

# 4

## Measurement theory and practice

*Rajender R Aparasu*

### *Chapter objectives*

- To explain the concept of measurement
- To discuss levels of measurement
- To describe reliability and methods to evaluate reliability
- To describe validity and methods to evaluate validity

### **Introduction**

Measurement forms the basis for empirical health services research including pharmaceutical practice and policy research. The measurement process is designed to record and capture the underlying construct and concept. In healthcare, critical policy and practice discussions are based on cost, quality, and access. These concepts or constructs have to be measured and analyzed using variables to evaluate the performance of a healthcare system at the patient, family, institution, and population levels. Each of these constructs is complex and often requires multiple measures or variables to capture the underlying concepts. Advances in measurement of each of these dimensions have led to improvement in the healthcare system. For example, the development of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM), diagnostic-related groups (DRGs), and Healthcare Common Procedure Coding System (HCPCS) has led to changes in healthcare delivery and reimbursement. In recent years, measurement of quality of healthcare based on structure, process, and outcome has gained significance due to its importance in healthcare delivery (Donabedian 2003).

The measurement process involves a systematic assignment of values or numbers to observations based on *a priori* rules of measurement (Viswanathan 2005). It is a critical step in quantitative research because it also defines the subsequent steps in conducting research such as analysis and

interpretation of the research findings. The measurement process involves recording observations that are manifestations of the underlying construct; this requires a good understanding of the construct and the measurement process. Often the variables are operationalized based on the methodology used to capture these variables. This can involve collection of data by the researcher, also called primary data, or data collected by others for reuse, also referred to as secondary data. The operationalization and collection of data are critical in the measurement process. Further evaluation of this measurement process will confirm whether the process is truly measuring the construct, which entails ensuring that the measurement process is reliable and valid. This chapter describes the concept and levels of measurement, and discusses reliability and validity, and methods commonly used to evaluate reliability and validity.

## Nature and level of measurement

The measurement process for any construct is based on the existing knowledge base regarding the construct. The measurement of physiological constructs such as blood pressure and blood glucose is often standardized. However, behavioral constructs such as compliance are rather complicated because the observed behavior may or may not reflect the intended construct. An understanding of the measurement process for behavioral constructs can provide a good framework to understand the measurement process in general. Summers (1970) suggests that the measurement process of an abstract construct such as behavior that is not directly observable involves three interlinked operational subprocesses:

- 1 Identification of acceptable behavior specimens that represent the underlying construct
- 2 Data collection of specimens
- 3 Conversion of specimens to a quantitative variable.

An acceptable specimen defines the data collection process, which in turn determines the type of quantitative variable. Often operational definitions, especially measured operational definitions, are based on these three operational subprocesses.

## Acceptable behavioral specimen

Identification of acceptable specimens that represent the underlying construct is a critical step in the measurement process. The current knowledge base defines ones that are acceptable and those that are not. Acceptable specimens for diagnosis of diabetes such as blood glucose and glycated hemoglobin are often standardized. The measurement process for constructs such as

compliance is complex. However, the concept of compliance with a medication regimen is important in evaluating treatment effects. Compliance refers to the extent to which the patient follows healthcare advice (Hayes et al. 1979). Not all patient behaviors that reflect compliance with a medication regimen are easily observable by researchers and clinicians. Compliance with a medication regimen can include behaviors that can capture the extent to which the patient is taking medication with respect to frequency and duration of therapy. The behavior specimens such as refill history which capture frequency and duration of therapy are considered acceptable for measuring compliance (Farmer 1999). Other medication-taking behaviors can include time of administration and avoidance of certain foods.

### Data collection

The data collection process entails capturing data based on specimens. Often multiple data collection processes can be linked to the behavior specimens. With respect to compliance, the data collection can involve self-reports of specimen behavior or observation of overt behavior such as refill history (Farmer 1999). The self-report methodology involves measurement of patient responses to a series of questions or items using a survey instrument to determine medication-taking behaviors. Self-report methodology is based on the assumption that the patients can report their medication use behavior. Overt behaviors of compliance can be assessed by direct observation of patient behavior or refill history as captured in prescription claims data. Refill history measures, such as medication possession ratio, reflect the number of doses filled by the patient for the dispensing period. Refill history is based on the assumption that prescription-filling behavior reflects the patient's medication use pattern. Compliance measurement based on biochemical instruments to evaluate drug levels can also be used as a behavior specimen (Farmer 1999).

### Assignment of values

Converting the behavioral specimen to a variable involves assignment of values based on the rules developed for data collection to evaluate the underlying concept. The assignment process should reflect variation in the underlying construct. For example, the value assigned to a compliance measure, based on self-reports or refill history, should reflect the extent to which the patient is following the healthcare advice for medications. Measurement and further analysis are dependent on the properties of the variable. The levels of measurement can be classified as nominal, ordinal, interval, and ratio, based on the properties of the quantitative variable. This is also referred to as Stevens's (1946) scales of measurement.

The nominal measure, also called the categorical measure, is the simplest and lowest form of measurement. It is used for naming or identification purposes. Examples include measurement of gender (male or female) and ethnicity (white, black, etc.). The subgroups in a nominal variable are mutually exclusive and exhaustive. All members of a group have the same identity and there is no relationship among the subgroups. For nominal measures the assignment of numbers, letters, or symbols is only for labeling and grouping purposes. Although numbers are often used for labeling or coding purposes, they do not have an arithmetic value. Consequently, none of the mathematical manipulations can be used for nominal measures. The nominal measures are often used for counting and to examine frequency distribution. The number of subgroups for classification or labeling is based on the extent of identification needed. For example, the respondents' residences can be grouped by state (50 states) or by region (northeast, midwest, south, and west). Nominal variables can be used as dichotomous variables where only two subgroups are recorded such as white and nonwhite for race.

The ordinal measure is rank ordering of a group membership with properties of transitivity. The group membership can be ordered from lowest to highest with a clear interrelationship of the levels, unlike the nominal measure. For example, the health status of a patient can be measured with levels or responses of excellent, very good, good, fair, or poor. Similarly, a patient's perception of compliance can be measured using a five-point scale from most compliant to least compliant. Ordinal measures are often used in survey instruments, and can be ordered from highest (excellent) to lowest (poor). In addition, the rank order relationship levels are mutually exclusive and exhaustive, similar to the nominal measure. In addition the interrelationship of the levels is known. The distance between the levels is, however, not equal and, at times, not known. The levels have properties of transitivity, i.e., if excellent is better than very good and very good is better than good, this makes excellent better than good.

The interval and ratio measurements have properties of equality of intervals in addition to characteristics of an ordinal measure. The key difference between interval and ratio measures is that there is an absolute zero level in ratio measure. Ratio is the highest form of measurement because it can represent the zero value of the measure. Mathematical manipulations such as multiplication and division can be used for ratio measures. Examples of ratio measures include prescription expenditure, hospitalization days, wait time, and weight gain. The best example of an interval measure is temperature measured on the Fahrenheit scale. The temperature is measured using the equidistant scale of Fahrenheit in which there is no zero amount or quantity of temperature; the assignment of zero temperature is arbitrary. The intelligence quotient (IQ) scale is another example of an interval measure where the distance between the scale values is the same. In psychometric research, it is

difficult to find a measure that has an absolute zero. The interval measures allow mathematical manipulations. In general, ratio and interval measures are used in the same way for statistical analysis. Ratio and interval measures provide great flexibility for mathematical manipulations as long as the original properties of measurement are satisfied. With respect to compliance, the number of missed doses is a ratio scale that can be obtained from patients using self-reports.

## Measurement issues

Ratio and interval measurements are the most desirable measures in health services research because they capture more information about the underlying construct, and are therefore more likely to reflect the variation in the underlying construct, than other ordinal or nominal measures. Ratio and interval scales also have all the necessary properties required for mathematical manipulations. Most importantly, these measures can be converted to low forms of measures such as ordinal measures. It is not possible to convert ordinal measures into ratio/interval measures. For example, family income captured in dollars (ratio measure) can be grouped into high, middle, or low income (ordinal measure) based on specific ranges of family income. However, it is not possible to convert family income captured as an ordinal measure into a ratio measure. In general, higher forms of measures should be preferred to lower forms of measures.

The goal of the measurement is to assign values to a variable based on the specific rules formulated to measure the underlying construct or concept. The measurement process is also designed to capture variation in the underlying construct. The values assigned in the measurement process are distinct due to the differences in the underlying construct or concept. If variations in the underlying construct or concept are not reflected in the measurement, it leads to measurement error, which is an error in the measurement process (Viswanathan 2005). Sources of measurement error can be due to the measurement process itself or factors outside the measurement process. For example, a measurement error may occur if the number of missed doses is captured using self-reports in the presence or absence of a physician. A good measurement process is designed to reduce such measurement errors. These issues are considered in detail in the discussion of reliability and validity.

Measurement and statistical analyses are based on the concept of analysis of the variation. The type of measurement of variables determines the statistical analysis to be used. This holds true for both dependent and independent variables. The interval and ratio measures capture greater variation in the dependent measures than the ordinal measures. Consequently, there is more opportunity to explain the extent of variation in interval and ratio measures

than ordinal measures. This can also improve to capture the sources of variation in statistical analysis.

## Reliability and validity

The goal of the measurement process is to ensure that the values assigned to variables are reliable and valid. Reliability and validity are different dimensions of the measurement process. Reliability ensures that the assignment of values is consistent or reproducible, whereas validity ensures that the assignment of values truly reflects the underlying construct or concept (Bohrnstedt 1970; DeVellis 1991; Trochim 2001). Both reliability and validity are important in the measurement process because reproducibility of a measure as well as the trueness of a measure is critical in research. Reliability of a measure does not, however, ensure its validity. For example, use of self-reports to capture compliance with medication may be reliable but may not be valid. Self-reports of missing doses may consistently provide the same measures but may not truly capture the patient behavior. The constructs in sociobehavioral research are often abstract and hence require evaluations of reliability and validity of the measurement process. Psychometric research has played a significant role in the evaluation of reliability and validity. Some of the reliability and validity evaluation methods are specific to survey research involving a survey instrument. However, the concepts of reliability and validity are relevant for both behavioral and nonbehavioral constructs.

## Measurement errors

Both reliability and validity of a measurement are affected by measurement errors. Measurement errors can be classified as random errors and nonrandom errors (Viswanathan 2005). Random or chance errors, as the term suggests, occur inconsistently, and cause the measures to deviate randomly from the true value. Random errors negatively affect the reliability of the measurement and are present in every measurement process. The goal of a measurement process is to minimize random errors and, thereby, maximize the reliability. The factors that influence reliability can be the individual, instrumentation, and environment. Individual or patient level factors include diurnal variation, education level, or biological variability. Instrument level factors include calibration of instrument, or misreading or mistakes in recording questionnaire responses. Environmental factors influence individual and instrumental factors, such as temperature, pressure, light, or electrical fluctuations. An ideal measurement process will minimize the influence of these factors to maximize the reliability. This can be achieved using standard and consistent data collection and administration methods. For example, the measurement can

be recorded at one specific time for all participants to minimize diurnal variation.

Nonrandom or systematic errors occur consistently by definition and hence cause the measures to deviate from the true value nonrandomly (Viswanathan 2005). The amount of systematic error directly influences the validity of the measurement. It is inversely related to validity: high nonrandom error decreases the validity of a measurement process. Systematic bias is a classic example of a nonrandom error that threatens the validity of a measurement process. Random bias does not influence the validity of a measurement.

There are different types of biases such as information, recall, and interviewer bias. Information bias occurs when there is consistently differential information among the participants of interest due to underlying factors. For example, in epidemiological research, test cases tend to provide more information than control cases that do not have the disease. Recall bias occurs where there is a differential ability to recall information about previous experience due to issues related to time or experience. For example, the measurement error in reporting the number of missed doses is likely to be less than the number of doses actually taken according to the instructions due to recall bias. Interviewer bias exists when interviewer perception or behavior influences the responses. The measurement process should preferably control for systematic biases at both data collection and study design stages to strengthen the validity. For example, blinding techniques are used to hide the procedural aspects of study design from respondents and interviewers to minimize the bias.

## Reliability

Reliability, as discussed earlier, ensures that the measurement is consistent or reproducible (Bohrnstedt 1970; DeVellis 1991; Trochim 2001). A reliable measurement process will consistently provide the same or a similar value for an unchanged variable, whereas a change in the underlying construct will reflect a change in the value assigned. According to Bohrnstedt (1970), the reliability assessments can be grouped into two major classes: measurement of stability and measurement of equivalence. Measures of stability evaluate relationship or correlation of measures across time or evaluators, and examples include the test–re-test method and interrater reliability. Measures of equivalence evaluate relationship or correlation between two sets of instruments, and examples include split-half, parallel form, and internal consistency methods.

### *Measures of stability*

The test–re-test method analyzes measures obtained from the same participants across time using the same instrument. It evaluates the stability of

measurements over time. The instrument for measurement can be any equipment such as a weighing scale or a survey instrument with a series of questions. For example, weight measured using a weighing scale at one time can be correlated with weight obtained after 1 hour. The time difference between the measurements is dependent on the type of measure. A longer time interval can influence the reliability for some measures due to changes in the underlying construct such as a person's weight. For other measures involving examinations, shorter time can influence the reliability due to knowledge of the previous test. The general rule is that the time interval should be long enough that respondents do not remember their responses without the change in the underlying construct. The correlation between the two measures on the same participants provides the correlation coefficient. This method is often used to ascertain the reliability of physiological measurements such as blood glucose and blood pressure. In fact, diagnostic criteria for a disease are often based on the test–re-test method. Pearson's correlation is used to calculate the correlation coefficient for interval–ratio measures, whereas Spearman's rank correlation is used for ordinal measures. In general, the correlation coefficient of  $>0.80$  is considered as a reliable determinant (Nunnally and Bernstein 1994). Disadvantages of the test–re-test method include inconvenience and reactivity. Due to multiple measurement processes it is inconvenient to participants and researchers. Reactivity refers to change in the underlying construct due to testing. For example, patients responding to a question related to compliance are sensitized to the issue of compliance and thus provide responses that might reflect an improvement in compliance.

Interrater reliability involves analysis of measures obtained from the same participants by different evaluators using the same instrument. For example, pharmacists' counseling time can be measured by two independent observers. The reliability of such measures can be evaluated by correlating the time measures obtained from the two observers. Cohen's  $\kappa$  coefficient is used for reliability involving two evaluators, whereas Fleiss's  $\kappa$  is used for measures involving multiple evaluators (Landis and Koch 1977). These measures are calculated based on the difference between percentage of agreement among evaluators and probability of chance agreement.

### *Measures of equivalence*

The split-half method is one of the earliest measures of equivalence to determine questionnaire reliability. This method involves dividing the number of items or individual questions of a survey instrument into two equivalent halves; the correlation between the two halves provides the correlation coefficient. There are two options for dividing the items or questions: one method involves dividing the items into even and odd questions; the other involves dividing the items into first and second halves. The decisions on the type of



splitting are based on practical considerations and type of items in the survey instrument. Irrespective of the approach, the split-half method provides two measures on each participant based on two equivalent-form measures, and the measures from the halves are then used to calculate the correlation coefficient. The basic underlying principle in the split-half method is that the two halves are designed to measure the same underlying construct. The strength of the split-half method is that it overcomes the problems of test–re-test methods such as reactivity and inconvenience. The weakness of the split-half method is that two halves must measure the same underlying construct and the reliability coefficient can vary based on the approach used to divide the items. The Spearman–Brown prophecy formula is generally used to obtain the correlation coefficient between the split halves (DeVellis 1991). The formula is based on the correlation coefficient between the split halves and correction needed to divide the items in the survey instrument into two halves.

The parallel-form method is an extension of the split-half method in which two parallel questionnaires are administered to each participant consecutively. The time interval between administration of the two questionnaires should be minimal to optimize changes in the underlying constructs. The scores from the two forms or questionnaires are used to calculate the correlation coefficient. Similar to the split-half method, the two forms should be equivalent and measure the same underlying construct but should not be identical. This method also overcomes the problems of test–re-test methods such as reactivity and inconvenience. However, it may be cumbersome to the respondents to complete two parallel questionnaires. In general it is easier to develop similar items in the split-half method than to create parallel forms. Consequently, the parallel-form method is seldom used in health services research but often used in educational research involving examinations due to experience in creating parallel exams and availability of a large pool of questions.

The internal consistency method is the most frequently used method in health services research. It involves correlation among all items or questions in a questionnaire without the need to divide items or create forms. The internal consistency method evaluates whether all items in a questionnaire are measuring the same construct. It also overcomes problems associated with split-half and parallel-form methods, such as varying reliability due to the process used to divide the items or create the forms. Internal consistency is based on the concept that items or questions designed to measure the same underlying construct should be highly correlated. This means that each item or question is used to compare consistency of responses with other items in the questionnaire for the study sample. As a result, the correlation coefficient is sample specific. There is a need to assess the reliability of the survey instrument with a change in the study sample. There are several ways to compute a reliability coefficient based on the internal consistency approach. Cronbach's  $\alpha$  is used

to calculate the internal consistency of measures based on continuous measures (Bohrnstedt 1970; DeVellis 1991). The reliability coefficient increases with an increase in the number of items and inter-item correlations. The Kuder–Richardson coefficient (KR20 or KR21) is used to calculate reliability for nominal measures. These coefficients are calculated based on the proportion of same responses for an item. High conformity of responses leads to high KR20 or a reliability coefficient of nominal measures. Cronbach's  $\alpha$  is an extension of KR20 and both calculations are based on classic test theory. Most statistical packages can calculate these reliability coefficients.

## Validity

Validity, as mentioned earlier, ensures that the instrument developed for measurement purposes truly represents the underlying construct (DeVellis 1991; Nunnally and Bernstein 1994; Trochim 2001). In addition to appropriateness, instruments are increasingly being evaluated for meaningfulness and usefulness, in recent years, in order to strengthen the validity. Although methods to ascertain validity have changed over the years, construct validity has remained the cornerstone of all types of validity assessments. Construct validity refers to the degree to which the instrument measures the underlying construct. The evidence to strengthen the construct validity is based on internal structure and external relationships. The internal structure evaluates the interrelationship of the items and underlying dimension of the construct. The external relationship evaluates the relationship of the instrument to other constructs.

The internal structure of the instrument should be consistent with the underlying dimensions of the construct. Factor analysis is commonly used in the development process of a survey instrument (DeVellis 1991; Nunnally and Bernstein 1994). Exploratory factor analysis helps to identify various factors or dimensions that are represented in the instrument. It also groups items that belong together representing the underlying construct. For example, this can be used to establish the dimensions of a quality-of-life scale. Exploratory factor analysis is usually followed by confirmatory factor analysis to determine the extent to which statistical validity is based on the underlying theoretical model. Although factor analysis approaches are useful, the internal structure should not be the only basis of construct validity. The relationship to other constructs also ascertains validity of the construct. External relationships should be empirically tested for hypotheses developed based on the theoretical relationships. This addresses the evidence of predictability and discernibility.

Trochim (2001) proposed that all types of validity testing methods should strengthen the construct validity of an instrument. This can be achieved using

translational and criterion validity. The translational validity addresses the translational or transformational aspect of construct validity, which includes face and content validity. These validity analyses are designed to ensure that the items in the instrument reflect the underlying construct. Face and content validity will ensure that the items represent the intended factors or dimensions of the construct, which can also be confirmed using exploratory factor analysis. Criterion validity refers to the relationship aspect of construct validity, and includes concurrent, predictive, convergent, and discriminant validity. Methods for criterion validity empirically test for theoretical relationships. The translational validity and the criterion validity ensure appropriateness, meaningfulness, and usefulness of the instrument. Multiple methods are needed to strengthen the construct validity of the instrument.

### *Translational validity*

Face validity is the simplest method to ensure translational validity. It addresses the question: Do these items and the overall instrument measure the underlying construct? This involves a judgment or an opinion of a layperson or an expert. A layperson will provide his or her perspective, mainly to address the issues related to the interpretation of items and administration of the instrument. Experts can provide detailed opinion about the appropriateness and wording of items, organization of items, and the overall look of the instrument with respect to the underlying construct. It primarily involves qualitative and subjective assessment of the instrument. Consequently, it is considered as the weakest form of validity assessment. It is often used in the development process to refine an instrument.

Content validity refers to the representative nature of the items to capture the underlying dimension of the construct. It presents the relationship of items to the dimensions of the underlying construct. The content validation process requires a clear definition of the dimensions of the underlying construct and ways to ensure that the selected items represent the relevant dimensions. The dimensions of the construct can be defined from the literature or expert opinion. For example, measurement of healthcare quality requires items or questions related to the structure of the healthcare system, process of obtaining healthcare, and outcomes of healthcare obtained (Donabedian 2003). A content expert's opinion can be sought to evaluate whether the measurement items represent the defined dimensions. Although there is some subjectivity in the process, seeking the opinions of multiple experts can reduce subjectivity and improve the face validity. Also, analytical measures such as content validation ratio, content validation form, and content validation index can strengthen the content validity (Shultz and Whitney 2005). Content validity will ensure that the items and the overall instrument reflect the dimensions of the construct.

### *Criterion validity*

Criterion validity addresses the relationship aspect of the construct validity by attesting to the relationships between the instrument and criterion, or other measures, based on theory and practice. Selection of the criterion plays an important role in criterion validity. The evidence and the extent of the inter-relationship of criteria strengthen or weaken the construct validity. A strong relationship means that the criterion is well validated and accepted. Often the criterion selected is external and considered the “gold standard.” The theoretical and practical knowledge about the issues are critical in selection of the criterion. For example, compliance measured using self-reports or refill history can be validated using the criterion of drug levels in blood or urine because these are considered the gold standard. As mentioned earlier, types of criterion validity include concurrent, predictive, convergent, and discriminant validity.

Concurrent validity is a type of criterion validity that refers to the relationship between the instrument and the criterion measured at the same point in time. The criterion selected for concurrent validity should measure the same underlying construct as that of the instrument. The selected criterion should be a standard measure. The rationale for concurrent validity is that, if the instrument and the criterion are administered at the same time and measure the same underlying construct, then there should be strong correlation between the two measures. For example, compliance measured using self-reports can be validated by comparing responses with the drug levels in blood or urine measured at the same time.

Predictive validity is a type of criterion validity that addresses the relationship between the instrument and criterion measured at a future time. The criterion and the instrument are not measuring the same underlying construct as in the concurrent validity. However, the instrument should be able to predict the criterion. For example, compliance measures based on refill history have been shown to predict the health expenditure in patients with diabetes. This is based on the hypothesis that disease is managed better in compliant than in noncompliant patients, and thus leads to a decrease in healthcare expenditure. The rationale for predictive validity is that, if the patients with diabetes are compliant, they will incur less expenditure due to better disease-state management. Therefore, the criterion selected for predictive validity should be based on theory and practice.

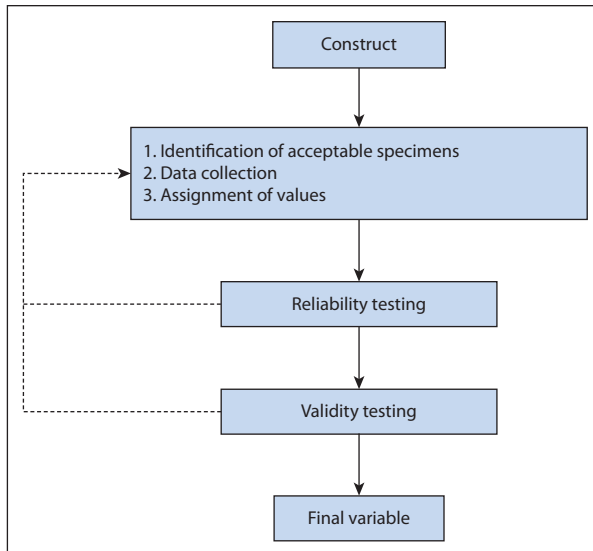
Convergent and discriminant validity are two sides of the same concept. Convergent validity refers to convergence or a strong relationship between the instrument and the criterion, which are theoretically similar. Discriminant validity refers to little or no relationship between the instrument and the criterion, which are theoretically different. Convergence validity is similar to concurrent validity but it is not restrictive with respect to time of

administration. The concept of convergent and discriminant validity is based on the principle that, if the instrument is valid, it will be strongly correlated with measures that are similar and will not be associated with measures that are dissimilar. For example, instruments to measure pain and overall quality of life will be strongly associated because they measure similar concepts. Conversely, pain measures are less likely to be associated with perceptions on economy as these are dissimilar concepts.

## Measurement process and practice

Understanding the measurement theory and practice is vital in conducting empirical research. An existing knowledge base defines the methodology to operationalize a construct, which includes identification of acceptable behavior specimens, data collection, and assignment of values to develop the variable for an underlying construct. Some constructs are easy to measure, such as expenditure; other constructs, such as quality, need a strong understanding of underlying theory in order to operationalize. Prescription expenditures are usually captured using secondary data sources such as claims data. The behavior specimen is reflected in claims data and it captures payment by insurance companies and other sources. The data collection involves use of secondary data and value assignment of expenditures involves a ratio scale. Other constructs such as quality are complex and require significant effort. The existing knowledge base suggests that quality measurements should be based on the underlying dimensions. For example, quality of medical care is based on measures of structure, process, and outcome (Donabedian 2003). The operational definitions and measurements processes for each of these dimensions are different. According to Donabedian (2003), measures of quality are relevant only when there is an interrelationship for structure, process, and outcome. Consequently, greater understanding of the underlying theory and advances in data collection methodologies play an important role in the measurement process and operationalization of healthcare constructs.

Reliability and validity issues are critical in the identification of behavior specimen and data collection phases. Concepts of reliability and validity are valuable to develop and improve a strong research instrument. Accordingly, they should be considered tools for continuous quality improvement of the measurement process. A well-defined construct helps to ensure the construct validity of an instrument, which includes translational and criterion validity. Translational validity includes face and content validity. The considerations of face and content validity can be incorporated in the identification of the behavior specimen phase to ensure that instrument development is consistent with the underlying dimensions of the construct. The criterion



**Figure 4.1** Measurement process and practice.

validity is difficult to incorporate into the behavior specification phase. In general the considerations of translational validity are likely to ensure criterion validity. Criterion validity can be tested only by correlating measures from the instrument with a criterion. The test findings will reveal needed improvements in the development phase. Figure 4.1 provides a schematic diagram for the measurement process and practice.

In the data collection phase, the developed instrument is utilized to collect the data. Various research techniques can be employed to ensure reliability and validity of data collection. Standardized data collection and administration methods can minimize random errors and improve reliability. Pre-testing of the instrument can help to identify items that require clarification. It can also help to improve the organization of the instrument. Response biases can be minimized using pre-tested items, techniques of blinding, utilization of trained interviewers, and consistent data collection methods. The data collection process should minimize random errors and control for nonrandom errors to maximize reliability and validity. The tests for reliability and validity will ensure that the instrument developed and utilized for research is reliable and valid.

## Summary and conclusions

The measurement process is designed to record and capture the underlying construct. This involves identification of acceptable specimens, data

collection of specimens, and conversion of specimens to a quantitative variable. The decisions made at each of these interrelated steps are based on the existing knowledge base of the construct. Reliability and validity issues are critical in the measurement process. Reliability addresses stability and equivalence of the measurement process. The tests of translational and criterion validity are designed to ensure construct validity of the measurement process. Construct validity addresses the extent to which the variable measures the underlying construct. A reliable and valid measurement process will minimize measurement errors and thereby strengthen the research. Measurement also forms the basis of subsequent steps in research such as statistical analysis.

### Review topics

- 1 Discuss levels of measurement using examples.
- 2 Describe the concept of reliability and methods to evaluate reliability.
- 3 Discuss common types of measurement errors.
- 4 Describe the concept of validity and methods to evaluate validity.
- 5 Describe the measurement process using an example in pharmaceutical practice and policy research.

### References

- Bohrnstedt GW (1970). Reliability and validity assessment in attitude measurement. In: Summers GF (ed.), *Attitude Measurement*. Chicago, IL: Rand McNally, 80–99.
- DeVellis RF (1991). *Scale Development: Theory and applications*. Thousand Oaks, CA: Sage Publications.
- Donabedian A (2003). *An Introduction to Quality Assurance in Health Care*. New York: Oxford University Press.
- Farmer KC (1999). Methods for measuring and monitoring medication regimen adherence in clinical trials and clinical practice. *Clin Ther* 21: 1074–90.
- Hayes RB, Taylor DW, Sackett DL (1979). *Compliance in Health Care*. Baltimore, MD: Johns Hopkins University Press.
- Landis JR, Koch GG (1977). The measurement of observer agreement for categorical data. *Biometrics* 33: 159–74.
- Nunnally JC, Bernstein IH (1994). *Psychometric Theory*, 3rd edn. New York: McGraw-Hill.
- Shultz KS, Whitney DJ (2005). *Measurement Theory in Action: Case studies and exercises*. Thousand Oaks, CA: Sage Publications.
- Stevens SS (1946). On the theory of scales of measurement. *Science* 103: 677–80.
- Summers GF (1970). Introduction. In: Summers GF (ed.), *Attitude Measurement*. Chicago, IL: Rand McNally, 1–21.
- Trochim WMK (2001). *The Research Methods Knowledge Base*, 2nd edn. Cincinnati, IL: Atomic Dog Publishing.
- Viswanathan M (2005). *Measurement Error and Research Design*. Thousand Oaks, CA: Sage Publications.

## Online resources

Agency for Healthcare Research and Quality (AHRQ). “National Quality Measures Clearinghouse.” Available at: [www.qualitymeasures.ahrq.gov](http://www.qualitymeasures.ahrq.gov).

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics. Series 2. “Data Evaluation and Methods Research.” Available at: [www.cdc.gov/nchs/products/series.htm#sr2](http://www.cdc.gov/nchs/products/series.htm#sr2).

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics. Series 6. “Cognition and Survey Measurement.” Available at: [www.cdc.gov/nchs/products/series.htm#sr6](http://www.cdc.gov/nchs/products/series.htm#sr6).

The Leapfrog Group. Available at: [www.leapfroggroup.org](http://www.leapfroggroup.org).