

# Measurement Issues

*Editors' note: This article is the sixth in a multipart series designed to improve the knowledge base of readers, particularly novices, in the area of clinical research. A better understanding of these principles should help in reading and understanding the application of published studies. It should also help those involved in beginning their own research projects.*

This part of the Basics of Research series discusses issues of measurement. Good decisions relative to data elements and the method selected for their measure are essential in ensuring a quality study. Because of space considerations, this issue can only present basic content. For those desiring more in-depth knowledge, the classic book on psychometric theory by Nunnally and Bernstein may be of value.<sup>1</sup>

## Data Collection Methods

Once the research question(s), study design, and subject selection procedures are determined, the researcher must consider how the variables of interest will be measured. All studies measure something. Numerous methods of data collection are available, depending on the research question and resources of the investigative team. Data collection methods vary in degree of structure, quantifiability, researcher obtrusiveness, and objectivity. Highly structured, objective methods are preferable when a specific, non-exploratory research question is being asked. For example, structured methods would work well for the question, "Is Approach A or Approach B a better method to accomplish X?" In contrast, less structured methods may be appropriate for the more general 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 and skill in intubation are more difficult. Measurement of all variables need not be quantifiable, but reliability and sometimes validity 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, the observer may not be able 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.

Measurement techniques also can vary in degree of objectivity. Objectivity is the degree to which a measurement is directly observable, without influence of opinions, emotions, or bias, regardless of who is making the observation. Different people determining end-tidal CO<sub>2</sub> as a measure of intubation success would be more objective than two people determining success by auscultation of the chest alone. Objectivity is increased when the measurement technique uses standard procedures or machine processing rather than subjective opinion. Objectivity of observations is also generally increased when the observer is not a direct participant in patient care or another activity that is being measured.

Biophysiologic measures, self-reporting, and direct observation are three common methods used to collect data for investigations and vary in their degree of structure, quantifiability, researcher obtrusiveness, and objectivity. To identify the measurement (data collection) methods best for a given project, the investigator should first list the variables (independent and dependent) of interest in the study, that is, those listed within the hypotheses or research question. The investigator should then select methods for data collection that are the most objective and quantifiable, balancing idealism and practicality.

*Biophysiologic measures* have become the most common in health care research. This trend is due partially to the increased technological nature of health care. The transport environment includes many biophysiologic devices, and air transport personnel are confident in the use of the equipment and interpretation of the data. Consequently, air transport researchers are generally comfortable with the technology and at ease with its use in their research. Biophysiologic measures include such things as blood pressure, oxygen saturation, end tidal CO<sub>2</sub>, and heart rate. Standards for the measurement of each of these variables are available, which increases the objectivity of the measures, as well as the ability to reproduce results from moment-to-moment or researcher-to-researcher.

One disadvantage of biopsychologic measures is that the presence of a quantifiable number may give a false sense of accuracy. If a temperature gauge reads 98.64 degrees, it may or may not actually be accurate to 0.01 degree. For laboratory tests, the hospital laboratory must undergo certification, which ensures the accuracy of their results. For other tests, researchers should establish, rather than blindly accept, the degree of accuracy present in their physiologic measures. With the increasing complexity of biopsychologic devices, be aware that they can provide inaccurate data if not used correctly.

*Self-report(ed) data* also are common within the health care environment. Self-report data are easy to obtain and, to variable degrees, can be quantifiable. 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 or highly personal. For example, attitudes toward specific policies may not be amenable to observation, but the subjects may be willing to express their views in a written or verbal format. An individual may be able to recall feelings or experiences from a previous time when observation or biopsychologic measurement was not possible. Surveys or mailed questionnaires are common forms of collecting self-report data because of their ease of development and analysis. A future article in this series will focus on survey research.

Another common approach to collecting self-report data is use of an existing *score* or *scale*. Researchers may be able to identify a previously developed and validated questionnaire to measure their variable of interest, such as for work satisfaction, self-esteem, depression, or quality of life. Using an existing scale or questionnaire has several advantages. First, time-consuming development of a new instrument construction is avoided. Second, instrument validity and reliability may already have been established. Third, an approach to the data analysis already exists. The disadvantages of using a pre-existing questionnaire tool include concerns that it may not precisely measure the variable of interest or that the originator might charge for use of the instrument (some are proprietary).

*Observation* is a third general category of measurement. In observation, the activity of interest is directly observed, described, and possibly recorded (eg, audiotape or videotape). The investigator then analyzes the observations 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. Recorded data can be objective and quantifiable, 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 less intrusive or more practical. Whenever possible, there should be prospective rules for how the data is recorded and later interpreted to increase objectivity and reliability.

Observations have the advantages of being easier, more available, able to maintain some of the context of the situation, and allowing for interpretation by the researcher at the time of occurrence. Observations that are recorded can be analyzed by more than one individual in an attempt to decrease subjectivity. Observation, however, has several disadvantages. Bias in recording and evaluation of the observations is more likely than with biopsychologic measurements. The presence of an observer or a recording device may make the subject more aware of their actions, causing alteration in their behavior, even if blinded to the study purpose. This is sometimes referred to as the *Hawthorne effect*.

## Accuracy of Measurements

In designing a research study, investigators attempt to use accurate tools for measuring the variables of interest. However, the true score, or value, of a variable is never really known. Any measurement or score obtained in a study always consists of two theoretical elements: the true value and "error." Most measurements should be assumed to contain a certain degree of "error" in measurement. Understanding common sources and categories of errors is useful in deciding which variables to measure in a study and how best to measure them.

## Validity

Validity refers to how well an instrument measures the concept it intends to measure.<sup>2</sup> In research terminology, it means the same thing as "accuracy." Biopsychologic measures generally have high validity because the measurement techniques are commonly based on set definitions and on accepted scientific principles. For example, blood pressure is the pressure in the cardiovascular system and the measurement is a relatively straightforward process. In contrast, development of a valid tool to measure pain is more difficult. Not everyone agrees on a definition of pain, and there is no way to directly measure it. The experience of pain is what is measured; developing a tool to address a nebulous and subjective entity is more difficult. The researcher might question whether the tool really measures pain or something else, such as anxiety. Validity can be difficult to "prove" because often absolute knowledge is not possible. However, several accepted alternative methods exist to demonstrate that an instrument is valid in measuring what it says it measures.

Of all of the measures of establishing validity, *face validity* may be the easiest to establish. Face validity means that the instrument simply looks like it is measuring what it should be measuring.<sup>2</sup> It is an intuitive and subjective judgment of whether the measurement makes sense, sometimes referred to as passing the "sniff test." At a minimum, measurement tools must have face validity. Because this is the weakest test of validity, other approaches also should be considered.

*Content validity* expands the concept of face validity and considers whether the questions asked or observations made actually address the entire variable of interest. Expert opinion and often a panel of experts (more faces) are used in this evaluation. Content validity is primarily used for self-report and observation types of data, rather than for biopsychologic measures. However, content validity also could be relevant when

looking at composite biophysiologic measures that are combined to make more complex scores or scales. For example, content validity of the revised trauma score could be established by determining whether the individual components of the revised trauma score covered all of the items necessary to describe the severity of the trauma.

Content validity cannot be measured directly, as is possible with predictive-related validity (see below). For educational assessment tools, comparison of the tool against the list of objectives or course outline might be an approach to establish content validity. Content validity is also commonly established through a trial-and-error period of pilot testing, then revision, before finalizing the instrument.

*Predictive validity*, often referred to as criterion-related validity, uses the process of comparing the new measurement tool of interest to another, already accepted, criterion that measures the same variable.<sup>1</sup> Although the term “predictive” implies a relationship to future events, a predictive validity approach can use current, previous, or future events for the criterion standard.<sup>1</sup> A criticism of this approach is that if there is another tool that can be used as the “gold standard,” perhaps it should be used instead. However, use of the “gold standard” may not be appropriate or even possible in all research environments. For example, obtaining an arterial blood gas is not practical while in a transport helicopter. Pulse oximetry is easier and was validated, using simultaneous arterial blood gases as the criterion standard for oxygen saturation. The values were consistently close enough that pulse oximetry could adequately “predict” arterial O<sub>2</sub> saturation, and so it had high predictive validity. Another approach to establish predictive validity is to obtain the measure of interest at time 1 and then at a future time measure an appropriately related variable. For example, if the revised trauma score truly measures severity of injury, it should predict mortality. A demonstrated correlation between the trauma score and patient mortality would be evidence of predictive validity for the revised trauma score.

*Construct validity* is considered the strongest type of validity for self-report type data, such as surveys. Construct validity examines how well the selected instrument measures the theoretical construct of interest.<sup>2</sup> In the cases of more abstract variables such as pain, the challenge is greater in using a construct validity approach to establishing a valid measure. In establishing construct validity, the researcher must demonstrate that the operational definition of the identified variable chosen for the study matches the conceptual definition of the research question.

## Reliability

Validity is concerned with how accurately a data point measures what it is intended to measure. In contrast, *reliability* is the degree of consistency with which an instrument repeatedly measures the variable it is designed to measure.<sup>2</sup> Reliability is necessary to have high validity. If an instrument gives different results on two separate readings and the underlying variable has not changed, it is not measuring it with much precision or consistency. In contrast, an instrument can be very reliable and still not have validity (ie, it can be consistently wrong). For example, if a stable heart rate is measured

multiple times, and each time the same number is obtained, there is a high degree of reliability. But if a study uses heart rate alone to determine level of stress or pain, regardless of reliability, it is not a valid measure of stress. In general, establishing reliability is easier than establishing validity.

As with validity, there are several types of reliability (eg, stability across time, inter-rater 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 is the same at both times. For example, the height of an adult can be expected to remain the same for relatively long periods. To measure the stability of a ruler as measure of height, a measurement could be taken today and tomorrow. If the measurement stays the same, the ruler has stability across time.

Other variables are expected to change more frequently. Measuring a pulse today and tomorrow would not be appropriate to evaluate test/retest reliability, as pulse is expected to change over time. Instead, it would be better to take the two pulse measurements within a short period while the subject remains stable. In determining test/retest reliability, the length of time over which stability reasonably can be expected is critical.

*Inter-rater reliability* is the degree to which two or more evaluators agree on the measurement obtained. For example, to test inter-rater reliability of a blood pressure measurement, a double stethoscope could be used to determine whether both researchers agree on a single blood pressure value. This method is most important in assessing methods that have a greater degree of subjectivity (eg, patient mental status). Researchers using observational methods, especially with subjective end points, should examine inter-rater reliability before or while collecting initial study data, to determine the reliability of those measures. The Kappa statistic is often used to measure this type of reliability.<sup>3</sup>

A related term is *intra-rater reliability*, the degree to which the same person agrees with himself or herself on the measurement obtained. For example, would the same investigator obtain the same Glasgow Coma Scale score each time if repeated in a patient with an unchanged level of coma?

*Internal consistency* is more complex and is the degree to which items on a questionnaire or psychological scale are consistent with each other. It is a concept that is included in “content validity” of survey instruments (see previous discussion). Consistent questionnaires have items that are directed at measuring the same thing. For example, a scale to measure self-esteem would have several 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. Overly long questionnaires may not be completed, so 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. Further information can be found in the referenced text.<sup>3</sup>

A final form of reliability is *parallel forms*. This is when two instruments have been shown to measure the same variable. For example, more than one “form” may be needed to measure a given variable when you want to reduce cheating across students or repeat an examination or other instrument at a short interval. To ensure that the instruments are reliable (consistently measuring the same variable), the researcher needs to have one group of subjects complete both forms at the same sitting. A correlation between the two forms is performed to determine the degree of reliability, using correlation coefficients.

## Conclusion

Many factors can affect the quality of a research study. Errors in measurement can invalidate the study results and resultant conclusions. Steps can be taken in the design phase to minimize such errors. Although not all sources of error can be entirely eliminated, an attempt should be made to minimize them whenever possible. Not all of these concerns are an issue with every study design. Common sense can help the researcher identify potential sources of error in the research design. Review of the research proposal by experienced researchers often can be very helpful. Comparison of a proposed research protocol to similar published studies also may provide insight.

This part in the series is meant to discuss issues associated with designing a research protocol that relates to making study measurements. This is only an introduction to subject or measurement error. A number of reference textbooks are available for those who wish to learn more on this subject.

## References

1. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.
2. Burns N, Grove SK. Practice of nursing research: conduct, critique, and utilization. St. Louise, MO: Elsevier Health Sciences; 2005.
3. Waltz CF, Strickland OL, Lenz ER. Measurement in nursing research. 2nd ed. Philadelphia: F.A. Davis; 1991.

*Cheryl Bagley Thompson, PhD, RN, is an assistant dean of informatics and learning technologies and director of the health informatics program at the University of Nebraska Medical Center, College of Nursing, in Omaha, Nebraska. She can be reached at [cbthompson@unmc.edu](mailto:cbthompson@unmc.edu). Edward A. Panacek, MD, MPH, is professor of emergency medicine and clinical toxicology and director of clinical trials at the UC Davis Medical Center in Sacramento, California.*

1067-991X/\$30.00

Copyright 2007 by Air Medical Journal Associates

doi:10.1067/j.amj.2007.02.001