Chapter 6

Experimental Research

In the most basic sense of the word, "experiment" denotes "a test, trial, or tentative procedure, esp. one for the purpose of discovering something unknown or of testing a principle, supposition, etc.; . . . the conducting of such operations; . . . or to try or test esp. in order to discover or prove something" (*Random House Webster's College Dictionary* 1991, 470). According to this definition, a child touching a stove to see if it really is hot is experimenting. She/he is testing whether a parent's claim about the stove being hot is true, as well as what "hot" means in reference to stoves.

Experimental research similarly engages in testing or trying for the sake of discovery or proof, but it employs more rigorous rules for testing/trying and discovery/proof than the child did. In fact, most people think of experimental methodology in terms of these rigorous rules, which govern:

- formulation of hypotheses for testing
- probability sampling of subjects for study
- operationalization of key terms in the hypothesis, or the formation of working definitions for what behaviors or objects count as evidence of the constructs being studied
- use of standardized instruments or other pilot-tested instruments of proven validity and reliability
- attempts to control for and minimize sources of error in measurement
- use of inferential statistical techniques on measurement data

Experimental research is almost always quantitative; it often generates studies described as "scientific"; and it can be terrifying for beginning researchers. This last is most unfortunate, as experimental methodology is a powerful heuristic. Some critics object that experiments, with all their controls, invoke an "artificial situation that may not reflect how people think and behave in the real world," but these controls are precisely what make experimental studies more replicable than others (Case 2002, 169). In the real world, librarians' responses to reference questions are unlikely ever to be reviewed by peers in comparison to

rating criteria as a measure of their quality. An experiment that measured the quality of reference by peer review/rating criteria would not capture the complexity of reality, where librarians and patrons have their own measures of quality and often bring their feelings toward each other into their informal calculations of quality. Because the procedures for peer review and the rating criteria have been codified, however, such an experiment could be repeated easily by other researchers to see whether they get similar results. Replicability of more "real life" approaches to measuring the quality of reference answers—ones having librarians and patrons apply their own rating procedures and criteria—would, in contrast, depend upon one researcher finding librarians and patrons who chanced to have the same standards and feelings as those in another study. This latter is an unlikely possibility given the diversity of human beings!

Moreover, by "applying a theory to a particular case in an attempt to test the theory" (Case 2002, 165), experimental methodology offers clear ways to bridge theory and practice. Do the "big ideas" of theories actually hold up when tested? Can "commonsense" practices be proven to have some basis in theory? Historically, only a small percentage of studies within librarianship have used experimental methodology (Grotzinger 1981, 39ff.), and it has long been hoped that this percentage would increase with concomitant benefits to the profession if only librarians would become more familiar with the benefits and techniques of experimental methodology.

FINDING A TOPIC

Some authors emphasize other aspects in defining experimental methodology, notably planned interventions or the manipulation of changes in the environment (Case 2002; Goldhor 1972; Green 2003) or the investigation of causality (Busha and Harter 1980; Goldhor 1972; Gustafson and Smith 1994; Peng 2003). We feel, however, that it is most important to keep the focus on the twin components of testing/trying and discovery/proof contained in the basic definition of "experiment." Other guides to research methodology support our decision to focus on a broader range of experiments here, as they, too, include experiments investigating correlation as well as causation. Sometimes they describe these as ex post facto experiments, quasi-experiments, or retrospective studies. (Cf. Busha and Harter 1980, and Carey 1998.) As these guides emphasize, a "true" experiment is not always possible (Busha and Harter 1980), or may take too long or be too expensive (Carey 1998). Moreover, studies of correlation serve as a "tool for uncovering potential causal links" for future studies (Carey 1998, 31): "Indeed, if two things actually are causally linked they will be correlated, so evidence of a correlation may provide some initial reason to suspect two factors are linked" as cause and effect (Carey 1998, 10).

These two components (testing/trying and discovery/proof) are what most characterize and distinguish experimental methodology. Experimental methodology is not content with descriptions of things as they are (e.g., usage statistics on library databases); rather, it seeks to predict relationships between things (e.g., does patron status relate to database usage?) Causal relationships between things (e.g., does being a graduate student cause higher levels of database use?) are a particular focus of experimental methodology in its most rigorous form, although we include in Figure 6-1 examples of some less rigorous studies that test relationships without seeking causation because we feel that beginning researchers may find these models more helpful. Researchers should also keep the basic testing/trying, discovery/proof orientation of experimental methodology in mind because it is technically possible to employ other research methodologies within an experimental framework. For example, Young and Ackerson (1995) actually employ content analysis within an experimental framework. The testing/trying, discovery/proof can be seen in their use of experimental and control groups to discover whether "modified" library instruction led to better bibliographies in student research papers than "traditional" library instruction did. The bibliography entries were examined and scored on a rating sheet using the methodology of content analysis, but the traditional/modified comparison was experimental.

Experimental studies within librarianship, such as those listed in Figure 6-1, typically take one of three basic forms:

- *Testing to see whether a particular technique or device works in a predicted way.* Often, experiments investigate whether a technique or device (the independent variable; for instance, a particular form of library instruction) is "effective" for accomplishing a given purpose by using statistics to compare the results (the dependent variable) obtained with that technique/device to those obtained by a control group not exposed to the technique/device (e.g., Vidmar [1998]; Young and Ackerson [1995]). At other times, a model of how a technique or device works is constructed, and actual performance is measured against projected performance (e.g., Gerhart [2004]; Nowicki [2003]).
- *Testing to see which of two or more techniques or devices for accomplishing a given purpose leads to better results.* The results (the dependent variables) for groups using the different techniques or devices (the independent variables) are compared statistically with each other, and sometimes also with control groups which were exposed to none of these techniques or devices (e.g., Kenney et al. [2003]; Weston and Lauderdale [1988]).
- *Testing statistically for relationships between various characteristics of items or individuals.* These characteristics can be thought of as independent variables, or intrinsic properties of the item or individual (the age of an item,

Figure 6-1: Studies Using Experimental Methodologies

Gerhart, Susan L. 2004. "Do Web Search Engines Suppress Controversy?" *First Monday* 9 (1). Available: http://firstmonday.dk/issues/issue9_1/gerhart (accessed May 18, 2007).

"Our starting assumption is: A controversial subtopic is revealed or suppressed to the degree its URLs are recognizable in the query for the broad topic." Compared the top 100 results in three search engines and two metasearch engines for searches on broad topics, controversial subtopics, non-controversial subtopics.

Kenney, Anne R., Nancy Y. McGovern, Ida T. Martinez, and Lance J. Heidig. 2003. "Google Meets E-bay: What Academic Librarians Can Learn from Alternative Information Providers." *D-Lib Magazine* 9 (6). Available: http://dlib.org/dlib/june03/kenney/06kenney.html (accessed May 18, 2007).

Selected set of 24 questions of sort typically encountered by reference librarians for study. Got answers to questions, difficulty rating of them, from Cornell University librarians and Google Answers. Then had answers blind-reviewed by librarians based on specified quality criteria. Found that overall Cornell rated a 3.43 and GA rated a 3.39—no real difference, although they did score more differently on different sub-sets of questions.

Moore, Deborah. 2005. "GCC Research Project on Information Competency." Available: http://glendale.edu/library/instruction/documents/ICEval05.pdf (accessed May 18, 2007).

Focus on Library 191, GCC's introductory course in information competency. Matching students who took LIB 191 with a randomly selected control group on enrollment status, prior GPA, primary language, units attempted. Compared two groups over several semesters on GPA, units completed, persistence to next semester. Found that "students passing Library 191 had significantly higher GPAs and completed significantly more units."

Nowicki, Stacy. 2003. "Student vs. Search Engine: Undergraduates Rank Results for Relevance." *portal: Libraries and the Academy* 3 (3): 503–515.

Investigated end-user judgments of relevance of search engine results, as well as performance of search engines in producing results rated as relevant. Seventy-five undergraduates searched topics of their choosing in six search engines. Students' ranking of relevancy of top ten results compared to that of search engine. Found no significant correlation between student rankings and search engines.

Pask, Judith M., and E. Stewart Saunders. 2004. "Differentiating Information Skills and Computer Skills: A Factor Analytic Approach." *portal: Libraries and the Academy* 4 (1): 61–73.

Survey given to incoming first-years at Purdue to determine computer skills and IL skills provided data sets. Used factor analysis on responses to 23 questions designed to measure these skills to investigate relationship of computer and IL skills. Thirty questions divided into six areas. Found that computer literacy and IL not the same.

Smalley, Topsy N. 2004. "College Success: High School Librarians Make the Difference." *Journal of Academic Librarianship* 30 (3): 193–198.

"This study examined levels of student achievement as recorded on Library 10 grade rosters and asked: Do students from high schools in the one school district that has library media teachers do better in the Information Research course when compared to students from the high schools that do not have librarians?" (pg. 194). Tracked and analyzed for correlations information about students' high school background, midterm class rank and grade in the Information Research course, final class rank and grade in the Information Research course.

(Cont'd.)

Figure 6-1: Studies Using Experimental Methodologies *(Continued)*

Vidmar, Dale J. 1998. "Affective Change: Integrating Pre-Sessions in the Students' Classroom Prior to Library Instruction." *Reference Services Review* 26 (3/4): 75–95.

Tests whether a "pre-session [which] involves the librarian going to the classroom of the students for 10 to 20 minutes prior to the students coming to the library for an actual library instruction session" is effective in "address[ing] the affective needs of students by helping to reduce anxiety and the resistance of students" (76). Compared scores from three pairs of classes: one class from each pair was the control group, which just got LI, while the other was the experimental group, which got the pre-session.

Weston, E. Paige, and Diane S. Lauderdale. 1988. "How Do We Learn What a Database Includes? A Case Study Using Psychology Dissertations." *RQ* 28: 35–41.

Hypothesis: "... that since PsycINFO includes such a vast number of dissertation records ... surely searches in PsycINFO would retrieve the majority of the dissertations of interest on any given psychology topic and would overlap extensively with searches in Dissertation Abstracts Online." Had two librarians search 14 psychology topics submitted by doctoral candidates in psychology. Analyzed, compared the number, kinds of items retrieved.

Young, Virginia E., and Linda G. Ackerson. 1995. "Evaluation of Student Research Paper Bibliographies: Refining Evaluation Criteria." *Research Strategies* 13 (2): 80–93.

Question was: "Does a course-integrated curriculum devised to develop skills in the use of print and automated information sources ... help students access, evaluate, and use current information to write better term papers?" (83). Study over five semesters with 251 students. Used a bibliography rating sheet to standardize scoring.

the sex of an individual), and dependent variables, or non-intrinsic properties of the item or individual (the number of circulations of an item, the score on a test of library anxiety for an individual). Researchers predict and test statistically for relationships between independent and dependent variables (e.g., Moore [2005]; Pask and Saunders [2004]; Smalley [2004]).

Within these three broad categories, experimental studies display great diversity in duration, their approaches to the experimental situation, and their models of data analysis.

- Most studies are cross-sectional, or focus upon subjects at single points in time. Young and Ackerson thus measure the effects of instructional methods only over the academic term in which students received the instruction, not over the longer term of their undergraduate education. However, other studies are longitudinal and track subjects over time, as Moore does when she relates student completion of an information literacy course to GPA and persistence over community college careers.
- In some studies, researchers must create the data-gathering situation and gather data before running statistical analyses. Kenney et al. (2003), for example, needed both the production of answers to reference questions and ratings of these answers for their experiment. In other experiments, though, researchers test hypotheses by looking for statistical relationships

in preexisting data. Moore (2005) thus used data on student enrollment status, GPA, and units completed that the college registrar had already gathered.

- Experiments can also be unobtrusive or obtrusive in their design, with subjects unaware or aware of being studied. Smalley's (2004) study is unobtrusive, because researchers can easily record students' high schools, GPA, and class rank without students being aware. In contrast, Vidmar's (1998) study is obtrusive because students completing pre- and post-tests are necessarily aware of doing so—and of being studied.
- When it comes to models of data analysis, some experiments can be described as correlation studies, which establish relationships between two or more variables without establishing causation (e.g., Kenney et al. 2003); factor-analytic studies, which reduce a large set of correlated variables to a smaller set of hypothetical traits or factors underlying the correlations (e.g., Pask and Saunders 2004); or two-condition experimental studies, which manipulate variables in testing cause and effect (e.g., Vidmar 1998).

FORMULATING QUESTIONS

When experimental methodology is used, the research question should take the form of a hypothesis, or a "declarative statement about the relation between two or more variables which can be observed empirically" (Busha and Harter 1980, 10). There are several different types of hypotheses. Non-directional hypotheses predict the existence of differences between groups in an experiment without predicting which groups will perform better, while directional hypotheses predict which groups will perform better. Null hypotheses predict that there will be no differences in performance between groups. Null hypotheses are only used for testing statistical significance. That is, researchers show that the null hypothesis cannot be true. Some sample hypotheses from prior library experiments posit that:

- "both skill and confidence levels increase as a result of cumulative exposure to the library and its services" (Greer, Weston and Alm 1991, 552);
- "a given, well-known specific controversy will not be revealed" in the top results for Web search engines (Gerhart, see Figure 6-1);
- "students in the experimental tutorial group will learn as much or more than the in-class group and will also report as much or more satisfaction with the learning experience" (Nichols, Shaffer and Shockey 2003); and
- "academic procrastination is positively related to library anxiety" (Onwuegbuzie and Jiao 2000).

Ideas for hypotheses, or likely relationships between variables, come from professional experience and from reading library literature. Because your reading shapes your predictions about possible relationships, always ground your hypothesis in the literature on the topic whenever writing about your research. Weston and Lauderdale (see Figure 6-1), for example, do a good job of explaining how reading of vendor documentation and prior studies lead to their hypothesis that PsycINFO should perform as well as Dissertation Abstracts in retrieving dissertations on psychology topics.

Too many experimental studies within librarianship are, regrettably, weak in formulating their hypotheses. Sometimes the hypothesis does not limit itself to the relationship between the variables but rather attempts to prove something broader. One group of researchers used as its hypothesis the claim that "research would confirm the viability of moving to an online format for introductory information literacy instruction." (The source of this quotation and those in the next few examples are deliberately not cited here in order to protect the identity of those whose hypotheses are being critiqued.) All their experiment investigated, however, was whether students receiving library instruction via an online tutorial learned more and were more satisfied than students receiving instruction via lecture/demonstration. Viability involves more than student test scores and satisfaction. Even if scores and satisfaction were higher, a tutorial would not be viable if the library had no staff or equipment to create or maintain it. As this example shows, a hypothesis that seeks to answer a value question (Should something be done? Is this good?) can be problematic. In other cases, researchers state their hypotheses only obliquely. For example, the following text could easily be written as an hypothesis:

> The authors supposed initially that since PsycINFO includes such a vast number of dissertation records . . . surely searches in PsycINFO would retrieve the majority of the dissertations of interest on any given psychology topic and would overlap extensively with searches in Dissertation Abstracts Online.

"PsycINFO provides as comprehensive and current coverage of psychology dissertations as Dissertation Abstracts" would make a fine hypothesis here. Questions about relationships can also be written easily as hypotheses, as can statements of "research objectives" and those predicted relationships that readers are left to infer. (See Figure 6-2 for further examples). The problem with poor hypothesis formulation is that it makes it harder for both researchers and readers to determine whether the predicted relationship between the variables does in fact hold true. ("If you don't specify a predicted event precisely, there are an indeterminate number for ways of an event of that general kind to take place" [Paulos 1988, 28].)

A hypothesis is simply an informed "best guess." You should not be vague in stating your hypothesis for fear that the experiment will not support it. If not enough is known about a topic to make even an informed "best guess," conduct an exploratory study using nonexperimental methods.

Figure 6-2: Examples of Hypotheses

Original Text	Possible Hypothesis
"Do students from high schools in the one school district that has library media teachers do better in the Information Research course when compared to students from the high schools that do not have librarians?" (Smalley 2004, 194)	Students from high schools in school districts with library media teachers will score more highly in a college level Information Research course than students from high schools in school districts without library media teachers.
"The project attempted to determine if a pre-session prior to a regularly scheduled library instruction session would have any effect upon the student attitudes toward the library, the librarians, the relevance of using the library, and the effectiveness of library instruction." (Vidmar 1998, 82)	Pre-sessions prior to a regularly scheduled library instruction session will positively impact student attitudes toward the library, the librarians, the relevance of using the library, and the effectiveness of library instruction.
"The main purpose of this study was to examine whether international students' acquisition of library skills was related to their English language proficiency." (Bilal 1989, 130)	International students' acquisition of library skills will be proportional to their English language proficiency: the more proficient an international student is in English, the more library skills she/he will acquire.

DEFINING THE POPULATION

Some methodologies are designed to work with the subjects available to the researcher, or with typical rather than representative subjects. Classroom research, for example, uses existing groups of students, and content analysis can focus upon the most famous or characteristic examples of discourse on a topic. Experimental methodology is not such a methodology. Because it seeks results that are widely generalizable, it typically uses research subjects that are representative of a broader population. The population is the largest "group to which the researcher feels it is appropriate to generalize or apply the results" (Gustafson and Smith 1994, 76). Because most populations are too large for every member to be studied, researchers get subjects from the population by sampling, or selecting a portion or subset of the larger population. How many subjects are needed in a sample and how subjects should be selected are the important and difficult questions here.

We will turn to those questions momentarily, but first a caution about representativeness. "Representative" essentially means that the members of the sample match the larger population in some demographic characteristics: age, sex, income, and so on. A researcher could potentially select a sample that is representative of its population in some characteristics (age, sex, ethnicity) but that is not representative in terms of some other characteristics (e.g., income) which may actually be the major variables underlying the experimental findings. For this reason, it is not enough for researchers to say that their sample is

"representative"; rather, they should specify of what population their sample is representative and why this population is relevant to the experimental focus.

Some researchers work under the misconception that 10 percent of the population represents an adequate sample size (e.g., Ware and Morganti 1986), or that larger samples are automatically better. The appropriate sample size is actually based upon the confidence level and confidence interval sought by the researcher, as well as upon the size of the total target population. The confidence level indicates the probability that subjects' responses do not represent the effects of chance alone. A 95 percent confidence level means that in only five cases out of 100 would the subjects have given that set of responses because of chance alone. In the 95 other cases, something other than chance—namely, the effects of the treatment or independent variable being studied—accounts for their responses. The lower the confidence level, the more likely the results could be due to chance alone. Ninety-five percent and 99 percent confidence levels are the most commonly used, with the 95 percent level being standard in social sciences research. The confidence interval is a plus-or-minus percentage indicating how often repeated random samples of a given size would be expected to measure a quality's "true" value. Certain confidence intervals are normally associated with certain size samples (as is shown in Figure 6-3), although the effects of target population size on sample size are somewhat more complex than can be discussed here. In general, "the larger the sample the greater our chances of getting a ratio close to that in the population [in the sample]; however, as sample size increases the chances of getting an exact match between sample and population frequencies decrease" (Carey 1998, 18). Larger samples generally mean smaller confidence intervals and hence greater assurance that the results obtained are not simply the effect of chance alone. A larger sample is also more likely to

Figure 6-3: Confidence Intervals Associated with Common Sample Sizes*

Sample Size	Approximate Margin of Error
25	+/–22%
50	+/–14%
100	+/–10%
250	+/– 6%
500	+/– 4%
1,000	+/– 3%
1,500	+/– 2%

*Based upon a 95% confidence level and a potentially infinite population.
(This figure is based on information from Carey 1998, 20.)