10-1995

Basics of research (Part 4): research study design (Part 2)

Cheryl Thompson
University of Nebraska Medical Center, cbthompson@unmc.edu

Edward A. Panacek
University of California, Davis

Eric Davis
Strong Memorial Hospital

Basics of Research (Part 4): Research Study Design (Part 2)

Cheryl Bagley Thompson, PhD, RN, CS,[1] Edward A. Panacek, MD,[2] Eric Davis, MD[3]

1. University of Utah College of Nursing, Salt Lake City, Utah
2. Division Of Emergency Medicine and Clinical Toxicology, University of California, Davis, Medical Center, Sacramento, Calif.
3. STAT Medevac, Center for Emergency Medicine, Pittsburgh, Pa.

Key Words: clinical research, research, research design, research methodology

Address for correspondence: Cheryl Bagley Thompson, PhD, RN, CS, University of Utah College of Nursing, 25 South Medical Dr., Salt Lake City, Utah 84112

After a research design has been decided, a number of details remain undetermined before the study is initiated (the protocol is "fleshed out"). This fourth article in this series on the basics of research discusses decisions related to the sample to be used, the type of data to be collected, the method to use to collect the data, and the potential sources of bias in the research procedures. Because this area can be complex, definitions of some of the new terms used within the paper are in Table 1.

Sampling Techniques

One of the first concerns of the researcher is how to enroll subjects in the study. The first step is to determine the subject population. The term *population* refers to all potential subjects for the study. For example, if a researcher is interested in stress levels of health-care providers who transport patients by air, *all* nurses, paramedics, EMTs, physicians and technicians employed by air transport programs would be included in the population. However, the population of interest may be more narrow. The researcher may wish only to investigate air medical personnel in the United States, or alternately just nurses and paramedics employed by transport programs within the United States.

In contrast, the *sample* used for the research project contains only the subjects who actually will participate in the study. In other words, the sample contains the small portion of the population selected for analysis. How this sample is selected from the entire population of subjects is important to the quality of the study. A poorly selected sample may yield biased results that cannot be applied to individuals outside of the sample (i.e., the results do not apply to the entire target population).

Random Sampling Methods

Several methods can be used to select a sample (Table 2). The most powerful sample is one that is selected randomly from the population. *Random selection* means that every potential subject in the population has a known probability of being selected for participation and that probability is quantifiable (i.e., it can be calculated). The most common way to approach a random sample is to give everyone an equal chance of participating in the study. This is called a *simple random sample*. However, if a small subgroup of subjects is the group of interest, the investigator may need to divide the population into major groups before random selection is applied to ensure that the smaller group of interest is included in the sample.

For example, an investigator may be interested in high school students. The investigator wants to be sure that some of the students are from the special education class. However, if only 2% of the students are in special education in the high school of interest, a simple random sample may not provide any special education students. Consequently, a *stratified random sample* may be drawn in which 98% of the subjects are selected randomly from the general student body, and 2% of the subjects are selected randomly from the group of special education students. This approach assures that both groups are included proportionally within the sample. The researcher also can elect to alter the proportions in the sample from the proportions present in

Table 1

Definition of Terms

Population - all subjects of interest to the researcher for the study
Sample - the small portion of the population selected for participation in the study
Sampling - the process used for selecting a sample from the population
Simple random sampling - a process in which a sample is selected randomly from the population with each subject having a known and calculable probability of being chosen
Stratified random sampling - a process in which a population is divided into subgroups and a predetermined portion of the sample is randomly drawn from each subgroup
Systematic random sample - a process in which a sample is drawn by systematically selecting every *n*th subject from a list of all subjects in the population. The starting point in the population must be selected randomly
Cluster sampling - a process in which the sample is selected by randomly choosing smaller and smaller subgroups from the main population
Convenience sampling - a process in which a sample is drawn from conveniently available subjects
Snowball sampling - a process in which the first subjects are drawn by convenience and these subjects then recruit people they know to participate, and they recruit people they know, etc.
Quota sampling - a process in which subjects are selected by convenience until the specified number of subjects for a specific subgroups is reached. At this point, subjects are no longer selected for that subgroup but recruitment continues for subgroups that have not yet reached their quota of subjects
Purposive sampling - a process in which subjects are selected by investigator to meet a specific purpose
Judgmental sampling - another name for purposive sample
Internal validity - the degree to which the changes or differences in the dependent variable (the outcome) can be attributed to the independent variable (intervention or group differences. This is related to the degree to which extraneous variables are controlled
History - where natural changes in the outcome variable is attributed to the intervention instead
Maturation - where changes in the dependent variable are a result of normal changes over time
Instrumentation - where changes in the dependent variable are the result of the measurement plan rather than the intervention
Loss of subjects - changes in the dependent variable are a result of differential loss of subjects from the intervention groups
Assignment of subjects - where changes in the dependent variable are a result of pre-existing differences in the subjects prior to implementation of the intervention
Blocking - assigning subjects to control and experimental groups based on an extraneous variables. Blocking helps to assure that one group will not get the preponderance of subjects with a specific value on a variable of interest
External validity - the degree to which the results can be applied to others outside the sample used for the study
Hawthorne effect - subjects respond in a different manner just because they are involved in a study
Biophysiologic measures - measures of biological function obtained through use of technology, such as electrocardiogram or hemodynamic monitoring
Self-report - the variables of interest are measured by asking the subject to report on the perception of the value for the variable
Psychological scale - usually a number of self-report items combined in a questionnaire designed to evaluate the subject on a particular psychological trait, such as self-esteem
Observation - the activity of interest is observed, described, and possibly recorded via audio or video tape
Validity - how well the tool measures what it is supposed to measure
Face validity - the instrument looks like it is measuring what it should be measuring
Criterion-related validity - the results from the tool of interest are compared to those of another criterion that relates to the variable to be measured
Concurrent validity - criterion-related validity where the measures are obtained at the same time
Predictive validity - criterion-related validity where measurement using one instrument is used to predict the value from another instrument at a future point in time
Content validity - is concerned with whether the questions asked, or observations made actually address all of the variables of interest
Construct validity - a form of validity where the researcher is not as concerned with the values obtained by the instrument but with the abstract match between the true value and the obtained value
Reliability - the degree of consistency with which an instrument measures the variable it is designed to measure
Stability - determination of the degree of change in a measure across time
Determination of stability - is only appropriate when the value for the variable of interest is expected to remain the same over the time period examined
Interrater reliability - the degree to which two or more evaluators agree on the measurement obtained
Internal consistency - the degree to which items on a questionnaire or psychological scale are consistent with each other

the population. In the example above, the researcher may instead select 90% from the general student body and 10% from the special education students. This approach would provide more information on a subgroup that constitutes a small portion of the population. In this example, although the chance of being selected is not equal for all students, the probability of being selected is known for each individual, and, thus, the sample is selected randomly.

Random selection can be accomplished in a variety of ways. One of the most common is to draw names out of a hat. If the researcher is interested in members of the National EMS Pilots Association, then the name of each member is placed on a piece of paper and put into the hat. One slip of paper is drawn for every subject required for the study. A second method uses a table of random numbers to select individuals from the list of the population. Another method uses the list of all possible subjects, but divides the total number of potential subjects by the number of subjects needed. The answer is used as the interval from which to pick names on the list. For example, if there were 1,000 names on the list, and 50 subjects were needed, 1,000 divided by 50 is 20. Consequently, every 20th name from the list would be selected. If the starting point for the selection is determined randomly (i.e., drawing one of the num-

Table 2

Sampling Methods

Probability
Simple random sample
Stratified random sample
Systematic random sample
Cluster sampling
Nonprobability
Convenience sample
Snowball sampling
Quota sampling
Purposive sample
Judgmental sampling

bers 1 to 20 out of a hat), and the list does not have a pre-established nonrandom order (e.g, if males and females were listed alternately), then this method of sample selection is considered to produce a *systematic random sample.* If the researcher starts at one and picks every 20th person, it is, instead, a *systematic nonrandom sample* and may be a source of bias.

A final method of obtaining a random sample is *cluster sampling.* This approach may be used in cases in which a list of all subjects in the population is not available. Instead of randomly selecting subjects, smaller and smaller groups of subjects are selected. For example, to select a random sample of nurses employed in emergency departments (ED) of major cities, a list of all states would be created and the desired number of states randomly selected. Next, a list of major cities in the selected states would be created and a set of cities selected randomly. A list of all hospitals in the selected cities would be created and a sample of hospitals selected randomly. Finally, from the list of hospitals a complete list of ED nurses would be constructed and the final sample randomly drawn. The advantage of this method is that random selection is preserved without having to obtain a list of every nurse in the United States employed in an ED in a major city. Consequently, sample selection is not only easier but less expensive.

Nonrandom Sampling Methods

Unfortunately, true random samples often are difficult to obtain. Rarely does the investigator know the names of all subjects of interest. If the population consists of all adult trauma patients transported by air, it is impossible to know ahead of time who will be a trauma victim. It is equally difficult to obtain access to all of these individuals once their identity is known. Consequently, although nonrandom sampling techniques are technically inferior, nonrandom samples are the type most commonly used for health-care research.

Convenience samples are the most common type of nonrandom samples. As the name suggests, they are subjects who are convenient to the researcher. In the case of adult trauma patients, the sample would consist of patients transported by the participating teams during the time period of the study. A variant of convenience sampling is *snowball sampling.* In this case, the initial subjects identify other individuals who also may be interested in participating. For example, the sample is obtained by recruiting air medical personnel at the annual conference who then talk with friends and encourage them to participate.

Quota sampling is similar to stratified random sampling, in that a specific number of subjects from different subgroups is recruited. The difference is that subjects are recruited by convenience rather than randomly. Once the quota for a subgroup is met, subjects are no longer recruited for that subgroup. So if 40 gunshot wounds, 40 abdominal blunt trauma, and 20 head injuries are required for the study, patients with gunshot wounds no longer will be recruited once 40 subjects meeting gunshot criteria have been enrolled. The advantage of quota sampling is that the researcher can be more specific about the type of subjects required for the study and assured that specific subgroups are represented adequately. As with convenience sampling, bias in the method of selection of subjects for the subgroups still may exist. An additional disadvantage is that subject recruitment may be more difficult if subjects from a subgroup are difficult to recruit.

Purposive sampling is even more restrictive than quota sampling. In this case, the researcher has specific requirements for the sample and picks subjects who meet these strict criteria. For example, the researcher may be interested in the behavior of experts but recognizes that there may be regional differences. The investigator then purposely could select a number of nationally recognized experts in air transport from each of the Association of Air Medical Services regions. Another case in which purposive sampling could be used is when the sample will be small and 100% cooperation is needed. In such cases, the researchers may ask specific subjects that they know will volunteer and follow through with the study protocol. This approach is also sometimes called *judgmental sampling,* because it is dependent on the judgment of the investigator as to who qualifies for inclusion. However, this "judgment" may lead to investigator bias in subject selection.

Since a nonrandom selection of subjects is much easier to obtain, and sometimes the only way to get subjects, why use more "expensive" random sampling techniques?

Random sampling techniques provide a higher quality of research results. First, selecting a sample at random helps to reduce bias from the process of sample selection itself. For example, your transport program may have a different philosophy, may have a different set of protocols, or may just differ in the quality of care provided to patients when compared to other air transport services. As a result, any study performed using subjects "convenient" to your program may give results that are biased by these factors. As a consequence, the results only would be applicable to your program and not to other programs.

A second reason for using a random sample relates to the statistical analysis of the data at the conclusion of the study. The inferential statistics commonly used for health-care data—*t*-tests, analysis of variance, multiple regression and correlations—were developed on the assumption that the sample under study is truly random. The tables used to determine if your results are different enough to be considered statistically significant were developed using random samples. Consequently, purists

may say that there are inaccuracies in your statistical analysis if your sample was not selected randomly.

What can you do if randomly selecting the sample is not possible? First, the researcher should try to avoid bias in sample selection. Patients should be selected using explicit inclusion/exclusion criteria without first trying to determine if they will be "good" subjects whose data likely will support your hypothesis. Second, the researcher should try to diversify the sample as much as possible. For example, multicenter studies have a wider range of subjects than do single-site studies. Finally, the researcher can look for bias once the data are collected and institute statistical controls, if necessary, to correct for characteristics that might bias the study results. For example, in a study comparing drug A with drug B, an analysis of covariance may need to be used instead of a simple analysis of variance. The analysis of covariance allows the researcher to control for the effect of an extraneous variable, such as gender, that might be causing bias in the results.

Table 3

Sample Size Example

Example	Values for SBP in Population							True Average
1	220	110	98	60	130	190	100	129.7
	Values for SBP in Sample 1							**Sample Average**
	220	110	190					173.3
	Values for SBP in Sample 2							
	110	60	100					90.0
	Values for SBP in Sample 3							
	110	60						85.0
2	114	120	128	132	138	136	140	129.7
	Values for SBP in Sample 1							
	114	120	136					123.3
	Values for SBP in Sample 2							
	120	132	140					130.7

SBP = systolic blood pressure

Sample Size

Once the researcher has determined where the sample will be obtained, he or she needs to determine how many subjects should be asked to participate. The larger the sample, the more the sample will resemble the target population of interest. However, large samples are expensive to recruit and expensive to use in the study. Consequently, the research needs to compromise so that the maximum good can be obtained from the smallest possible sample.

Power analysis is a commonly used technique for determining adequate sample size for intervention-type studies. This technique uses information about the size of the change or difference between the study groups that is expected, how much of a chance the researcher wants to take that the results may be wrong, and the statistical techniques that will be used.[1]

There are a number of software programs for personal computers that can be used to do a power analysis and don't require a background in statistics. However, time spent with a statistician during the planning phase of a study is recommended and often helps avoid problems later in the study.

Individuals not conducting a formal power analysis by hand or by computer can use heuristics and other rules of thumb to determine an adequate sample size. The first factor to consider is the strength of the relationship that is expected. If the intervention will not cause a very large effect, then a larger sample size is needed. The same is true if two or more groups are being examined. The smaller the difference between the groups, the larger the number of subjects needed to find the differences.

Another factor to be considered is how much difference between the subjects at baseline is expected. If the study population is heterogeneous, a small sample may look only at a few of the subjects on the fringes of the population. However, if most of the subjects will be similar, then fewer subjects will be needed to get an accurate idea of the nature of the population. See Table 3 for an example. Other factors to consider in determining sample size are the number of subjects expected to drop out before the study is concluded (increase sample with increased attrition), the number of variables being examined (increase sample with more variables), the number of subgroups into which the sample will be divided (increase sample with more subgroups), and the sensitivity of the tools used to measure the expected effects (increase sample with insensitive measurement tools).

Survey research using mailed questionnaires requires a particularly large sample to obtain valid results. Mailed surveys have a relatively low return rate, often averaging below 50%.[2] However, air medical programs and personnel have exhibited a much higher return rate for mailed questionnaires.[3]

Defining the Study Population

Once the number of subjects has been determined, attention must be focused on developing the criteria that define the subjects in the target population (Table 4). Inclusion criteria to determine the specific characteristics of subjects must be itemized. Exclusion criteria to specify subjects that are not to be included in the sample also must be itemized.

Inclusion criteria help to ensure that subjects fulfill the needs of the researcher. Common inclusion criteria include demographic parameters, clinical characteristics, geographic considerations and the temporal setting. Demographic parameters help to ensure a degree of homogeneity in the sample. For example, when studying the effect of surfactants on neonatal respiratory distress, an upper age limit will be necessary as part of the definition of a neonate. Clinical characteristics help to narrow the sample to subjects appropri-

Table 4

Defining the Target Population

Inclusion Criteria

Goals	Specific and focused on a target population
	Demographic Parameters
	Clinical characteristics
	Geographic considerations
	Temporal setting
	Informed consent

Exclusion Criteria

Goals	Attempt to predict and eliminate analysis problems
	Probable confounding variables
	High risk of "lost to follow-up"
	Inability to provide good data
	Ethical constraints

ate to the study. For example, subjects who are hospitalized may not be good candidates for a study on the effect of a new drug on long-term blood-pressure control. Geographic considerations may help to limit subjects to an area accessible to the researchers or to ensure geographic diversity.

Temporal setting may be important in a number of ways. First, sleep-research subjects may need to be available in the evening. Second, inclusion criteria could specify that patients be at least 24 hours post-op. Third, a study could require that the subjects be divided into two groups based on a temporal factor. For example, patients whose asthma symptoms lasted less than 24 hours would be in one group, and individuals whose symptoms lasted more than 24 hours would be in another group. Finally, temporal requirements could be part of the randomization plan. For example, only patients seen during the first week of the month might be included in a long-term study.

A final consideration for inclusion criteria is that of *informed consent.* Ethics will be discussed in further detail in a future segment of the series. However, consent is important when considering inclusion criteria. A common inclusion criterion is that subjects must provide verbal or written consent to be eligible for the study.

Exclusion criteria are as important as inclusion criteria because they help to predict and/or to eliminate potential analytic problems. Probable confounding variables commonly are used as exclusion criteria. For example, if patients taking digoxin are known to react differently to the new blood-pressure medication being studied, all patients on digoxin should be excluded from the study.

Exclusion criteria also help to facilitate the research process. Subjects that may provide poor quality data or who are difficult to recruit into the study or keep in the study do not make the researcher's job easier. Consequently, exclusion criteria often are developed to keep these individuals out of the sample. Two common examples are the ability to speak English and the ability to read. Individuals who cannot do one or the other may not be able to comply with the research protocol and might be excluded from the sample. An example of subjects at risk of "lost to follow-up" might be patients transported by air to a facility other than the base hospital. The geographic distance between the research team and the patient may be too great, and these patients might be excluded as potential subjects.

Finally, *ethical constraints* may dictate specific exclusion criteria. Prisoners often are viewed as individuals at risk for violation of their personal rights. Because of the risk that the prisoner did not feel free to refuse to participate in the study, they may be excluded to eliminate possible hints of ethical violations.

Inclusion and exclusion criteria should be considered carefully before initiating a study. Too strict of criteria limit the ability to solicit a sufficient number of subjects. Alternately, too few criteria put the researcher at risk for confounding variables or a difficulty in obtaining an acceptable data set.

Other Sources of Bias in Research Design

Threats to Internal Validity

Many factors potentially can introduce bias into a research study. *Internal validity* is the degree to which changes or differences in the dependent variable (the outcome) can be attributed to the independent variable (intervention or group differences). In other words, are the study results really true? Although the terms sound the same, internal validity is not related to instrument validity. Internal validity is study-specific rather than instrument-specific.

Extraneous (or confounding) variables are factors that can influence study outcomes but that are not part of the study itself. Extraneous variables threaten the internal validity of the study and may include history, maturation, instrumentation, loss of subjects and assignment of subjects.

History, one potential extraneous variable, occurs when natural changes in the outcome variable are attributed mistakenly to the intervention instead. For example, if a program added a second team member to the transport team, improvement in quality of patient care after an educational program may not be due to the education, but rather to other reasons, such as the added personnel.

Another related extraneous variable is *maturation.* Maturation refers to changes in the dependent variable as a result of normal changes over time. For example, after a surgical procedure, pain naturally decreases over time. Thus, an investigator could not necessarily attribute a decrease in pain after surgery to the intervention because of the role of maturation threatening the internal validity.

Repeated measurement of the dependent variable (outcome variable) can be used to control for the effects of history and/or maturation. Analysis of trends over time can help identify changes due to the intervention versus changes that would occur even without intervention.

Instrumentation can be a threat to internal validity in several ways. One method is when the researcher uses a different tool to measure the variable of interest at time X than was used at time Y. However, a threat to internal validity also could exist if the same instrument is used at a short interval, and the subject could learn from time X how to react at time Y. Care must be taken that the tool itself does not act as an intervention separate from the intended intervention.

Loss of subjects during the study artificially can impact outcome, if some variable besides chance effects mortality of subjects. For example, if only the subjects who do not like your approach or

who do not respond successfully to your treatment drop out, you will have an artificial approval of your intervention. See Table 5 for an example.

Finally, the method of *assignment of subjects* to experimental and control groups could influence the outcome of the study. If, for example, all subjects were assigned to the experimental group until that group had enough subjects, and the second half of the volunteers were assigned to the control group, the first group might be significantly different than the second group even before the intervention was applied. Individuals who volunteer early may be inherently different than those who dawdle and volunteer later after much encouragement. Consequently, study results only may reflect beginning group differences rather than true effects of the intervention.

The most successful method for dealing with selection difficulties is randomly to assign individuals to the groups. This avoids any pre-existing bias in the subject assignment. However, randomization may not be "kind" to the investigator. For example, even if the investigator flips a coin to assign subjects to groups, the investigator could be unlucky and get 14 out of 20 heads rather than an even distribution of heads and tails, or could by chance get more males in group No. 1 than in group No. 2.

If the investigator needs to assure that subjects with a particular characteristic are distributed evenly, subjects can be blocked before assignment to groups based on pre-existing characteristics. *Blocking* entails setting up groups based on specific characteristics. For example, if gender is expected to make a difference, the investigator can assure that an equal number of males and females are assigned to each group. Control for extraneous variables also could be handled by using statistical control. Statistical control is the process of using pre-existing variables as covariates or additional factors in the statistical analysis. (See Cohen and Cohen[4] for further information.) In addition, the investigator can try to sort out effects of instrumentation by adding additional groups that do not receive a pretest (e.g., Solomon 4 group designs, see Part 3 in this series[5]).

Table 5

Mortality Example

Results with Mortality

Control	Experimental
75 80 82 78 84 (Average = 79.8)	85 81 (Average = 83)

Results without Mortality

Control	Experimental
75 80 82 78 84 (Average = 79.8)	60 65 72 85 81 (Average = 72.6)

In the example, if only the subjects that were averse to the intervention dropped out of the study, then the results would support the fact that the intervention improved scores. However, if all subjects remained in the study, the results would instead suggest that the intervention decreased scores.

If the investigator does not wish to block on a potentially confounding variable, a homogeneous sample that does not vary may be used. For example, if gender is expected to cause differences in the outcome, the investigator could study only males or only females.

Threats to External Validity

In contrast to internal validity, *external validity* is the degree to which the results can be applied to others outside the sample used for the study. In many cases, the results can be generalized only to individuals included in the study because of something unique about the group or the situation. Environmental variables, such as the temperature of the room, the frequency of rest for the subjects or even the investigator's presence, may influence the status of the subject or the measurement activities of the researcher. For example, if the study was done in a hot room at the end of the day, the results may be generalized only to tired, irritable subjects, but not to subjects who do not have these characteristics.

The *Hawthorne effect* is another factor that may influence external validity. The Hawthorne effect occurs when subjects respond in a different manner just because they are involved in a study. For example, it may be the influence of having a researcher paying attention to the transport program that causes the subjects to change their attitude and performance rather than as the result of the study intervention (independent variable).

Repeated measurement of anything can take a toll on the subjects. If the subjects are exposed to a large number of questionnaires, observations, etc., they may become tired of the procedures or are so accustomed to them that their performance is altered. Consequently, data obtained from a complex study may apply only to others involved in similar complex studies.

One overriding goal of research is to investigate a small sample of subjects and then to be able to apply the findings to a broad group (i.e., the population). External validity is the degree to which this goal can be met. Nonrandom sampling techniques inherently limit the external validity of the study because of greater potential for bias in subject selection. In addition, the inclusion/exclusion criteria are a two-edged sword. Designed to maximize the internal validity of the study by minimizing potential confounding variables, if they excessively narrow the study population, they can limit the external validity.

Measurement Collection Methods

Once the design and subject selection procedures are determined, the researcher must consider how the variables of interest will be measured. Many different methods of data collection are available depending on the research question and resources of the investigator. Data-collection methods vary in the degree of structure, quantifiability, researcher obtrusiveness and objectivity. Highly structured methods are preferable when a specific, nonexploratory research question is being asked. For example, structured methods would work well for the question "Is heparin or normal saline a better agent to maintain

Table 6

Data Collection Methods

	Quantifiability	Objectivity	Structure	Obtrusiveness
Biophysiological	XXX	XXX	XXX	XX
Self-Report	XX	XX	XXX	XXX
Observation	X	X	Depends	Depends

Number of Xs symbolizes the degree to which the characteristic is met.

patency of a heparin lock?" In contrast, less structured methods may be appropriate for the question "What is the experience of being transported by helicopter for acute chest pain?"

Some variables are inherently more quantifiable than others. Blood pressure and other vital signs are easily quantifiable. However, level of stress or skill in intubation are less readily quantifiable. Measurement of all variables need not be quantifiable, but reproducability and reliability are usually higher when the measure can be quantified.

Obtrusiveness of the research protocol can impact the quality of the data obtained. Individuals under scrutiny by a researcher may alter their usual behavior, either for better or for worse. If observation during flight is used as a research method, it may be difficult for the observer to remain unobtrusive because of the small space involved. The observer should make every attempt not to interfere with the normal process of events. In addition, participant bias is reduced if the purpose of the observer is blinded to the participants.

Finally, measurement techniques can vary in degree of objectivity. *Objectivity* is the degree to which two individuals can provide the same measure on a specific variable. Two people determining end-tidal CO_2 as a measure of intubation success would be more objective than two people determining success by visual inspection alone. Degree of objectivity is increased when the measurement technique relies more on standard procedure than on subjective opinion. Objectivity also is increased when the observer is not involved in provision of patient care or other research activity being measured.

Biophysiologic measures, self-report and *observation* are three common methods used to collect data for investigations, and vary in their degree of structure, quantifiability, researcher obtrusiveness and objectivity (Table 6). To identify the measurement-collection methods best for the project, the investigator should first list the variables of interest in the study and included within the hypotheses or research questions. Once the methods for data collection are identified, the researcher should become aware of the limitations of the particular method of data collection chosen and implement procedures to limit the difficulties whenever possible. There are generally two ways to accomplish this. One approach is to have the protocol and data collection sheets reviewed before the study by as many people as possible. The other approach is to "pilot test" the data-collection method, before the full study, using old charts or a few actual patients.

Biophysiologic measures are increasingly common with health-care research. This trend is due partially to the increased technological nature of health care. The transport environment includes many biophysiological devices, and air transport personnel are confident in the use of the equipment and interpretation of the data. Consequently, air transport researchers are comfortable with the technology and at ease with its use in their research. Biophysiological measures include, but are not limited to, blood pressure, weight and heart rate. Standards for the measurement of each of these variables are available, increasing the objectivity of the measures, and the ability to reproduce results from moment-to-moment or researcher-to-researcher.

A primary disadvantage of biophysiological measures can be high reliance on their validity and reliability. The presence of a quantifiable number may give a false sense of accuracy. If a temperature gauges reads 98.64 degrees, it may or may not actually be accurate to 0.01 degree. Researchers should establish, rather than blindly accept, the degree of accuracy present in their physiological measures. Another limitation results from increasing complexity of biophysiological devices. Such devices can provide inaccurate data unless they are used correctly. With increased complexity, it may be more difficult to detect equipment malfunction.

Self-report data also are common within the health-care environment. Self-report data are easy to obtain and, with some approaches, can be given at least the appearance of quantifiability. Self-report data can be in the form of diaries, interviews or completion of a list of written or verbal questions. Self-report can be used to measure attitudes, psychological tendencies and behaviors. In some studies, self-report is the only way to measure the variable of interest, especially when the variables are subjective. For example, attitudes towards specific policies may not be amenable to observation, but the subjects may be willing to express their views in a written or verbal format. Self report is not as constrained as other methods. An individual may be able to recall feelings or experiences from a previous point in time when observation or biophysiological measurement were not possible. As an example, this approach can be used to measure amounts of "pain."

Surveys or mailed questionnaires are common forms of self-report data because of their ease of development and analysis. The usual format is to pose a question and leave a space for the subject's response. The more specific the answer requested, the easier data analysis, but the more stilted the responses might be. For example, it is easier to tabulate the number of people who support use of helmets for air transport versus those who don't support, rather than summarizing opinions regarding helmet use by air transport personnel.

Another common approach to collecting self-report data is a *psychological scale*. Researchers have developed specific questionnaires to measure variables, such as work satisfaction, self-esteem and quality of life. The ad-

vantage of this approach is that usually the validity and reliability of the instrument has been established previously, a method for data analysis is predetermined, and time-consuming instrument construction is avoided. The disadvantages include concerns that the tool does not precisely measure your variable of interest and that the originator might charge for use of the instrument.

Observation is the final approach to data collection to be discussed. In observation, the activity of interest is observed, described and possibly recorded via audiotape or videotape. The investigator then analyzes the episode for the variables of interest. For example, a researcher interested in infection-control activities during transport may ride along and note each occurrence in which an appropriate precaution is taken and each occurrence in which a principle of infection control is violated. The data can be quantifiable, as in the previous example, or of a more subjective nature. Studies examining administration of cardiopulmonary resuscitation may collect data, such as observed depth of compression or adequacy of chest rise during ventilation. Although more intrusive methods could be used to provide quantitative data, such as measured depth of compression or tidal volume, observation and a subjective appraisal may be used to minimize intrusiveness of the data collection.

Observation methods have the advantage of being usable in many settings, of maintaining some of the context of the situation, of providing a way to re-examine the situation after it occurred, and of allowing for interpretation by the researcher. Observations that are recorded can be analyzed by more than one individual in an attempt to decrease the subjective nature of data analysis. Observation, however, has several disadvantages. Bias in recording and evaluation of the observations is a possibility, even with a conscious effort to increase objectivity. The presence of an observer or a recording device may make the subject more aware of their actions, causing alteration in their behavior.

Validity of Measurements

In designing a research study, the investigator attempts to use the best tool for measuring the variable of interest. Unfortunately, the true score of the variable is never known absolutely. An obtained score always is altered to a certain degree by "error in measurement." The error in measurement can have multiple causes, including validity and reliability of the instrument.

Validity is the "degree to which an instrument measures what it is supposed to be measuring."[6] Biophysiological measures have "relatively" high validity because the measurement technique may be based on the definition and on basic scientific principles. For example, blood pressure is the pressure in the cardiovascular system. The measurement of pressure is a relatively straight- forward process. In contrast, development of a valid tool to measure pain is more difficult. Not everyone agrees on a definition of pain, so developing a tool to address a nebulous and subjective entity is more difficult. The researcher might question whether the tool measures pain, or whether it really measures something else, such as the related concept anxiety.

Validity is difficult to ensure because absolute knowledge cannot be obtained. Researchers use several "roundabout" methods to try and demonstrate that an instrument is valid and measuring what it says it measures. Of all of the measures of validity, *face validity* may be the easiest to establish. Face validity means that the instrument looks like it is measuring what it should be measuring[6] and is an intuitive and subjective judgment. At minimum, a tool must have face validity. As this is the weakest test of validity, other approaches also should be used.

Criterion-related validity uses the process of comparing the tool of interest to another criterion that relates to the variable of interest. A critique of this approach is that if there is another tool that can be used as the "gold standard," why not use it instead. Use of the "gold standard" may be suitable in most cases, but sometimes the better instrument may not be appropriate in the research environment. For example, to establish the criterion-related validity of pulse oximetry as a measure of blood oxygenation, the values obtained from pulse oximetry might be compared to the values obtained from an arterial blood gas. During air transport, blood gases are not available, so the less-invasive pulse oximetry might be the only way to obtain SAO_2 data for a study. In the above example, the blood gas and the pulse-oximetry measure would be obtained at the same time to establish criterion-related validity.

The above method is considered establishment of *concurrent validity,* as the two measures were done at the same time. Another form of criterion-related validity is *predictive validity.* Here, the measure of interest is obtained, and, at a future time, another criterion is measured. If X leads to Y with a certain frequency, and you measure X, then you should be able to measure Y to verify the validity of X. For example, if the revised trauma score measures severity of injury and should predict mortality, then a proven correlation between trauma score and patient mortality would be evidence of predictive validity for the revised trauma score.

Content validity deals with whether the questions asked or observations made actually address all of the variable of interest. Content validity relates more to self-report data and observations than to biophysiological measures. However, content validity also would be relevant when looking at composite biophysiological measures that are combined to make more complex assessments. For example, content validity of the revised trauma score would be established by determining if the individual components of the revised trauma score covered all of the items necessary to describe the severity of the trauma.

Unfortunately, content validity cannot be measured directly in most cases, as is possible with criterion-related validity. Establishment of content validity relies mostly on the opinion of experts. For educational assessment tools, comparison of the tool against the list of objectives or course outline might be an approach to the establishment of content validity. In this way, content validity is similar to face validity. The difference is that face validity often involves the same people both as the subjects and as the experts. Also, content validity is more concerned

with the question of whether everything is covered and nothing is left out. As a result, content validity uses more specific and objective criteria.

Construct validity is, perhaps, the most difficult to understand and measure. Establishment of construct validity is an abstract process, as the researcher is not as concerned with the values obtained by the instrument, but with the abstract match between the true value and the obtained value. Further discussion is beyond the scope of this series. (For more information, see Polit and Hunger, 1995.[6])

Instrument Reliability

In contrast to validity, *reliability* is the degree of consistency with which an instrument measures the variable it is designed to measure.[6] Fortunately, establishing instrument reliability is easier than establishing validity. It is important to note that an unreliable instrument cannot be valid. If the instrument does not measure something the same way twice, the instrument cannot be measuring what it is supposed to measure. In contrast, an instrument can be very reliable and yet not have validity. For example, if you take a blood pressure multiple times, and each time it is the same, that is a reliable measure. But if you say you are determining level of stress, measuring blood pressure by itself is not a valid measure of stress, despite its obvious reliability.

As with validity, there are several types of reliability, (e.g., stability across time, interrater reliability, internal consistency and equivalence). *Stability* across time is measured using the test-retest approach. A measurement is taken at one point in time and then repeated using the same situation, instrument, etc., at a second point in time. This approach to measuring reliability is only appropriate when the variable being measured can be considered stable across the chosen period of time. For example, the height of an adult can be expected to remain the same for relatively long periods of time. To measure the stability of a ruler as measure of height, one height could be taken today and another in a month. If the measure, such as weight, could be expected to change more frequently, placing the two measurements at a one-month interval could not be expected to provide test-retest reliability. Instead, having the individual step off the scale, wait a minute or two, and then step back on the scale would be a more appropriate evaluation of stability because weight does not fluctuate over a one- to two-minute period of time. When a researcher wishes to examine test-retest reliability, careful consideration must be made of the length of time over which stability reasonably can be expected.

Interrater reliability is the degree to which two or more evaluators agree on the measurement obtained. For example, to test interrater reliability of a blood-pressure measurement, a double stethoscope would be used to determine whether both researchers would agree on a single blood-pressure value. This method is most important in assessing methods that have a greater degree of subjectivity (e.g., patient mental status). Researchers using observational methods should examine interrater reliability before collecting study data to assure that everyone is looking for the same thing.

Internal consistency is more complex and is the degree to which items on a questionnaire or psychological scale are consistent with each other. Questionnaires that are consistent have items that are directed at measuring the same thing. For example, a scale to measure self-esteem would have a number of questions directed at measuring a component of self-esteem. Achieving a questionnaire with internal consistency is a balancing act. The goal is to be consistent without being redundant. Long questionnaires may not be completed; the goal is to ask as few questions as possible that provide a valid measure of the variable of interest.

Two main techniques are used to measure internal consistency, *split-half reliability* and *Cronbach's Coefficient Alpha*. A discussion of the two methods is beyond the scope of this series. (Further information can be found in Polit and Hunger.[6])

A final form of reliability is *parallel forms*. Parallel forms is an examination of two instruments used to measure the same variable. For example, you may not want all students to have the exact same test if they are sitting close to each other when taking the exam. You also may want to repeat the exam at a short interval and do not want subjects to remember questions from the first time. To ensure that the instruments are reliable, the researcher needs to have one group of subjects complete both forms at the same sitting. A correlation between the two forms is done to determine the degree of reliability.

Conclusion

There are many factors that impact the quality of a research study. Not all points must be addressed with a given design. In many cases, common sense will help the researcher identify potential sources of bias in the research design. Not all sources of bias can be eliminated, but an attempt should be made to eliminate or reduce bias when possible.

Submitting the research proposal to others is a helpful method for determining sources of bias. Comparison of your research protocol to published reports of other similar studies also may be helpful. The methods section of a research report should present the steps taken by the researchers to minimize bias. Similar approaches then can be used in the proposed study.

This part in the series is meant to discuss the many issues associated with "fleshing out" a research protocol. The subject can become complex because of the broad spectrum of clinical research. It is impossible to go into each area in great detail, but a number of reference textbooks are available for those who wish to learn more on this subject. The important planning phase of a study can take longer and be more difficult than the study itself.

As discussed in the first parts of this series, a research proposal should be based on sound scientific principles. However, the quality of the science and the ethics of a study are two different issues. The next article in the series will discuss the ethics of research and methods for assurance that the rights of human subjects are protected within the research design.

References

1. Cohen J: Statistical Power Analysis for the Behavioral Science. Hillsdale, N.J. Lawrence Erlbaum Associates, 1988.
2. Heberlein TA, Baumgartner, R: Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review* 1978;43:447–462.
3. Balazs K, Thompson CB: Quality assurance and continuous quality improvement within air transport programs. Submitted for publication 1995.
4. Cohen J, Cohen P: Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences., Hillsdale, N.J., Lawrence Erlbaum Associates, 1983.
5. Panacek EA, Thompson CB: Basics of Research (Part 3): Research study design Part 1. *Air Medical Journal* 1995;14:139–146.
6. Polit DF, Hungler BP: *Nursing Research: Principles and Methods, 5th edition,* Philadelphia Pa., J.B. Lippincott Company, 1995.

Recommended Texts

1. Polit DF, Hungler BP: *Nursing research: Principles and methods, 5th edition.* Philadelphia Pa., J.B. Lippincott Company, 1995.
2. Bailey DM: *Reserach for the Health Professional: A Practical guide.* Philadelphia, Pa., F.A. Davis Co, 1991.
3. Hulley SB, Dummings SR: *Designing Clinical Research.* Baltimore, Md., Williams and Wilkins, 1988.
4. Okolo En: *Health Research Design and Methodology.* Boca Raton, Fla., CRC Press, Inc., 1990.