Conducting Research in Technical Communication: The Application of True Experimental Designs

Jan H. Spyridakis
University of Washington

SUMMARY
An increasing number of technical communicators are making document or screen-design decisions that they would like to test empirically. However, many have not been trained in research design and statistics. This paper explains the use of true experimental designs in technical communication research. Major steps in the research process are discussed in detail:

- Identifying and defining a problem
- Reviewing relevant literature
- Formulating hypotheses
- Defining variables
- Constructing a research design
- Selecting subjects
- Creating experimental materials
- Collecting and analyzing data, and
- Arriving at conclusions based on the empirical work.

From this discussion, readers should be able to read research studies critically and, if desired, design empirical studies after further reading in texts on research design and statistics.

Contemporary research methods in technical communication have been influenced by a broad range of disciplines, theories, and practices, from linguistics, to learning theory, to educational practice, to reading comprehension, to human factors, to cognitive and physiological psychology, and so on. Regardless of the discipline, the question remains: Exactly what is research? Research is a systematic approach to provide answers to questions, answers that may be abstract and general, as is often the case in basic research, or concrete and specific, as is often the case in applied research.

Focusing on the conduct of basic research in technical communication, this paper seeks to:

1. Provide a starting point for beginning researchers
2. Offer some reminders to those already conducting basic research
3. Furnish a foundation to those who must evaluate the quality or validity of research—either

This article has been peer-reviewed.

Technical Communication, Fourth Quarter 1992
as general consumers of research or as professionals who hire researchers to conduct studies.

Basic research is generally concerned with identifying and understanding relationships among variables. In attempting to explain some part of the events of our world, basic research endeavors to develop knowledge consistent with or leading to theories. The development of comprehensive theories that explain past or present events allows us to formulate hypotheses about future events. And then the goal of subsequent research becomes that of testing these new hypotheses, which in turn continues the cyclical process of theory development leading to further hypotheses.

While basic research often has little applicability to specific situations beyond the immediate research area, the results of basic research often lead to the development of a product—a textbook, a teaching approach, a software program. In contrast, applied research, relying on the theories and findings of basic research, uses systematic evaluation to find immediate answers to problems or issues surrounding documents, screen designs, products, methods, or techniques.

Research types are also differentiated on the basis of their methods. Although discussions of research methods classes use different classification schemes [1-3], two commonly cited categories in the technical communication field are qualitative and quantitative. Other articles in this special issue discuss qualitative methods; this article examines the experimental research methods and designs used in basic research.

In experimental research, researchers carefully manipulate variables in controlled environments (perhaps laboratories) with the goal of testing hypotheses about group differences and examining cause and effect relationships. Conclusions about cause and effect relationships require true experimental designs that allow attribution of effects to the manipulated variables. In general, the adjective true signifies the random assignment of subjects to experimental conditions as well as the use of other controls.

The process of basic research and the application of true experimental designs is similar across disciplines. Although this process might appear to be linear, it is often iterative, particularly in new research arenas. Briefly the process involves the following steps:

1. Identifying a research problem
2. Reviewing relevant literature
3. Constructing hypotheses
4. Defining independent and dependent variables
5. Ensuring validity and reliability
6. Defining the subject population and selecting a sample
7. Creating a research design to examine the problem
8. Collecting and analyzing data
9. Drawing conclusions about relationships among variables and suggesting the direction of future research.

This paper discusses the research process in this order, except that complete research designs are discussed last so that the terminology associated with them can be easily understood in the context of earlier sections.

Choosing A Problem and Reviewing the Literature

In pursuing a problem of interest, researchers must examine the relevant literature, evaluate whether the problem or topic merits further exploration, and determine what aspects should be investigated. Before researchers wed themselves to a problem, they should consider whether it meets certain criteria. A problem should have value in theory and application and be workable, given time, money, and personnel constraints. Its potential answer should have sufficient magnitude to lead to advancements in the field. Finally, the methodology necessary for examining the problem should be accomplishable and an adequate pool of potential subjects should be available.

The relationship of the problem statement, the literature review, and the hypotheses readily becomes apparent if one examines the introductions and literature reviews of empirical studies. Most introductions to empirical work state a topic area to be examined and quickly explain the relevance of the topic to the given audience. For example, if an experiment examined the effect of topc sentence placement on reading comprehension, the researcher, depending on his or her interests and audience, could relate the topic to document design decisions or even writing instruction.

Problem Statement

Introductions to empirical studies usually contain a problem statement, clarifying what variables are of
interest. Problem statements describe the purpose of the study. They ask about the relationship among variables. They should be testable by empirical methods and should not represent the researcher’s expectations for the study’s results. While not all studies present a concrete problem statement, the “ideal” problem statement is expressed as either a declarative or interrogative sentence that specifies four variables:

- Independent variables (factors to be manipulated)
- Moderator variables (subject characteristics to be compared)
- Dependent variables (constructs to be measured)
- Control variables (conditions to be held constant)

Correctly worded problem statements ask about the effect of the independent and moderator variables on the dependent variables in the context of control variables (see Figure 1). Some researchers announce the problem statement in the introduction in only a general way, offering a more traditional purpose statement, and hold the specific problem statement for the end of the literature review.

Note that the problem statements do not state a directional bias and thus do not hint at the expected answer to the problem. A statement such as “This study investigates whether text difficulty reduces the comprehension of college readers” would be inappropriate if it conveys the researcher’s expectations. Such directionality should be saved for the hypotheses.

**Review of Relevant Literature**

The goal of reviewing relevant literature should be to determine what advances have been made in a specific area, what limitations exist with previous research, and what theories suggest that a given area needs to be examined. In examining prior studies, a researcher should focus on whether the research designs—from subject selection and assignment to conditions, to materials and test design, to the actual administration procedures—were valid and reliable and whether the researchers have drawn appropriate conclusions [4-5].

When validity is lacking, careful readers can conclude that a previous researcher has misinterpreted the reason for certain effects or overinterpreted the applicability of the findings, perhaps because the design was insufficient to test stated hypotheses, or perhaps because the study controls were too loose.

When reliability is lacking, one can conclude that the experimental instruments (tests) would not obtain similar results in the future.

Two approaches for reporting the results of the literature search are common. One reviews previous empirical studies, focusing on the subjects, variables, materials, and overall procedures that others have employed. Researchers who use this approach often identify new areas to examine or areas in need of reexamination. For example, in an empirical study of signaling, the researcher reviewed previous studies that examined the effect of headings, previews, and logical connectives on reading comprehension [6-10]. The review revealed that previous studies had used only short, easy texts of familiar content to assess the reading comprehension of young readers. This literature review led to new hypotheses for the effects of signals on the reading comprehension of competent college-age readers who read texts differing in length, difficulty, and familiarity.

A second common approach to reviewing the literature examines competing theories or hypotheses on a topic and then arrives at a research design to determine which theory or hypothesis is correct. For example, many studies of comprehension of abstract versus concrete words seek to distinguish between two hypotheses:

1. The dual-encoding hypothesis, which suggests that readers have both visual and verbal traces for concrete words, and
2. The single-encoding hypothesis, which suggests that readers use only a verbal trace to encode both word types.

---

**Figure 1. Problem Statements in Two Forms**

**Declarative:** This paper investigates the effect of text difficulty on comprehension, learning, and reader preference.

- **Independent variable:** college age readers
- **Mediator variable:** reading ability
- **Control variable:** college age readers

**Interrogative:** What effect do text difficulty have on comprehension, learning, and reader preference?

- **Independent variable:** college age readers
- **Mediator variable:** reading ability
- **Control variable:** college age readers
While the literature review in such cases certainly reviews the relevant studies, it focuses on the competing hypotheses.

**Constructing Hypotheses**

Regardless of the approach used is the literature review, the goal should be to develop the logic that will support the hypotheses and ensuing research design. Usually found at the end of the literature review, a well-constructed hypothesis should—
1. Speculate on the relationship of two or more variables
2. Be stated clearly and unambiguously in the form of a declarative sentence
3. Be testable.

Unlike problem statements, hypotheses state the expected answers to be found in the study. They are more detailed than problem statements in that they usually name the levels of the independent, moderator, and dependent variables under investigation. The levels of the independent and moderator variables represent the various conditions to be studied. For example, the independent variable of heading phrasing could have two levels: interrogative form and declarative form.

Given this view of headings, one might state the following directional hypothesis: Headings in interrogative form will improve recall with low-ability readers more than headings in declarative form. Alternatively, one could state this even more concretely: Low-ability readers will score higher on free recall tests after reading passages with interrogative heads than after reading passages with declarative heads.

A nondirectional hypothesis might state that both reading-ability groups will score equally well on a free recall test with declarative heads. The decision to make hypotheses directional or nondirectional stems from what the literature search suggests will occur in the planned study.

**DEFINING AND OPERATIONALIZING VARIABLES**

At this point or in the methods section that follows, the researcher must operationalize the variables, that is, clearly describe how the variables and their levels are defined and how they will be brought into action.

In experimental studies, investigators are interested in determining how independent and moderator variables affect dependent variables given certain control variables. The way in which these variables are operationalized and measured influences the statistics used.

Operationalizing variables consists of deciding how the existence of a concept or quality is to be identified or measured. For example:
- A researcher examining the independent variable of paragraph coherence must be concrete about what manipulations represent changes in coherence: e.g., the addition or deletion of a fixed number of conjunctions, conjunctive adverbs, or pronouns.
- A researcher examining the independent variable of reading ability must define how reading ability is to be determined—through a test the researcher constructs and describes, or perhaps through a standardized reading inventory.
- A researcher wanting to measure the dependent variable of reading speed must define speed, the units of speed, and the actual measurement method.

A researcher must be precise about how the existence of a stated variable is to be perceived—and how that variable is to be assessed. The tasks of defining and operationalizing variables are further addressed here while measurement and data-analysis issues are discussed later.

**Independent and Moderator Variables**

Researchers want to assess the effects of independent and moderator variables, by either manipulating the independent variables or selecting different levels of the moderator variables. The distinction between these two types of variables is that independent variables are directly manipulated by the researcher, while moderator variables, which are actually a type of independent variable, represent subject characteristics that are present before the experiment (e.g., age, gender, ability, expertise). Many researchers never use the term moderator variables and simply view them as independent variables.

A researcher’s interest in independent and moderator variables arises not only from the identification of the variables but from the identification of the levels of those variables—the different categories or subsets of the variables. Logically one does not have a variable (something that can vary) unless it has at least two levels. The researcher must decide what aspects and levels of the variable to investigate in a given study. For example, a researcher conducting a comprehension study could assess the effects of three variables with two levels each: headings (pres-
One does not have to examine all possible levels of an independent or moderator variable; however, the decision not to test all possible levels may affect the chosen research design and statistical analyses.

The levels of moderator and independent variables are used to group subjects and to create experimental conditions. Imagine a study on the effect of computer interaction style (independent variable) and subjects' computer expertise (moderator variable) on task completion (dependent variable) in a database program. Let's also imagine that the researcher decides to investigate only two levels of the independent variable (pull-down menus and commands) and two levels of the moderator variable (experts and novices).

A typical factorial experiment based on these variables, where the total number of treatment conditions is derived by multiplying the number of levels of all independent and moderator variables, would result in four testable and mutually exclusive conditions (2 levels of interaction style × 2 levels of expertise = 4 conditions):
1. Experts using pull-down menus
2. Novices using pull-down menus
3. Experts using commands

One does not have to examine all possible levels of an independent or moderator variable; however, the decision not to test all possible levels may affect the chosen research design and statistical analyses. In a factorial experiment, too many variables or levels per variable can create so many conditions or subject groups that the study becomes difficult to execute and the results become almost impossible to interpret. Consider a factorial experiment of four independent (or moderator) variables, two with two levels each and two with four levels; it would result in 64 conditions to test (2 × 2 × 4 × 4 = 64). To avoid obtaining results and interactions that are virtually unintelligible, researchers often decide to conduct subsequent studies to assess other related variables.

**Dependent Variables**

Dependent variables represent the concepts researchers seek to measure; in experimental studies they are the variables one expects to be influenced because of manipulations of the independent variables in the context of the moderator variables. Dependent variables typically represent such concepts as comprehension, efficiency, discriminability, etc. Dependent variables may have multiple measures. For example, comprehension, a dependent variable, could be measured through factual and inferential multiple-choice tests, free or cued recall tests, and problem-solving tests. Efficiency could be measured through the number of tasks completed, the number of screens accessed, the number of commands used, or even speed. These measures could also have various levels; for example, in the multiple-choice test, subscores could be calculated for questions assessing subordinate versus superordinate information.

**Control Variables**

The term control variables is actually oxymoronic in that control variables do not vary—their variability is in fact what is controlled. They are in reality potential independent variables that the researcher consciously decides to neutralize so that they cannot influence the dependent measures. Typical concepts that are controlled are prose length, computer platform, reading ability, user expertise, age, and so on. If a researcher chooses to control some of these potential variables, he or she does not assess various levels of these, but, instead selects one level and holds it constant across all conditions. Thus, the researcher would determine a fixed length for all prose to be examined; use a certain computer; or obtain subjects with similar reading ability, computer expertise, or age. In contrast, if the researcher wants to examine different levels of these control variables, they would become independent or moderator variables.

**CONTROLLING THREATS TO VALIDITY**

Before constructing a true experimental design, one must consider the concepts of internal and external validity. Internal validity refers to whether the results of a study represent what the researcher purports to be measuring. External validity refers to whether an experimental design and its results are generalizable to other situations in the real world.
Researchers want to ensure that results occur because of experimental treatments as opposed to uncontrollable factors and that the implications of the results will apply beyond the narrow confines of a specific study.

While no experiment is ever completely free from threats to validity, careful researchers can greatly reduce the threats and at least be aware of what threats may exist and acknowledge them in their conclusions. Some of the most common threats to internal and external validity are summarized here. (For further information, see Campbell and Stanley [11].

Internal Validity

Eight threats to internal validity are discussed here.

Subject History. All events that occur in the study environment, except for the experimental conditions, should occur for all subjects. To avoid the threat of subject history, researchers use control groups, where all subject groups have the same experience except for the experimental manipulations, which of course differ. Further, if subjects are tested separately, a researcher must ensure that all subjects have the same experience, except for the planned treatment. Instructions should be printed or read from a script, and ideally the same test administrator should interact with all subjects.

Subject Selection. To ensure that the selection and assignment of subjects to conditions does not result in biased groups, researchers should identify one subject pool that contains subjects with similar backgrounds and then select all experimental subjects from that pool. If moderator variables are used, one subject pool for each level of the variable is needed, but researchers must then randomly assign subjects to the different experimental conditions.

If a researcher were to assess two versions of a computer manual and select one group of subjects from the technical division of a company to interact with manual A, and another group of subjects from the R & D division to interact with manual B, he or she would encounter the threat of subject selection. While there are two groups of subjects looking at two different versions of a manual, the two groups represent different populations: They have different backgrounds, knowledge sets, education, work environments, and work managers. This researcher should have selected one subject pool to represent the desired population and then randomly assigned the subjects to manual A or manual B.

Subject Maturation. If an experiment is conducted over time, there is a danger that subjects may have differing experiences outside the confines of the experiment and mature or change between experimental sessions. This threat is of particular concern with longitudinal studies. Again, the random assignment of subjects to conditions should remove this threat. The logic behind random assignment is quite clear in this case: If there is an effect on subjects caused by changes over time, and if subjects are randomly assigned to conditions, then that effect can be assumed to be equally distributed across conditions. The researcher can correctly conclude that group differences are due to treatments.

Testing. In many experiments, researchers want to gather information on subjects prior to the actual experimental conditions. This is often done with a pretest; however, a threat arises if the pretest has an effect on the posttest that is not known to the researcher. For example, many researchers measure subjects' familiarity with a topic before the subjects read a prose passage on the topic. It is possible that the pretest itself could easily prime subjects' memory for the information and enhance their performance on the posttest [12].

Whenever possible, unobtrusive measures should be used to obtain pretest data. When direct pretests are used, for whatever reason, researchers should use appropriate statistics to assess or neutralize their effect. Using control groups also helps reduce this threat; if both the control and experimental groups are pretested, any differences between groups can be attributed to the experimental treatments.

Instrumentation. Threats from instrumentation occur whenever experimental instruments vary across conditions. Those who have conducted experiments involving computers should well understand this threat: Software can and often does malfunction and thus provide subjects with unanticipated, different experiences. When such events occur, researchers cannot determine whether differences between treatment groups are due to the planned treatments or the instrument's variation.

This threat can also occur when multiple test administrators, observers, or raters administer conditions or evaluate test performance. When different people help conduct an experiment, they should be trained and their inter-rater reliability assessed. If, for example, five graduate students help score recall tests, the researcher should train them in scoring methods and check to see that their scoring methods are similar. The researcher could also use blind raters...
by not telling the rates about the differences in conditions, thus preventing any bias in scoring different conditions. Other reliability checks are available to ensure that the test itself functions consistently across subjects and conditions [13].

**Statistical Regression.** To avoid the threat of scores regressing to the mean, researchers should avoid selecting subjects on the basis of extreme scores on pre-measures. If subjects are chosen on the basis of extreme scores on same pre-measure, they will never remain as extreme on future tests. This is simply the nature of chance.

**Experimental Mortality.** Any study that occurs over a period of time inevitably loses subjects after the initial experimental session[s]. If a researcher uses a large initial sample (and perhaps attempts to locate dropouts to determine whether they share some characteristic that could affect the results), he or she will be less concerned that experimental mortality has created a bias in the remaining sample.

**Expectancy.** Subjects' or experimenters' expectations about experimental conditions can create results in and of themselves. With subjects, demand characteristics—subjects' performing differently from normal because they believe certain responses are expected—can create this threat from expectancy. Additionally, if experimenters know which conditions are being administered, they can unknowingly influence subjects' responses. To avoid this threat, some researchers use a double-blind experiment, where neither the test administrators nor subjects know what the conditions are.

The threat to expectancy can easily occur when subjects are given too much information in the introduction to a study. In many reading comprehension studies, for example, subjects are told prior to the experimental tasks that after reading they will take a recall test and write down everything they can remember. It is difficult, in such cases, to interpret whether significant effects are caused by the treatment or the instructions.

**External Validity** Researchers must also be concerned with external validity. That is, they must determine whether results would be generalizable to other situations in the real world. For example, if a researcher finds that subjects with more education recall more information after reading than subjects with less education, he or she must consider whether this relationship is true only for readers who are nonnative English speakers, enrolled in the second week of a beginning English course, living in the Pacific Northwest, and reading extremely long technical documents on quarks. Such a finding would certainly have limited external validity.

One commonly described threat to external validity is the Hawthorne effect, where all subjects perform differently than normal simply because they are being observed in an experiment [14]. The belief is that the presence of experimental conditions can restrict generalizability to only a laboratory setting.

**Other Controls to Ensure Validity** Beyond the multitude of precautions already discussed, a few other precautions can ensure solid designs, and interpretable and generalizable results. Expectancy should be constant for all subjects except for the experimental treatment, from the instructions subjects receive, to the time subjects spend on tasks, to the screen or document quality that subjects read from.

If subjects perform more than one task or take more than one test, researchers need to control for the effect of order. They can randomly order the tasks or tests, or counterbalance the order within conditions so that the effect of initial tasks or tests on subsequent tasks or tests is removed or controlled for. If the order is counterbalanced within conditions (half the subjects receive task 1 before task 2, and half receive task 2 before task 3), the researcher might decide to treat order as another independent variable and measure its effect. As will become more apparent in the next few sections, the methods of selecting subjects and designing materials and tests can also have a powerful effect on validity.

**USING HUMAN SUBJECTS** Researchers need to use a careful approach in selecting the relevant population and sample, assigning subjects to conditions, and using control groups.

**Selecting and Assigning Subjects to Conditions** Ideally, in selecting subjects for an experiment, researchers must attempt to follow the principle of randomness. The first task is to identify the population. If the population a researcher wants to generalize to is very small (for example, a 25-person R & D group in a given company), the researcher would probably study the whole population. Usually,
though, the populations that researchers want to generalize to are much larger—adult Americans who read and use technical information at work, experienced adult computer users, etc.

While true random sampling of the designated population is ideal, it is often not possible simply because of limited resources. A researcher wanting to randomly sample adult, American English-speaking readers would theoretically need to draw a sample from across the country. Instead he or she will probably identify a more available sampling pool that represents the population in all critical ways (e.g., education, prior task experience, topic familiarity), and then randomly select subjects for the study from that pool. (However, this discussion is not advocating simply accessing the most available or first sample one bumps into [151].)

An ongoing argument regarding sample selection concerns the use of college students when a researcher wants to generalize to a broader population. Certainly, tradeoffs exist in such situations. While researchers may appear to use college students because they are ready, willing, and able to participate in experiments, it may be that these researchers have also examined the reality of the world. With many studies, it is virtually impossible to identify one population of professionals on the job who are sufficiently homogeneous or critical variables (e.g., computer experience, writing ability) yet still have the time or desire to participate in an experiment.

While one wants to sample from a heterogeneous pool of subjects, one needs subjects who are not homogeneous on variables that could influence and perhaps confound the study result. Further, to avoid self-selection bias, one wants to access a population where all or most subjects selected are likely to participate.

In some environments, random selection is further complicated by an institution's human subjects requirements, often enforced by federal agencies. These requirements, established to protect subjects from unethical researchers, necessitate the use of informed consent, competent subjects over 18 years of age or parental permission, voluntary participation, the right to withdraw from a study, etc. While a researcher can obtain potential subjects in multiple sections of classes at the university for which he or she works, forcing them to participate is a study is illegal.

However, effective, persuasive approaches can encourage subject participation—from offering financial incentives, to giving extra credit, to requiring class participation (but usually there must be alternative methods of meeting the class requirement). In any of these scenarios, a slight selection bias can occur in the sample in that those who agree or refuse to participate, accept or ignore the monetary or extra credit incentives, or meet the class requirements by participating in the study or by fulfilling an alternative requirement may represent some subset of the sample. Careful planning, humanistic introductions of the experiment to the subjects, and sensitive researchers can help ensure an unbiased group of subjects.

While random selection can be difficult, as discussed above, random assignment of subjects to treatment groups is either simple. For example, randomly ordering six versions of a passage and then handing one version to each subject would ensure random assignment. In using such a method one also avoids using intact groups in which preexisting subject groups (e.g., classes, employee divisions) are used as separate treatment groups. The use of intact groups and the resulting adjustment to the data analysis leads to a considerable loss of power in examining group differences.

In general, fewer than 10 subjects per condition is risky; the performance of this few subjects can represent considerable variation and mask differences among groups. When the variance were reduced. The effect and size of such variance decreases as sample size increases.

When moderator variables are used, researchers need multiple sampling pools from the desired population with one pool per level of the moderator variable. Then subjects who represent each level of the variable should be randomly assigned to all other conditions. For example, a researcher wanting to test both expert and novice computer users on two types of online help would need to take these steps:
1. Identify the relevant population of computer users
2. Identify two subject groups representing the moderator characteristics
3. Randomly select two sampling pools, and
4. Randomly assign an equal number of subjects to the two online help conditions.

Technical Communication, Fourth Quarter 1992
Researchers generally attempt to place an equal number of subjects in all treatment conditions because many statistical routines require modification to accurately handle unequal cell sizes (the number of subjects in each treatment group). When unequal cell sizes occur, researchers deal with the issue in different ways, from randomly deleting some subject data to assigning the group mean to the missing subjects. Some of these methods are questionable in some situations. If a researcher's best efforts to obtain equal cell sizes results in unequal cells, he or she should consult a research design text to identify the appropriate solution.

One final concern regards sample size. As will become apparent in the data-analysis discussion, small samples lack the statistical power to identify significant differences between groups. When samples are too small, too much variance in subject performance may preclude identifying significant relationships, and researchers may erroneously conclude that a treatment has no effect (a Type II error) [16, 26].

Many research design or statistics texts contain formulas for calculating minimum sample size; however, to use these formulas, the researcher must know some population parameters; usually the variance on relevant dependent variables [17-18]. Unfortunately, since most research in technical communication is charting new ground, such information is often unavailable. Some pilot studies are conducted for the specific purpose of obtaining the needed information to calculate an adequate sample size.

In the absence of the needed population parameters, many researchers rely on experience or common sense. In general, fewer than 10 subjects per condition is risky; the performance of this few subjects can represent considerable variation and mask differences among groups that would have been apparent if the variance were reduced. The effect size of such variance decreases as sample size increases.

Furthermore, results may be unreliable (unlikely to occur again in another time and place) when sample sizes are small. A quick examination of experimental studies of technical communication issues reveals that successful studies generally use 10 to 20 subjects per condition (this does not mean 10 to 20 subjects for the entire study). It is misleading to discuss a study's total sample size without stating how many conditions (treatment groups) are examined.

One final caution: If a study is to be conducted over time, to reduce the threat of subject mortality, researchers should start with enough subjects in the initial sessions to ensure an adequate number of subjects per condition after the virtually inevitable, though perhaps small, reduction in the number of subjects over subsequent conditions.

**Subject groups: Experimental Versus Control**

Many texts on research design define the control group as a group that receives no treatment. Perhaps a better definition emanates from medical research, where one group receives an experimental drug while the control group receives a placebo. The concept behind the use of a placebo is very important. A control group should not simply receive no treatment—no attention from the researcher—or the threat of subject history, as well as other threats to validity, arises.

---

**True experimental designs use control groups, whether the group is seen as a no-treatment group or an alternative-treatment group representing another level of an independent or moderator variable.**

If a researcher spends time and interacts with the experimental group, then a similar amount of interaction must occur with the control group. If the experimental group is given a set of instructions to study and memorize, then control group subjects should have a similar experience—they should be given some unrelated materials to study and memorize. One should realize, though, that the unrelated materials could, in and of themselves, affect performance.

True experimental designs use control groups, whether the group is seen as a no-treatment group (yet in reality some type of treatment akin to a placebo is administered), or an alternative-treatment group representing another level of an independent or moderator variable.

Once subject selection and assignment methods have been determined, the materials for the experimental conditions can be constructed.

**Designing Materials and Test Measures**

A researcher must follow careful steps in designing experimental materials and tests. Frequently the
starting point for test materials is some naturally occurring passagelographic manual, or software program. From there, the researcher revises and redesigns the materials as necessary to construct the various versions of the materials. The goal in selecting naturally occurring passages is to avoid experimenter bias that can occur when the researcher constructs the original version and also alters it to suit the experimental conditions. When researchers design original materials, they should employ objective rates to ensure that the variables are being implemented as operationalized.

One should note, however, that in experiments examining software variables, researchers often create their own software so that only the intended variables differ across software versions. Software versions and even software manuals, are often hard to test if naturally occurring materials are used because of the number of variables that differ across different products. If, for example, a researcher wants to compare two word-processing programs—Microsoft Word versus Word Perfect—he or she would find it impossible to isolate the variables and levels of variables that differ, thus making it difficult to interpret results. The ideal here, if one used existing software or manuals, would be to use one product as the naturally occurring version and then modify it to represent the other levels of the independent variables.

Ultimately the goal in creating multiple versions of the experimental materials is to construct one version per experimental condition, as represented by the various levels of the independent variables. The differing versions should vary on these levels only. Imagine a study of paragraphs with and without topic sentences, where the group reading the paragraphs without topic sentences is given paragraphs that, on average, are 35 words shorter than their topic sentence counterparts. The researcher cannot conclude whether the results are due to the presence/absence of the topic sentences or to the longer/shorter paragraphs. Theoretically, the non-topic sentence paragraphs should have words added to them to equalize paragraph length across treatments.

The dependent variables show themselves in the test materials. Researchers must decide what measures best represent the concepts they want to assess—what measures have the test construct validity. For example, a researcher wanting to assess comprehension can use multiple choice tests (also called forced choice recognition tests), free recall, cued recall, close procedures (where sentences have blanks for subjects to complete), problem solving, etc. Researchers must carefully decide what type of test to use and ensure that the concept they are interested in is actually assessed by the given tests.

For example, some psychologists argue that a simple recall test where the researcher tallies the number of idea units recalled does not measure comprehension. A researcher might decide to improve construct validity by separately scoring superordinate versus subordinate idea units recalled. Similarly, a multiple choice test can be used to measure different constructs, from factual recognition to inference formation. Researchers frequently use multiple dependent measures to assess one construct.

The researcher checks to see that these subjects’ performance is equal to chance, across response possibilities and the whole test: With a 20-question multiple choice test with four possible responses per question, a chance score would be five questions correct.

If the measures represent open-ended responses, as is the case with recall measures, then a scoring scheme for identifying acceptable responses must be designed and raters must be trained in its use. If the measures represent closed responses, as in the case with multiple choice tests, then the test should be pretested. In pilot tests, one subject group, which is later excluded from the experimental conditions, is given the test without participating in any experimental conditions. The goal is twofold: to assess (1) the different questions and possible answers and (2) the overall test.

The researcher checks to see that these subjects’ performance is equal to chance, across response possibilities and the whole test: With a 20-question multiple choice test with four possible responses per question, a chance score would be five questions correct. The researcher also assesses item difficulty and discriminability so as to ensure that certain questions and responses are not too obvious or obtuse. Such test assessment steps can lead to considerable revision and rereading of the text.

Researchers can pilot-test tasks, task materials, and tests to help refine instruments and avoid ceiling and floor effects. Ceiling effects occur when materials or tests are too easy for the selected population,
Readers who plan to conduct studies should be aware of the critical factors affecting the value of a given study and be sufficiently informed to seek further information on relevant topics in one of the research design or statistics books listed in the bibliography.

resulting in most subjects obtaining very high scores. Floor effects are the exact opposite—because of the difficulty of tasks or tests, most subjects obtain very low scores. In either scenario, real treatment effects may not be detected.

MEASURING EFFECTS

The measurement scales of the independent and dependent variables influence what statistics researchers use to analyze relationships among variables. Measurement scales vary in precision and are usually grouped into four categories. Ranked from least to most precise, the four scales relevant here are nominal, ordinal, interval, and ratio:

- With nominal scales, the researcher assigns different values or names to the properties being measured. For example, user help can be measured on a nominal scale, having two groups: online or hard copy. With nominal scales, the points on the scale have no numerical relationship (e.g., female = 1, male = 2).

- With ordinal scales, the researcher uses numbers that indicate order by magnitude. For example, test difficulty could be measured by the number of abstract words in a document—with level 1 being defined by 25 abstract words out of 1,000, level 2 being 50 abstract words out of 1,000, and level 3 being 100 abstract words out of 1,000.

- With interval scales, equal differences between measurements represent equal differences in the measured property. While all of the points on interval scales are separated by equal intervals, the zero point on an interval scale is arbitrary. The dependent measure of recall, for example, is viewed as an interval measurement in that each idea unit recalled adds one more point to a subject’s score, yet many would argue that a zero on this measure does not reflect the absence of the property.

- With ratio scales, measurements are separated by equal intervals and there is also a true zero point. For example, a count of the number of errors made would possess the quality of equal intervals as well as the quality of a true zero point—with no errors receiving a score of zero and being a conceptually correct judgment.

ANALYZING DATA

The distinction among measurement scales is important because the statistical procedures researchers rely on have been designed to work with data measured on specific scales. The scientists who created different statistical routines made restrictive assumptions about the type of data their routines could accurately manipulate. When researchers violate these assumptions, the routines can produce misleading statistics. One should remember that any higher-order scale can be converted to a lower-order scale (e.g., ratio-scale data can be converted to nominal-scale data).

Before a review of some of the more common statistics that researchers use, it will prove useful to define two terms: descriptive versus inferential statistics. Descriptive and inferential statistics differ by virtue of their function: whether it is to describe or to infer. While descriptive statistics allow researchers to describe a sample and discuss basic relationships among independent, moderator, and dependent variables, inferential statistics allow researchers to probe the strength of those relationships and make inferences about causality. Readers should view the following discussion as a starting point from which they can evaluate a researcher’s use of statistics and pursue further knowledge in relevant texts (see the bibliography).

Descriptive Statistics

Descriptive statistics use quantitative information to describe samples and their variables. They allow researchers to summarize a large set of data efficiently and describe a phenomenon in an orderly manner. Six commonly used descriptive statistics are discussed here: frequency distributions, means, medians, variances, standard deviations, and correlations.

Frequency Distributions. Frequency distributions allow researchers to summarize a set of scores by presenting the number of scores that fall into each category of the dependent variable. Frequency distributions can also be used to describe other character-
istics of a study, such as the number of Macintosh versus IBM users in a study.

Means. The most commonly used descriptive statistic is the mean or average, a measure of central tendency. Means (X or M) are calculated by summing all numerical measurements on a dependent variable and then dividing by the total number of measurements. Some researchers argue that means can be calculated only for data measured on interval and ratio scales, in that the values from any lower-order scales should not be manipulated mathematically.

Medians. Another measure of central tendency, medians are often used for data measured on ordinal scales and for skewed distributions (nonnormal curves). Medians divide a set of scores in half by virtue of their position in the distribution, thus separating the lower half of scores from the higher half. Medians are basically calculated by ordering scores from low to high, and then counting in from the two ends to the middle.

Variance. The variance (σ²), another common descriptive statistic, is itself a type of mean. It is the averaged, squared difference between each individual’s score and the mean of the sample. The variance is calculated by (1) subtracting the mean from each score and squaring the result, (2) adding the squared differences, and (3) dividing the obtained sum by one less than the total number of measurements.

Standard Deviation. The standard deviation (SD), another commonly used descriptive statistic, is the square root of the variance. The SD indicates how widely spread individual scores are: a mean of 10 with an SD of 2 indicates that the scores are much more tightly clustered (hence, less variance and a smaller SD) around the mean than, a mean of 10 with an SD of 4, assuming a normal distribution.

Correlations. One final common descriptive statistic is a correlation. Researchers seeking to examine the relationship between two measurement sets can calculate the correlation of the two sets—how one variable changes in relation to another. The most commonly used correlation statistic, Pearson’s r, identifies the relationship of data measured on interval or ratio scales. Other correlation statistics, such as Spearman’s r or Kendall’s tau, are appropriate if data are measured on ordinal scales.

Correlations range from a -1 to +1. Negative correlations indicate an inverse relationship between variables, implying that as one variable increases, the other decreases. Positive correlations imply that as one variable increases, the other variable also increases, or that both are decreasing. A correlation of computer experience (number of hours of computer use monthly) and use of help (number of times users access help in a given task) could be negative, indicating that subjects with more experience tend to access help less often. A positive correlation would indicate that subjects with more experience access help more often, and vice versa.

From r, one can calculate r², a measure of the amount of variance in the dependent variable that the relationship between the two variables accounts for. For example, if the r for the correlation between computer experience and help use were 0.80, r² would be 0.64, meaning that the relationship between the two variables accounts for 64% of the variance in the scores. Simply stated, r² allows one to assess the practical significance of the correlation.

Researchers also rely on probability tables for various correlations to determine the correlation’s significance (discussed in the next section). Such a step causes some researchers to identify correlations as inferential statistics; however, many group them with descriptive statistics because they do not generally allow for conclusions of causation (19).

Inferential Statistics

Researchers test hypotheses about populations through the measurement of samples by using inferential statistics. With such statistics, researchers make inferences, inferring information about groups that have not been observed (the population) from groups that have been observed (the sample). When a researcher analyzes data using inferential statistics, he or she calculates a specific statistic to represent the relationship among the variables within the sample data. To determine whether these statistics are significant—whether observed differences among groups are due to more than chance—the researcher consults a table of probability values for various values of a statistic or obtains a probability value from the computer printout created by a statistical package.

Probability levels represent the likelihood of obtaining a certain value of a statistic under a variety of conditions, specifically, different numbers of levels of independent and moderator variables, and different numbers of subjects. Generally, the smaller the number of subjects, the larger the statistic needs to be in order to be significant. This is one reason why small sample sizes are risky. The numbers listed in probability tables or on computer printouts are referred to as the p-values (probability values).
SELECTED READINGS ON RESEARCH METHODS


Roughly translated, the p-value (or alpha level) indicates the probability of calculating a statistic as extreme as, or more extreme than, the observed one. It indicates the probability that the observed relationship between variables occurred by chance. The smaller the p-value, the more likely the relationship is to be significant: a p-value of 0.05 indicates there is only a 5% probability that the relationship between the variables occurred by chance, while a p-value of 0.01 indicates that there is only a 1% probability.

These conclusions are merited unless Type I or Type II errors occur. With Type I errors, the researcher is committed to the p-value and concludes there is a significant relationship when in fact there is not. With Type II errors, the opposite occurs: the researcher relies on the p-value and concludes there is no significant relationship when in fact there is.

Researchers must decide in advance what probability level to use as a criterion to determine significance and hold to it. In certain situations, a p-value of 0.05 might not be sufficiently rigorous, given certain aspects of the experimental design or the state of knowledge in the field. This is particularly true with much medical research, where researchers are usually unwilling to be wrong because the cost of mistakes can be high. In such cases, the researcher may decide to accept only those findings that have p-values of less than or equal to 0.01 or 0.001.
The number of different inferential statistics is seemingly endless. Readers of this paper should realize that the brief coverage provided here is not sufficient to train them in using these procedures. Readable books on statistics abound, as do courses at universities. Those who have "that phobia" should seek learner-friendly statistics courses; they can often be found in university departments not traditionally known for such activities (e.g., Departments of Educational Psychology or Psychology).

The choice of the appropriate statistic for identifying differences and other types of relationships among groups is influenced by the following factors:

1. The scales used for the independent and dependent variables
2. The number of independent variables and levels
3. The number of dependent variables to be analyzed simultaneously
4. The type of inference to be made.

The following focuses on some of the most commonly used inferential statistics: chi-square tests, t-tests, and analysis of variance.

Chi-Square Goodness-of-Fit Tests. To compare a set of observed frequencies with a set of theoretical frequencies, a researcher can calculate a chi-square ($\chi^2$) statistic. The $\chi^2$ goodness-of-fit test assesses how well the actual distribution of the dependent variable matches an expected distribution. The $\chi^2$ test can also assess the relationship between two variables, specifically, whether two variables are independent of each other. The $\chi^2$ test accepts variables measured on any scale, but researchers must transform all data into nominal data before calculating a $\chi^2$. More specifically, the $\chi^2$ test compares observed values to expected values; the larger the difference, the larger the $\chi^2$ statistic, and the more likely it will be significant. $\chi^2$ tests are frequently used to assess subjects' answers to survey questions. Imagine a question with three possible response categories: never, sometimes, and frequently. The expected frequency for each response category, based on the distribution of the subjects is about 33% of the subjects responded never, 65% responded sometimes, and 10% responded frequently. The $\chi^2$ test would assess whether the observed and expected frequencies significantly differ.

t-Tests and Analysis of Variance. To compare group differences when the dependent variable is measured on an interval scale and independent variables are measured on nominal scales or have been converted from higher-order scales, researchers commonly use a t-test or an analysis of variance (ANOVA). Both tests assume a normal distribution of the dependent variable and homogeneous variance across treatment groups. If the study contains only two treatment groups (two levels of one independent variable), a t-test can assess whether the means of the groups differ significantly. If a study contains more than two treatment groups, ANOVA is appropriate. Both statistical procedures can examine only one dependent variable at a time.

The difference between treatment groups in our hypothetical study of interaction styles can be assessed with a t-test, where a t-statistic is calculated. The independent variable (interaction styles) and, its two levels (command versus pull-down menus) are based on a nominal scale, and the dependent measure of number of errors is measured on an interval scale. When the dependent variable is measured with ordinal or nominal scales, different statistical methods are used.

What is important is that the conclusions that researchers draw from their work be honest. They should admit the limitations of their own studies as they were conducted.

If the study contains more than two treatment groups, an ANOVA can assess differences among the groups means. With ANOVA, an $F$-statistic is calculated. When one independent variable has more than two levels, a one-way ANOVA is used; when more than one independent variable exists, an $n$-way ANOVA is used, with $n$ representing the number of independent variables (e.g., two-way ANOVAS contain two independent variables, three-way ANOVAS contain three independent variables). One-way ANOVA identifies a main effect for one independent variable (an effect of the variable on the dependent measure). N-way ANOVAS identify a main effect for each independent variable and interactions for all combinations of the independent variables.

Consider a one-way ANOVA assessing group differences for three text-difficulty conditions (A, B, and C) on comprehension. One $F$-statistic is calculated for the main effect of text difficulty. If this statistic is significant, the researcher can determine which groups differ with post-hoc multiple-comparison techniques (e.g., Tukey HSD, Fisher LSD, Bonferroni-t, Dunnet, Scheffé, Dunncan, Newman-Keuls, and polynomial orthogonal contrasts).
Imagine that condition A has a mean of 12, condition B a mean of 10, and condition C a mean of 8. Given a significant F value from the ANOVA, certainly the highest and lowest means (A and C) differ significantly, but only a multiple-comparison technique can determine whether A and B differ, and whether B and C differ.

Multiple comparisons should be run whenever ANOVA identifies a significant main effect for independent variables with more than two levels. Researchers should consult a statistics book to determine which multiple-comparison technique is most appropriate for a given design.

Many times researchers use t-tests to interpret interactions that occur in n-way ANOVAS. In a two-way ANOVA that consists of two independent variables with two levels each (a 2 x 2 ANOVA), a significant interaction would indicate that some of the four treatment means significantly differ from each other. Consider a study examining the effects of test length (long and short) and test familiarity (unfamiliar and familiar) on recall. While a significant main effect might be, “Short tests elicited significantly higher recall scores than long tests,” a significant interaction might be, “Familiarity with test content significantly improved subjects’ recall in long tests but not in short tests.” Researchers often graph significant interactions to identify the trend suggested by the treatment means—when the trend is obvious in terms of which means seem to be extremely high or low, many rely on descriptive interpretations as opposed to t-tests.

A concern with both ANOVAs and t-tests arises in situations where researchers are required to obtain multiple scores per subject. A statistical application that accounts for the relationships among the multiple scores must be used. Multiple scores can occur when both pretests and posttests are administered, or when any test is given more than once and is measured with the same metric (e.g., the same 20-question multiple choice test immediately after reading a passage and again 2 weeks later).

If the independent variable is nominal and the dependent variable is interval, a repeated-measures ANOVA or a matched or paired t-test can successfully analyze the within-subjects designs, which occur whenever subjects have multiple scores on the same metric. The use of regular ANOVAS or unpaired t-tests would essentially deny the relationship of subject performance on multiple tests, and significant results could be misleading. In fact, within-subjects designs increase statistical power by allowing individual subjects to act essentially as their own controls; thus, the individual subjects’ performance across tests helps explain the overall variance and, in turn, reduces random error.

A typical example of the need for within-subjects designs occurs when the same subjects are tested on simple versus complex tasks. Only a statistical procedure such as a repeated-measures ANOVA or a paired t-test can accurately assess subjects’ scores on the multiple tasks. If the simple and complex tasks are randomly ordered, then researchers can draw conclusions about the effect of sizele versus complex tasks on the dependent variable. However, if the researcher wants to reach conclusions about the order of simple or complex tasks, he or she would need to counterbalance their order, giving half the subjects simple tasks before complex tasks, and the other half complex tasks before simple tasks. This researcher would still have a within-subjects design.

In reporting both t-tests and ANOVAS, the researcher should report—

1. The t or F statistic.
2. The degrees of freedom for the statistic (the number of values of the variable that are free to vary). With t-tests, the number of degrees of freedom equals the total number of subjects for all treatment groups, minus one subject per group. With ANOVAS, additional degrees of freedom values are reported: for example, the number of levels of each independent variable minus one for the main effects and the product of these differences for the interactions.
3. The probability of obtaining such a statistic.
4. The means and SDs for the treatment groups. With two-way (or greater) ANOVAS, the researcher also needs to report the means and SDs associated with significant main effects. These means are obtained by viewing the data as if there were only one independent variable at a time (collapsing all independent variables but one) until all main-effect means and SDs are obtained.

Other Statistics

A number of other techniques have been developed to extend the analysis of the relationship between multiple independent variables and a single dependent variable measured on an interval or ratio scale. These methods include multiple regression and analysis of covariance (ANCOVA).

Multiple regression, like correlation, allows for
the prediction of a value in the dependent variable

given certain values of the independent variable.

ANOVAs are actually a special kind of regression

where uncorrelated independent variables are as-
sessed, yet multiple regression can assess the relative

contribution of multiple independent variables to

scores on the dependent measure. Multiple regres-
sion offers advantages because it can handle both
discrete (nominally measured) and continuous (ordi-
nal, interval, and ratio) independent variables as op-

posed to ANOVA that require nominal independent

variables or the transformation of higher-order scales
to the lower-order nominal scale, a transformation

that can weaken relationships [18, 6].

ANOVA's are useful when one wants to remove

the effect of a known variable (e.g., subject differ-

ences in reading ability, typing speed, pretest scores)

by using that variable as a covariate. ANOCOVAs use

the covariate to adjust for subject differences and ac-


tually correct for these differences, thus reducing

overall variance.

Readers who may never conduct studies

should have sufficient knowledge to
critically consume the findings of empirical
research or to understand the quality of
research they may hire a consultant to
do.

If the research question involves more than one

dependent variable or correlated independent vari-

ables, the researcher might make use of multivariate

statistics such as multivariate regression analyses,
canonical correlations, factor analyses, etc. These

analyses are beyond the scope of the present discus-
sion, and interested readers should consult the bibli-
ography in the box for further sources. Some texts
contain decision trees or tables that help researchers
select appropriate statistics [3, 269; 18, 63; 20, 205].

CREATING THE RESEARCH DESIGN

Given the preceding discussion of the factors in-
volved in empirical research, the examination of a

few common research designs becomes relatively

straightforward. One should note that all true experi-

mental designs rely on random assignment of sub-
jects to conditions. The classical experimental design,
described in numerous texts using a rotation system
developed by Campbell and Stanley [11], is concept-
ually equivalent. The posttest-posttest control

group design shown in part I of Table 1. This design
allows for the comparison of two groups, represent-
ing either one level of an independent variable and

treatment or two levels of one independent variable.

The difference between pretest and posttest

scores is also computed. If the independent variable

is measured on a nominal scale and the dependent

variable is measured on an interval scale, a two-way

repeated-measures ANOVA could be used to analyze

the scores. Treatment would be seen as a between-

subjects factor (different subjects participated in sep-

date treatments), and the two test scenarios would

be seen as a within-subjects factor. Making this a

mixed design because of the existence of both types

of factors.

Like the pretest-posttest control group design (I),

the posttest-only design (part II of Table 1) controls

for threats to validity and removes sources of bias. If

one takes design I and removes the pretest, the simi-

lar design II occurs; two treatment groups (one

treatment and one control) and a posttest for both
groups. If the independent measure is nominal and

the dependent measure is interval, a simple t-test

could analyze the results.

Extending the simple posttest-only design with

the examination of more than two levels of one inde-

pendent variable creates a one-way design, a very

useful and common design (see part III of Table 1). This

design, given a nominal independent variable and an

interval dependent variable, could be ana-

lyzed with a one-way ANOVA if a significant F-

value occurred. Follow-up multiple comparisons

should also be run.

The final design shown in part IV of Table 1 is a

factorial design where two independent variables are

analyzed: one with two levels and one with three

levels. The factorial nature of the design means that

six (2 x 3) experimental conditions are created.

Given nominal independent variables and an interval

dependent variable, a 2 x 3, two-way ANOVA could

analyze the data. A multiple comparison would clas-

sify a significant main effect or the second variable

(because it has more than two means), and t-tests
could help interpret a significant interaction.

Other research designs known as pre-experimen-
tal or quasi-experimental designs are used under cer-

tain circumstances. Such designs do not protect

against all threats to validity or sources of bias. For

example, time-series designs (multiple observations

of one group, a treatment, and more observations)}
### Table 1

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>Random assignment of Subjects</th>
<th>Pretest</th>
<th>Treatment</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Pretest-Posttest Control Group Design</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>Yes</td>
<td>Level A of independent variable I</td>
<td>Yes</td>
</tr>
<tr>
<td>Control (or experimental)</td>
<td>Yes</td>
<td>Yes</td>
<td>Placebo, no treatment, or Level B of independent variable I</td>
<td>Yes</td>
</tr>
<tr>
<td>II. Posttest Only Design</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level A of independent variable I</td>
<td>Yes</td>
</tr>
<tr>
<td>(or control)</td>
<td>Yes</td>
<td>No</td>
<td>Level B of independent variable I, or no treatment</td>
<td>Yes</td>
</tr>
<tr>
<td>III. Expanded Posttest Only Design</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level A of independent variable I</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level B of independent variable I</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level C of independent variable I</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level D of independent variable I, or no treatment</td>
<td>Yes</td>
</tr>
<tr>
<td>IV. 2 x 3 Factorial Design</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level A of independent variable I</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level A of independent variable II</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level A of independent variable III</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level B of independent variable I</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level B of independent variable II</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level B of independent variable III</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level C of independent variable II</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level D of independent variable I</td>
<td>Yes</td>
</tr>
<tr>
<td>Experimental</td>
<td>Yes</td>
<td>No</td>
<td>Level D of independent variable II</td>
<td>Yes</td>
</tr>
</tbody>
</table>

are used when it is impossible to use a control group. A typical example would be when a researcher wants to evaluate a curriculum change that has occurred across all sections of a course. In such instances, pre-experimental design such as a one-shot case study or a one-group pretest-posttest design might also be used.

Another common quasi-experimental design is a non-equivalent control group design, where the treatment and control groups represent intact groups that cannot be randomly assigned to conditions. Of course a pretest must be administered to reveal whether the groups are essentially equal on critical measures.

Numerous other designs are discussed in research design texts.

### Conclusion

Clearly, the conduct of true experiments is an elaborately detailed task. This paper has presented many critical factors involved in sound experimental design. Readers who may never conduct studies should have gained sufficient knowledge to critically consume the findings of empirical research or to understand the quality of research they may hire a consultant to conduct. Readers who plan to conduct studies should have become aware of the critical factors affecting the value of a given study and be sufficiently informed to seek further information on relevant topics in one of the research design or statistics books listed in the bibliography.

Careful planning and the observation of a mass of
details can and does pay off in valid and reliable studies that advance theory and contribute to the field. True experimental designs, by virtue of their isolation of experimental variables, allow for the conclusion that changes in performance are attributable to experimental treatments. Further, true experiments are replicable.

One should realize at this point, though, that no study is ever perfect and that many decisions that researchers make amount to trade-offs. It seems at times that as internal validity increases, external validity decreases. This situation reflects one of the tradeoffs between laboratory and field research. What is important is that the conclusions that researchers draw from their work be honest. They should admit the limitations of their own studies as they were conducted and suggest the need for further research to assess the areas untapped by their work.

No one will have a problem with a study that has reasonable limitations if the researcher acknowledges them and suggests the next step. In fact, one can only admire the researcher who intelligently designs and conducts a study while admitting to its limitations.

**REFERENCES**