

Basics of research - part 4B: Sample size, data collection, and bias

CV Pollack and EA Panacek

Measurement collection methods

Once the design and subject selection procedures are determined the researcher must next consider how the variables of interest will be measured. Many different methods of data collection are available depending upon 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, non-exploratory 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 patency of a heparin lock"? In contrast, less structured methods may be appropriate for the question "what is the experience of being transported by ambulance for acute abdominal 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 reproducibility and reliability are usually higher when the measure can be quantified.

Obtrusiveness of the research protocol can impact the quality of the data obtained. An individual under scrutiny by a researcher may alter his usual behaviour, either for better or for worse. If

observation during a resuscitation 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 CO₂ 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 an investigation and vary in their degree of structure, quantifiability, researcher obtrusiveness, and objectivity. (Tables 1 & 2) In order 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 of 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 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.

Correspondence to:

Charles V. Pollack, Director

Marisopa Medical Center, Department of Emergency Medicine, Phoenix, Arizona, USA

UC-Davis Medical Center, Division of Emergency Medicine
Edward A. Panacek

*Part 4A of this article was published in the last issue of this Journal.

Biophysiologic measures are increasingly common

Table 1. Definition of terms

Biopsychologic measures	measures of biological function obtained through use of technology, such as electrocardiogram or haemodynamic monitoring
Self report	the variables of interest are measured by asking the subject to report on their 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 its 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 in which the measures are obtained at the same time
Predictive validity	criterion-related measurement using one instrument to to predict the value from another instrument at a future point in time
Content validity	concerned with whether the questions asked, or observations made actually address all of the variable of interest
Construct validity	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
Inter-rater reliability	the degree to which two or more evaluators agree upon the measurement obtained
Internal consistency	the degree to which items on a questionnaire or psychological scale are consistent with each other

Table 2. Data collection methods

	Quantifiability	Objectivity	Structure	Obtrusiveness
Biophysiological	XXX	XXX	XXX	XX
Self Report	XX	XX	XXX	XXX
Observation	X	X	Varies	Varies

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

with healthcare research. This trend is due partially to the increased technological nature of healthcare. Biophysiological measures include but are not limited to blood pressure, weight, and heart rate. Standards for the measure 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 measurements can be overly high reliance on their validity and reliability. 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 0.1 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 correctly used and with increased complexity it may be more difficult to detect equipment malfunction.

"Self report" data also is common within the healthcare environment. Self-report data is 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 behaviours. 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 amendable to observation but the subject 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 measurements 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 because of the ease of development and ease of 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.

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 advantage of this approach is that usually the validity and reliability of the instrument has been previously established, a method of 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 audio or videotape. The investigator then analyses the episode for the variables of interest. For example, 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 subjective appraisal may be used in order to minimise 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 analysed 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 legitimate effort to increase objectivity. The presence of an observer or a recording device may make the subject more aware of their actions, causing an alteration in their behaviour.

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 absolutely known. An obtained score is always 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." Biophysiological measures have "relatively" high validity because the measurement technique may be based on the definition and upon sound scientific principles. For example, blood pressure is the pressure in the cardiovascular system. The

measurement of pressure is a relatively straightforward process. In contrast, development of a valid tool to measure pain is much more difficult. Not everyone agrees on a definition of pain so developing a tool to address a nebulous and highly subjective entity is much more difficult. The researcher might question if the tool measures pain or if it really measures something else such as anxiety.

Validity is very difficult to insure because absolute knowledge cannot be obtained. Researchers use several "round about" 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 only means that the instrument looks like it is measuring what it should be measuring⁴ and is an intuitive and subjective judgement. At a very minimum a tool must have validity, but as this is the weakest test of validity, other approaches also should be used to establish validity.

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 transport blood gases are not available, so the less invasive pulse oximetry might be the only way to obtain SaO₂ 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 then at a future time another

criterion is measured. The philosophy behind this approach is that if X leads to Y with a certain frequency, and one measures X, then one 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 directly measured in most cases as is possible with criterion related validity. Establishment of content validity relies most commonly 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. Secondly, content validity is more concerned with the question of whether everything is covered and nothing left out. As a result, content validity uses more specific and objective criteria.

Construct validity is perhaps the most difficult type of validity to understand and to measure. Establishment of construct validity is a more 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 Hungler, 1995.)

Instrument reliability

In contrast to validity, **reliability** is the degree of consistency with which an instrument measures the variable it is designed to measure.⁴ 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, at least one of those times 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 one takes a blood pressure multiple times and each time it is the same, that is a reliable measure. But if one is 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: 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. 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 measurement could be taken today and another in a month. If instead the measure could be expected to change more frequently, such as weight, 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 1-2 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 can reasonably be expected.

Inter-rater reliability is the degree to which two or more evaluators agree upon the measurement obtained. For example, to test inter-rater reliability of a blood pressure measurement a double stethoscope would be used to determine if 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 inter-rater reliability before collecting study data to assure that everyone is looking for the same behaviours or measurements.

Internal consistency is a little 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 all directed at measuring a component of self esteem. Achieving a questionnaire with internal consistency is a balancing act. The goal is to arrive at a questionnaire that is consistent without being redundant. Long questionnaires may not be completed and 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. A discussion of the two methods is beyond the scope of this series.⁴

A final form of reliability is parallel forms. Parallel forms is an examination of two instruments used to measure the same variable. For example, one may not want all students to have the exact same test if they are sitting very close to each other when taking the exam. You also may want to repeat the exam at a very short interval and do not want subjects to remember questions from the first time. To assure 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 then 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 an eye to common sense will help the researcher to 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 where possible.

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

This article in the series is meant to discuss the many issues associated with "fleshing out" a research protocol. The subject can become very 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. It is important to understand the importance of the planning phase of a study, which 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 upon 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: NJ Lawrence Erlbaum Associates, 1988.
2. Heberlein TA, Baumgartner, R. Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *Amer Soc Rev* 1978;43:447-62.
3. Cohen J, Cohen P. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Hillsdale: NJ Lawrence Erlbaum Associates, 1983.
4. Polit DG, Hungler BP. *Nursing Research: Principles and Methods*, 5th edition. Philadelphia PA: JB Lippincott Company, 1995.

Recommended Texts

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