, S. W. (2005). Evaluating the quality of evidence from correlational research for evidence-based practice. *Exceptional Children, 71*, 181-194.

Towne, L. (2005). Scientific evidence and inference in educational policy and practice: Defining and implementing "Scientifically based research". In C. A. Dwyer (Ed.), *Measurement and Research in the Accountability Era* (pp. 41-58). Mahwah, NJ: Routledge.

U.S. Congress. (2001). No Child Left Behind Act of 2001, *Public Law No. 107-110.* Washington, DC.

U. S. Department of Education. (2002). Lead & manage my school: Student achievement and school accountability conference. Retrieved December 27, 2007, from http://www.ed.gov/admins/lead/account/sasaconference02.html

Waytowich, V. L., Onwuegbuzie, A. J., & Jiao, Q. G. (2006). Characteristics of doctoral students who commit citation errors. *Library Review, 55*, 195-208.

Whitehurst, G. J. (2002a). Archived evidence-based education (EBE). Retrieved December 20, 2007, from http://www.ed.gov/offices/OERI/presentations/evidencebase.html

Whitehurst, G. J. (2002b). Evidence-based education (EBE). On *Student Achievement and School Accountability Conference.* Retrieved December 20, 2007, from http://www.ed.gov/nclb/methods/whatworks/eb/edlite-index.html

Vol 466|1 July 2010

(Source 2)

nature

OPINION

Most people are not WEIRD

To understand human psychology, behavioural scientists must stop doing most of their experiments on Westerners, argue **Joseph Henrich**, **Steven J. Heine** and **Ara Norenzayan**.

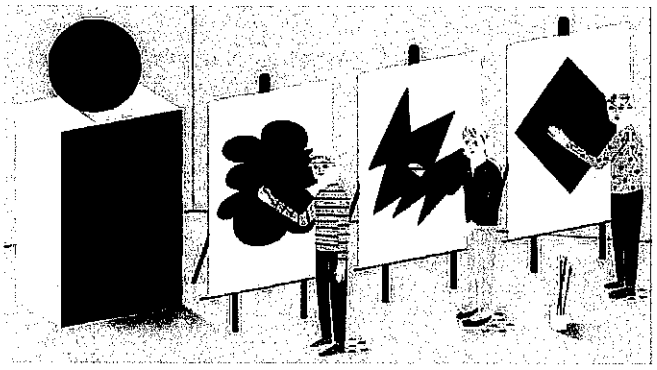
Much research on human behaviour and psychology assumes that everyone shares most fundamental cognitive and affective processes, and that findings from one population apply across the board. A growing body of evidence suggests that this is not the case.

Experimental findings from several disciplines indicate considerable variation among human populations in diverse domains, such as visual perception, analytic reasoning, fairness, cooperation, memory and the heritability of IQ[1,2]. This is in line with what anthropologists have long suggested: that people from Western, educated, industrialized, rich and democratic (WEIRD) societies — and particularly American undergraduates — are some of the most psychologically unusual people on Earth[1].

So the fact that the vast majority of studies use WEIRD participants presents a challenge to the understanding of human psychology and behaviour. A 2008 survey of the top psychology journals found that 96% of subjects were from Western industrialized countries — which house just 12% of the world's population[3]. Strange, then, that research articles routinely assume that their results are broadly representative, rarely adding even a cautionary footnote on how far their findings can be generalized.

The evidence that basic cognitive and motivational processes vary across populations has become increasingly difficult to ignore. For example, many studies have shown that Americans, Canadians and western Europeans rely on analytical reasoning strategies — which separate objects from their contexts and rely on rules to explain and predict behaviour — substantially more than non-Westerners. Research also indicates that Americans use analytical thinking more than, say, Europeans. By contrast, Asians tend to reason holistically, for example by considering people's behaviour in terms of their situation[1]. Yet many long-standing theories of how humans perceive, categorize and remember emphasize the centrality of analytical thought.

It is a similar story with social behaviour related to fairness and equality. Here, researchers often use one-shot economic experiments such as the ultimatum game, in which a player decides how much of a fixed amount to offer a second player, who can then accept or reject this proposal. If the second player rejects it, neither player gets anything. Participants from industrialized societies tend to divide the money equally, and reject low offers. People from non-industrialized societies behave differently, especially in the smallest-scale non-market societies such as foragers in Africa and horticulturalists in South America, where people are neither inclined to make equal offers nor to punish those who make low offers[4].



GRACIA LAM

Recent developments in evolutionary biology, neuroscience and related fields suggest that these differences stem from the way in which populations have adapted to diverse culturally constructed environments. Amazonian groups, such as the Piraha, whose languages do not include numerals above three, are worse at distinguishing large quantities digitally than groups using extensive counting systems, but are similar in their ability to approximate quantities. This suggests the kind of counting system people grow up with influences how they think about integers[1].

Costly generalizations

Using study participants from one unusual population could have important practical consequences. For example, economists have been developing theories of decision-making incorporating insights from psychology and social science — such as how to set wages — and examining how these might translate into policy[5]. Researchers and policy-makers should recognize that populations vary considerably in the extent to which they display certain biases, patterns and preferences in economic decisions, such as those related to optimism[1]. Such differences can, for example, affect the way that experienced investors make decisions about the stock market[6].

We offer four suggestions to help put theories of human behaviour and psychology on a firmer empirical footing. First, editors and reviewers should push researchers to support any generalizations with evidence. Second, granting agencies, reviewers and editors should give researchers credit for comparing diverse and inconvenient subject pools. Third, granting agencies should prioritize cross-disciplinary, cross-cultural research. Fourth, researchers must strive to evaluate how their findings apply to other populations. There are several low-cost ways to approach this in the short term: one is to select a few judiciously chosen populations that provide a 'tough test' of universality in some domain, such as societies with limited counting systems for testing theories about numerical cognition[1,2].

A crucial longer-term goal is to establish a set of principles that researchers can use to distinguish variable from universal aspects of psychology. Establishing such principles will remain difficult until behavioural scientists develop interdisciplinary, international research networks for long-term studies on diverse populations using an array of methods, from experimental techniques and ethnography to brain-imaging and biomarkers.

Recognizing the full extent of human diversity does not mean giving up on the quest to understand human nature. To the contrary, this recognition illuminates a journey into human nature that is more exciting, more complex, and ultimately more consequential than has previously been suspected ■

Joseph Henrich, **Steven J. Heine** and **Ara Norenzayan** are in the Department of Psychology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. Joseph Henrich is also in the Department of Economics.
e-mail: joseph.henrich@gmail.com

1. Henrich, J., Heine, S. J. & Norenzayan, A. *Behav. Brain Sci.* doi:10.1017/S0140525X0999152X (2010).
2. Henrich, J., Heine, S. J. & Norenzayan, A. *Behav. Brain Sci.* doi:10.1017/S0140525X10000725 (2010).
3. Arnett, J. *Am. Psychol.* **63,** 602–614 (2008).
4. Henrich, J. *et al. Science* **327,** 1480–1484 (2010).
5. Foote, C. L., Goette, L. & Meier, S. *Policymaking Insights from Behavioral Economics* (Federal Reserve Bank of Boston, 2009).
6. Ji, L. J., Zhang, Z. Y. & Guo, T. Y. *J. Behav. Decis. Making* **21,** 399–413 (2008).

Assignment (2)

Prior to beginning work on this assignment, be sure to have read all the required resources for the week.

Find an experimental research study on the topic chosen in Week One for your Final Research Proposal. You may choose to include an experimental study which was included in the literature review you used in the Week One assignment by searching the reference list for experimental research studies on the topic. However, it is also acceptable to find and include an experimental research study on the topic that is not included in that literature review.

Identify the specific experimental research design used in the study. Summarize the main points of the experimental research study including information on the hypothesis, sampling strategy, research design, statistical analysis, results, and conclusion(s). Evaluate the published experimental research study focusing on and identifying the specific threats to validity that apply to the chosen study. Explain whether or not these threats were adequately addressed by the researchers. Describe how the researchers applied ethical principles in the research study.

The Research and Critique an Experimental Study

- Must be three to four double-spaced pages in length (not including title and reference pages) and formatted according to APA style as outlined in the Ashford Writing Center (Links to an external site.).
- Must include a separate title page with the following:
  - Title of paper
  - Student's name
  - Course name and number
  - Instructor's name
  - Date submitted
- Must use at least one peer-reviewed source in addition to those required for this week.
- Must document all sources in APA style as outlined in the Ashford Writing Center.
- Must include a separate reference page that is formatted according to APA style as outlined in the Ashford Writing Center.

Carefully review the Grading Rubric (Links to an external site.) for the criteria that will be used to evaluate your assi

I have attached the Book Source

Please include another peer-Reviewed Source →

Martin, W.E., & Bridgmon, K.D. (2012). Research methods for the social sciences, Volume 42: Quantitative and Statistical research Methods: From hypothesis to results. Somerset, NJ: John Wiley & Sons.

Book Source

Chapter 4

RESEARCH AND STATISTICAL DESIGNS

LEARNING OBJECTIVES

- Understand the importance of carefully formulating experimental conditions and procedures, reducing the imprecision in measurement, and controlling extraneous experimental influences.
- Explore methods to control for extraneous variables.
- Understand the threats to internal validity of experimental and quasi-experimental designs.
- Examine commonly used experimental and quasi-experimental designs and examples of statistical methods that can be used in the designs.
- Understand basic models used in correlation methods.
- Understand issues to consider in choosing statistics to use in different research problems.

The purpose of this chapter is to identify important components of establishing experimental and quasi-experimental research by examining various designs. Correlational research models are also discussed.

Statistical designs are essentially linked to research designs. Research designs provide the blueprint for researchers to find answers to research questions. The choice and interpretation of statistics used in a particular study is dependent upon the research design selected by a researcher. A class of research designs called experimental designs has the purpose to assess outcomes of cause and effect among variables. Generally, the purpose of experimental research is to manage and understand the change among variables by *maximizing experimental variance*, *minimizing error variance*, and *controlling extraneous variance* (*MaxMinCon*) (Kerlinger & Lee, 2000).

A researcher maximizes experimental variance by carefully formulating experimental conditions and procedures. A researcher minimizes error variance by strategically reducing the imprecision in measurement and random error fluctuations. Controlling extraneous variance is dealt with by identifying and reducing the effects of unwanted variables through the identification of well-suited design, experimental procedures, and statistics (Martin & Bridgmon, 2009). A researcher considers several issues when choosing statistics to use for varying research problems.

FORMULATING EXPERIMENTAL CONDITIONS

Experimental research is initiated based on well-developed research questions, with alternative and null hypotheses usually connected to theory and previous research. *Independent (cause) variables* and *dependent (outcome) variables* are *operationally defined*. Operation definitions entail describing the variables in detail, especially as to how they function and are measured.

For example, suppose a treatment method for seasonal affect disorder (SAD) is an independent variable in a study and there are three treatment conditions (light therapy, cognitive-behavioral therapy (CBT), and a minimal contact-delayed treatment control condition). The dependent variable is the intensity of symptoms of a current SAD episode.

The independent variable is assessed as to whether it was implemented as planned, which is known as *treatment integrity (treatment fidelity)*. A researcher

operationally defines every aspect of the experimental conditions to ensure an accurate and consistent implementation of the independent variable (*treatment delivery*). Equally important is defining how to assess whether the participants received (*treatment receipt*) and followed through with (*treatment adherence*) the intended treatment (Shadish, Cook, & Campbell, 2002).

The dependent variable of intensity of symptoms of the current SAD episode also needs to be operationally defined to know what the variable is and how it will be measured. The dependent variable might be operationally defined as the scores on a psychometrically validated instrument such as the Structured Interview Guide for the Hamilton Rating Scale of Depression and SAD subscale (SIGH-SAD). The researcher comprehensively identifies research that substantiates the validity and reliability of the SIGH-SAD.

Operationalism is vital to understanding research and maintaining objectivity. Moreover, clearly defined experimental procedures and measurements make it easier for researchers to replicate studies to confirm or repudiate important findings.

REDUCING THE IMPRECISION IN MEASUREMENT

All research includes fluctuations between an observed quantity and its true quantity. These fluctuations or differences are referred to as *errors* that represent imprecision in measurement. Two important sources of error are sampling error and error of measurement.

Sampling Error

A sample selected for study is intended to represent a population. Inferences about sample statistics such as a mean and standard deviation are used to estimate population parameters. It is highly unlikely that sample statistics will be exactly the same as population parameters. The difference between a sample mean and a population mean is referred to as *sampling error*. Indexes representing sampling error are available to help researchers understand the differences between sample statistics and population parameters. All parametric statistical procedures have sampling error terms to account for sampling error. For example, the sampling error index for the difference between a sample and a population mean (one-sample *t*-test) is called the *standard error of the mean*. The standard error terms associated with some other statistics include independent *t*-test (standard

error of difference), analysis of variance *F*-test (mean square error), correlation coefficient-*r* (standard error of estimate), and multiple regression (standard error of regression coefficient).

The use of *random sampling* is the best way to reduce sampling error. Random sampling involves taking a random sample from a larger population by selecting cases by chance to represent the population. This probabilistic approach will result in the most representative sample if it is done correctly. For example, the first step for a researcher who wants to study third graders in a school district is to obtain a complete list of all third graders. The researcher would choose a desired sample size and then use a random number generator to select the number for the sample from the complete list of third graders (population).

Ideally, the two-step process of first randomly selecting a sample of participants and then *randomly assigning* them to a treatment condition is used in research for generalized causal inference. However, it is uncommon for participants to be randomly selected to participate in experimental studies (Shadish, Cook, & Campbell, 2002). Random assignment of participants to treatment conditions is required for a study to be considered experimental. The distinguishing characteristic of *quasi-experimental* research is that random assignment is not used.

Shadish et al. articulate an alternative selection method to random selection used by experimental researchers known as *purposive sampling*. Purposive sampling is a deliberate process by a researcher to identify population characteristics, including participants, treatments, outcomes, and settings, and then select a sample that embodies the desired population characteristics.

Error of Measurement

A second source of error is *error of measurement*. When using measures from participants to assess the dependent variables in a study, there are differences between the participants' observed scores and their true scores. An *observed score* is the actual measured score of a participant, and a *true score* is the actual score (theoretical) on the measure. Conceptually, an observed score is similar to a sample characteristic, and a true score is similar to a population parameter.

Error of measurement can be attributed to both *random error* and *systematic error*. Random error includes fluctuations is the characteristics of individuals, such as the motivation levels of participants changing randomly from one assessment to the next.

A novice examiner may depart from standardized assessment procedures in scoring or use an unreliable test in a study; these are examples of sources of random error. Measurement reliability is especially affected by random error.

Systematic error is represented by a decrease or an increase of scores on a dependent measure in a predictable way. This type of systematic error often relates to procedural irregularities in the research process. A weight scale calibrated incorrectly and consistently measuring three pounds heavier than it should represents systematic error. Systematic error most adversely affects measurement validity.

Measurement error exists in all studies. We cannot eliminate errors in measurement, but we can minimize their effects.

CONTROLLING EXTRANEOUS EXPERIMENTAL INFLUENCES

Experimental research strives to accurately and consistently measure whether changes in a dependent variable can be attributed to a specified independent variable. *Extraneous variables (nuisance variables)* confuse the assessment of the real effects of the independent variable on the dependent variable. Extraneous variables are first identified and then efforts are made to control their confounding effects.

For example, a researcher wants to compare the effects of a hypnosis treatment program to a nicotine gum therapy among smokers to see which program is more effective in producing smoking abstinence. It is discovered that the group of participants receiving the hypnosis treatment have a significantly lower average number of years smoking compared to the nicotine gum therapy participants. Prior number of years smoking is an extraneous variable that is salient in its potential impact on the dependent variable of smoking abstinence. This extraneous variable needs to be accounted for and controlled to accurately understand the cause-and-effect relationship between treatment and outcome.

Methods of Controlling Extraneous Variables

There are several general ways to control for extraneous variables. One way is to *eliminate an extraneous variable.* For example, a researcher is concerned about the extraneous effects of below-average and above-average IQ scores among

individuals on correct responses in completing a task based on three methods of instruction. So the researcher decides to select participants with IQ scores between ± one standard deviation from the mean.

Another way to account for extraneous variance is to *build the extraneous variable into the design.* Prior evidence suggests that gender may confound the results of interventions on math achievement. Gender can be included as an independent variable in a design where its main and interaction effects can be assessed with the primary independent variable of interest.

The effects of extraneous variance also can be *statistically controlled* using techniques such as analysis of covariance and multiple regression. The extraneous variable needs to be measured and then the dependent variable is adjusted to account for its influence.

Matching participants across condition groups based on one or more extraneous variables can be used to rule out or hold constant their confounding effects. A researcher might match participants in a treatment and control condition based on gender, height, and age. A woman 5′ 0″ to 5′ 2″ in height who is between the ages of 20 and 25 years is matched to participant with similar characteristics in the other group. All participants would be matched on homogeneous characteristics. Sophisticated propensity score matching methods have been refined in recent years, which have enhanced the rigor of quasi-experimental designs (see Cook & Steiner, 2010).

The most important way to control extraneous variance is using *random assignment.* Each participant is assigned to experimental and control group conditions totally by chance. Tables of random numbers or computerized random number generators are used to accomplish random assignment. Random assignment makes groups randomly similar to each other. It is less likely that one group of participants will have more or less extraneous variance than another group when groups are randomly assigned from the same sample. Shadish, Cook, and Campbell (2002) reported the benefits of random assignment to generalized causal inference as: "equating groups before treatment begins, by making alternative explanations implausible, by creating error terms that are uncorrelated with treatment variables, and by allowing valid estimates of error terms" (p. 252).

Another method of controlling extraneous variables is to use *blinded procedures.* Participants are not informed about the nature or purpose of a study in a *single-blind procedure* to minimize participants behaving in reaction to the purpose of the study. In a *double-blind procedure* neither the participants nor the

researchers are informed of the treatment conditions that the participants receive. This procedure helps minimize experimenter expectancy effects and participant reaction (demand) effects. This procedure usually requires that other people select groups, administer treatment, and record results. In a *triple-blind procedure* neither the participants nor the persons administering the treatment nor those evaluating the response to treatment know which participants are receiving a given treatment. A *partial-blind procedure* is used when the researcher finds out which treatment conditions are administered just before administering the treatment conditions.

Deception can be used to shift the participants' attention away from the actual purpose of a study so they do not react to the perceived purpose in a certain way. The use of deception requires careful consideration of ethical standards and approval from an Institutional Review Board (IRB). There are several other procedures to control for extraneous variables (see Kirk, 1995; Shadish, Cook, & Campbell, 2002).

INTERNAL VALIDITY AND EXPERIMENTAL DESIGNS

For over 50 years, Campbell (1957); Campbell and Stanley (1963); Cook and Campbell (1979); Shadish, Cook, and Campbell (2002); Shadish and Cook (2009); and Cook and Steiner (2010) have developed sophisticated methods in experimental and quasi-experimental research. Two of many experimental methodological areas that have evolved are internal validity and research designs.

Internal Validity

Determining *internal validity* is assessing whether the A (cause) → B (outcome) relationship is indeed causal based on the manipulations and measures of the variables used in a study. This assessment process includes knowing which variable is the cause and which is the effect, which is referred to as *ambiguous temporal precedence* (Shadish, Cook, and Campbell, 2002).

An important part of assessing for internal validity is ruling out threats to internal validity. Thirteen threats to internal validity are presented in Table 4.1. The first eight threats (*THIS MESS*) can be assessed and ruled out through the use of experimental designs, and the last five threats (*DREAD*) are corrected using

TABLE 4.1 Threats to Internal Validity: THIS MESS DREAD

Testing effect	The treatment effect may be confounded when changes in posttest scores of participants are influenced by their experience from taking a pretest.
History	An external but concurrent event to the experiment (e.g., a hurricane) may affect scores on a dependent measure that are unrelated to the treatment.
Instrumentation	Pretest and posttest scores may change because of a faulty measurement instrument, irrespective of the treatment.
Selection (differential)	Differences in characteristics of participants assigned (usually not random) to treatment conditions may confound attributing the changes in the dependent variable to the treatment.
Maturation	Psychological and physical changes within participants may occur in an experiment, especially over time. These participant maturational changes may have an extraneous effect on the dependent measure.
Experimental mortality (attrition)	The loss of participants to treatment or measurement of the dependent variable can produce unbalanced attribution of the effects of the treatment on the dependent variable.
Statistical regression	This is the phenomenon that extremely high or low group scores on a variable tend to regress to the mean (get lower or higher) on a second measurement of the variable, confounding the treatment effect.
Selection-maturation Interaction	This is a combination of two threats to internal validity. For example, some participants in one assigned treatment condition group may have matured in math self-efficacy more than participants in another treatment condition group, and the purpose of the study is to increase math achievement. Additionally, other combinations of threats could interact to confound the treatment effect.
Diffusion of experimental effect	The treatment may diffuse to the control group over time because the control group may seek access to the more desirable treatment.
Rivalry (compensatory)	The control group participants may perform beyond their usual level because they perceive that they are in competition with the experimental group.

Equalization of treatments (compensatory)	A treatment group may receive experimental rewards that appear more desirable than those received by the control group participants. Efforts are made by individuals outside of the experiment to compensate the control group participants with similar desirable goods. This would obscure the results of the treatment.
Ambiguous temporal precedence	A lack of clarity is provided by the researchers as to which variable occurred first, leading to a question of which variable is the cause and which is the effect.
Demoralization (resentfulness)	Lower performance of control group participants on the dependent measures may result from their belief that the treatment group is receiving a desirable treatment.

Source: Reprinted by permission from W. E. Martin Jr. & K. D. Bridgmon. (2009). Essential elements of experimental and quasi-experimental research. In S. Lapan & M. Quartaroli (Eds.), *Research essentials: An introduction to designs and practices* (pp. 35–58). San Francisco, CA: Jossey-Bass.

experimental procedures. Next, we will discuss how experimental designs are used to rule out threats to internal validity.

Experimental Designs

There are many *experimental designs*, *quasi-experimental designs*, and *correlation designs* available for use by researchers to execute their studies. Two fundamental and commonly used experimental designs are discussed first. Explanations are provided as to how the designs reduce the effects of THIS MESS. Examples of statistical designs that can be used with the designs are identified. The other five threats to internal validity (DREAD) are also discussed.

A few rules and symbols are used to describe experimental designs. Each row represents a group of participants receiving a different experimental condition. An *R* designates that the participants were randomly assigned to an experimental condition (treatment or control). All *O*'s (observations) represent assessments of one or more dependent variables. An *O* before an experimental condition is a pre-experimental condition assessment, and an *O* after the experimental condition is a postexperimental condition assessment. An *X* stands for a treatment condition and *C* symbolizes a control condition.

Randomized Multiple Treatments and Control with Posttest-Only Design

R	X_1	O
R	X_2	O
R	C	O

This design is used in the one-way analysis of variance (ANOVA) example in Chapter 7. The purpose of the study is to compare the effects of psychotherapeutic treatment (independent variable) on changes in depressive symptoms (dependent variable). Psychotherapeutic treatment has three experimental conditions: (1) cognitive-behavioral therapy (CBT), (2) interpersonal therapy (IPT), and (3) control group. Each line in the diagram represents one of the three groups. The R symbolizes random assignment to each group condition, which represents a study sample of participants being randomly assigned equally to each condition. The conditions by groups are represented by X_1 (CBT), X_2 (IPT), and C (Control). The O (observation) is the posttest following eight weeks of treatment implementation using the Center for Epidemiological Studies Depression Scale (CES-D) to measure differential change in depressive symptoms resulting from the different treatment conditions.

This *randomized multiple treatments and control with posttest-only design* strongly benefits from the essential elements of experimental research: (1) random assignment, (2) control groups, and (3) a manipulated independent variable. The design does reduce the threats of THIS MES but a concern arises with the last S (interaction of two threats). For example, more participants may be lost (attrition) from one or more randomly assigned groups (selection) during the course of an experiment. This affects the expected outcome of random assignment to make groups randomly similar to each other at the onset of the study. If there were a pretest in the design, information could be derived to help explain how the participants dropping out of the experiment affected changes in the dependent variable. The interaction effect of attrition and selection would not be an issue if there were no loss of participants using this design. Without a pretest, it is impossible to know the causal effects of the independent variable on the dependent variable because of the disproportionate loss of participants from one condition compared to the other condition.

A one-way ANOVA can be used in this design to compare the means of CES-D scores across the three psychotherapeutic conditions for significant differences. A post hoc analysis can be used to assess the differences between the three pairs of means.

Randomized Multiple Treatments and Control with Pretest and Posttest Design

R	O	X_1	O
R	O	X_2	O
R	O	C	O

The *randomized multiple treatments and control with pretest and posttest design* adds a pretest to the previous design. So, the participants could be measured on their depressive symptoms using an equivalent form of the CES-D before administering the experimental conditions. This reduces the concern about understanding the effects of interaction between attrition and selection. The pretest scores on depressive symptoms of people lost during the research process can be used to assist in controlling differences in participants by group on the posttest scores. The design helps control for all of the threats associated with THIS MESS.

A statistical design that could be used with this design is a 3 × 2 repeated-measures ANOVA. The 3 represents the independent variable (factor) of *psychotherapeutic treatment* with three conditions of CBT, IPT, and control. The 2 represents a second independent variable of *time* with two conditions (pretest and posttest). This statistical design allows for a comparison of the differences among the means of CES-D scores across the three experimental conditions. A post hoc analysis can be used to assess the differences between the three pairs of means.

Also, differences between the means at pretest and posttest can be analyzed. Finally, the interaction effects of psychotherapeutic treatment and time can be compared to see if the CES-D average scores differ at conditions of psychotherapeutic treatment when compared to pretest and posttest conditions. The interaction effect analysis could be followed by a simple effects analysis to determine which levels of the first independent variable interact with which levels of the second independent variable to produce significant effects on the dependent variable.

Experimental Research Procedures

Effective experimental designs are used to reduce the threats to internal validity of THIS MESS. Seven recommended *experimental research procedures* that aid to reduce the threats of DREAD (see Table 4.1) are identified next.

1. Use different persons to implement each treatment condition.
2. Arrange treatment conditions so that contact between experimental and control group participants is minimized.
3. Use single-blind, double-blind, or triple-blind procedures to remove participant or experimenter bias.
4. Debrief participants using interviews to assess their experiences and expectations of the treatment conditions.
5. Consider not using experimental rewards.
6. Clearly define treatment conditions.
7. Use past research and theory to guide the evidence of an A (independent variable) → B (dependent variable) causal relationship.

Quasi-Experimental Designs

The purpose of using quasi-experimental designs is to discover causal relationships similar to using experimental designs. However, quasi-experimental designs are considered compromise designs because random assignment is not used. It is important for researchers using quasi-experimental designs to provide carefully detailed, logical arguments in ruling out alternative explanations (*rival hypotheses*) to account for their findings of a causal relationship.

Quasi-experimental designs are improved by (1) adding control or comparison groups, (2) adding pretests and posttests, (3) removing and reinstituting treatments, (4) adding replications, (5) reversing treatment, and (6) case matching (Shadish, Cook, & Campbell, 2002). Various statistical procedures also have been developed that have improved quasi-experimental designs. *Propensity scores* are used in case matching to equate individual cases from different populations to

reduce selection bias. Propensity scores are a weighted composite of measured covariates that correlate with treatment (Cook & Steiner, 2010; Rosenbaum & Rubin, 1983). *Regression discontinuity designs* use cutoff scores to assign participants to conditions. The cutoff score is a measure obtained before treatment is undertaken. Then, posttest scores are analyzed with ordinary least squares regression using the cutoff scores (Shadish, Cook, & Campbell, 2002). Two commonly used quasi-experimental designs are presented next.

Repeated-Treatment Design with One Group

$$O_{\text{Baseline}} \qquad X_1 O_1 \quad X_{2\text{Removed}} O_2 \quad X_3 O_3 \quad X_4 O_4$$

The focus of the research example of *repeated-treatment design with one group* is to assess whether the use of friends (support partners) in behavioral weight control treatment (BWCT) can improve the weight loss outcomes of persons who are overweight (this example is used in Chapter 8). A randomly selected group of participants who are overweight are weighed at baseline before treatment (O_{Baseline}). Then, each participant's weight loss is measured and recorded after three months of treatment with partner support (X_1 O_1 at 3 months), after three months when partner support is removed ($X_{2\text{Removed}}$ O_2 at 6 months), after three months when partner support is added back to treatment (X_3 O_3 at 9 months), and after the treatment with partner support is continued for three months (X_4 O_4 at 12 months).

The most important result of this design to demonstrate some degree of causal relationship is measuring the change of weight loss when the treatment is removed and reinstated. External events (history) and physical/psychological changes affecting participants may confound the understanding of weight loss as it relates to the treatment. It would be difficult to tease out confounding without using a control group and random assignment. Moreover, participants' awareness of the cycle of treatment, treatment removal, and observations might behave uncharacteristically in reaction to the study cycle. Since the same participants are measured repeatedly over time under various conditions, a repeated-measures ANOVA would be an appropriate statistic to use in this study. The changes in mean weight and covariances at baseline, three months, six months, and 12 months would be assessed for differences.

Nonequivalent No-Treatment Control Group Time-Series Design

NR	O_1	O_2	O_3	*X*	O_4	O_5	O_6
NR	O_1	O_2	O_3	*C*	O_4	O_5	O_6

There is a control group in the *nonequivalent no-treatment control group time-series design* and no removal of treatment. The *NR* stands for nonrandom assignment. This design controls for most of the threats to internal validity, THIS MES. However, an interaction (S) between selection and history may be a concern. It is possible that one group of participants is affected differently by external events (history) that influence their scores more on a dependent variable than the other group. The control group and multiple pretests and posttests in this design provide valuable information to rule out threats to internal validity. Visual analyses of graphs are useful for understanding the results. Other statistics that can be used include repeated-measures ANOVA, randomization tests, and bootstrapping methods.

Correlational Research Methods

Correlations methods are used in experimental and quasi-experimental designs, but the primary purpose of correlation research is to explore *bivariate relationships*, *multiple relationships*, and predictions among variables. Variables in correlational research are often referred to as *X* and *Y* variables. The *X* variable is called either an independent or a *predictor variable*, and the *Y* variable is identified as either a dependent or a *criterion variable*. A bivariate correlation (r) assesses the relationship between two variables, and a multiple correlation (R) assesses the relationships among more than two variables. A *simple linear regression* uses a bivariate correlation to predict the dependent variable (Y) from the independent variable (X). A *multiple linear regression* uses a multiple correlation to predict a dependent variable (Y) from two or more independent variables (Xs).

In Chapter 12, an example is presented about a researcher who assessed whether lower interest in scientific activities is a better predictor of higher dissertation stress reported by counseling and clinical psychology doctoral students when compared to their interests in practitioner activities.

The first independent variable in the study is doctoral students' interest in scientific activities as measured by the Scientist Scale of the Scientist Practitioner

Inventory (SPI). The Scientist Scale measures student interests in activities related to research, statistics and design, teaching/guiding/editing, and academic ideas.

The second independent variable used in the study is students' interest in practitioner activities measured using the Practitioner Scale of the SPI. The Practitioner Scale assesses interests in activities involving therapy, clinical expertise and consultation, and testing and interpretation. Higher scores on the Practitioner Scale reflect higher interests.

First, we consider only one independent variable (scientific activities) and its prediction of dissertation stress. An analysis of the extent that students' interest in participating in scientific activities (X) relates and predicts dissertation stress (Y) could be answered using a bivariate correlation and simple linear regression model. The model could be written symbolically as $Yf(X)$, where Y (dissertation stress) is a function of X (scientific interests). A prediction model can be written as $Y' = A + B_1X_1$, where Y' is the predicted value on the dependent variable (dissertation stress), A is the intercept, the value of Y when all the X values are zero, B is the unstandardized regression coefficient assigned to each X value, and X is the measured value of the independent variable (scientific interests).

A multiple correlation and multiple regression analysis can be used to analyze the two independent variables with the dependent variable. It can be written symbolically as $Yf(X_1, X_2)$. The prediction model would be $Y' = A + B_1X_1 + B_2X_2$.

CHOOSING A STATISTIC TO USE FOR AN ANALYSIS

The consideration of five issues can be useful in deciding on a statistic to use in a particular analysis: (1) the focus of the interplay among variables (e.g., relationships or differences), (2) the number of independent and dependent variables used in an analysis, (3) the scales of measurements of the dependent variables, (4) the number and relationships (dependent vs. independent) of participant groups being compared, and (5) the extent that underlying assumptions of the statistic are met. These issues are illustrated in relation to statistics covered in this book (also see Figure 4.1).

One-Way Analysis of Variance

1. *Focus of analysis.* The purpose of a one-way ANOVA is to compare mean differences of a dependent variable among groups.