Chapter

3 Statistical Essentials II: Measurement

Learning Objectives

From this point on, we will be building on earlier principles. In this chapter, we will move on to statistical concepts related to measuring factors that are of interest to clinicians and investigators. This chapter will prepare you to:

- Distinguish between independent and dependent variables
- Differentiate the levels of measurement and understand why this idea is important in statistics
- Discuss the application of levels of measurement to evidence-based practice
- Define instrument reliability and validity and discuss the importance of addressing them prior to measuring variables with an instrument
- Understand how the quality of research is influenced by measurement decisions

Key Terms

Categorical variables
Confounding variable
Construct
Construct validity
Content validity
Continuous variable
Criterion-related validity
Data
Data set
Dependent variable
Discrete variable
External validity
Independent variable
Instrument
Internal consistency
Internal validity
Interrater reliability
Levels of measurement
Measurement error
Qualitative variables
Quantitative variables
Reliability
Test–retest reliability
Tool
Validity
Variable

VARIABLES AND DATA

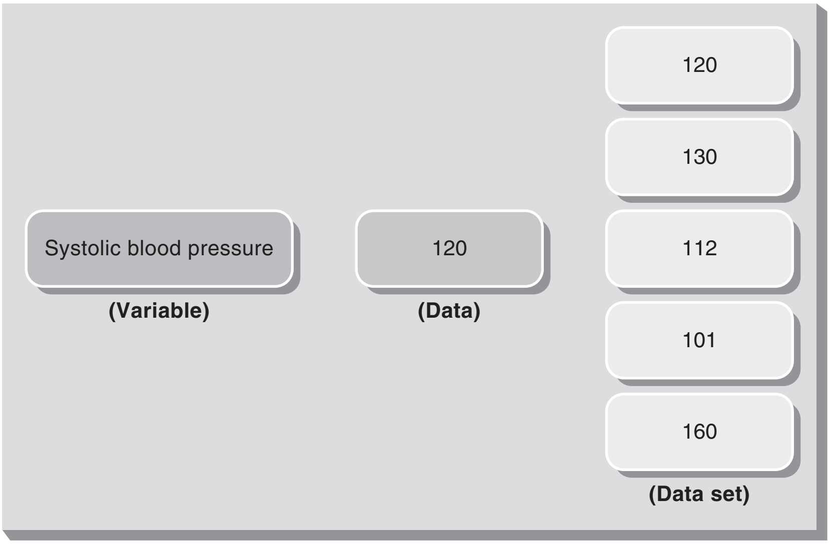
Investigators define variables and collect data to answer their research questions. A **variable** is a trait or characteristic that varies or changes, and **data** are the values of variables when they vary. For example, systolic blood pressure is a variable because it is a characteristic that fluctuates, both from one person to another and at different times within the same person (**FIGURE 3-1**). Each blood pressure measurement is a data value. A collection of these data values is a **data set**.

Investigators classify variables according to a variety of characteristics and these suggest how variables may be used in research and evidence-based practice. Variables may be *qualitative* or *quantitative.* **Qualitative variables** have values that are nonnumeric, while **quantitative variables** have values that are numeric. Systolic blood pressure may be seen as either qualitative—high, normal, or low—or quantitative—120 mmHg. In this text we will confine ourselves to numeric variables, as these are amenable to statistical analyses.

Numeric variables may be either discrete or continuous. **Discrete variables** have values that are countable but do not include numeric

Figure 3-1

Variables and data.



values, such as fractions, between countable categories. **Continuous variables** have every possible value on a continuum. Gender may be counted, as in there are 10 women in a waiting area, and is an example of a discrete variable; this is because, in the real world, there is no such thing as 10.5 women. On the other hand, systolic blood pressure ranging from 0 to 200 mmHg is an example of a continuous variable since we could measure the pressure anywhere between 0 and 200 mmHg. Conceivably, we could measure systolic blood pressure to the nearest hundredth—120.05 mmHg.

Variables can also be characterized as independent or dependent if an investigator is investigating the interaction between variables for statistical hypothesis testing. The variable that is manipulated by the investigator or affects another variable is the **independent variable**. The **dependent variable** is affected by an independent variable. For example, suppose an investigator is examining whether hepatitis B antigen affects liver function test results. The presence or absence of the hepatitis B antigen is the independent variable, as it affects the liver function test results, and the liver function test result is the dependent variable, as it is affected by hepatitis B antigen.

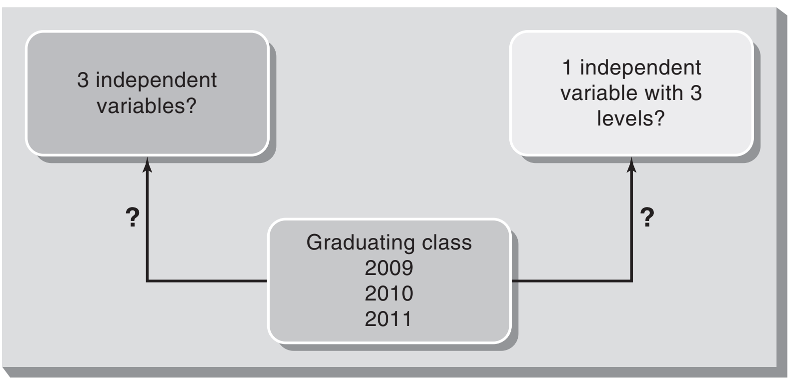
Consider another example. An investigator is studying the effectiveness of a newly developed medicine to treat constipation. The investigator devises an experiment in which the treatment group receives the new medication and the control group receives a placebo. The investigator measures the number of days between taking the drug and the first bowel movement among participants in both control and treatment groups. Here, the group assignment—treatment or control—is the independent variable since it is manipulated by the investigator and it affects the length of time until the first bowel movement. The number of days until the first bowel movement is the dependent variable, since it is affected by the group assignment or whether the participant received the new drug or the placebo.

In the previous two examples, it seems pretty clear how independent variables differ from dependent variables. However, determining the number of independent variables in a study can be confusing. Suppose an investigator is studying the licensure examination passing rates of graduating nurses from classes 2009, 2010, and 2011 at a public university. In this case, the graduating class is an independent or *grouping* variable and the passing rate is a dependent variable. However, it can initially seem like the three graduating classes are three *different* independent variables. Actually, graduating class is a single independent variable with three levels, (**FIGURE 3-2**).

Let us consider one more example of confusing independent and dependent variables. An investigator is studying the relationship

Figure 3-2

Types of variables.



Evaluating the Use of Variables

- Has the investigator explicitly identified and defined the variables in the study?
- Is there a logical connection between the variables so that the reader can correctly identify independent and dependent variables?

between social support and quality of life in older adults in assisted-living environments. The investigator hypothesizes that strong social support influences the quality of life of these older adults. In this example, it seems clear that social support is the independent variable and quality of life is the dependent variable. However, the investigator could propose the equally valid hypothesis that quality of life influences social support; as such, we have flipped independent and dependent variables. In high-quality studies, the investigator provides a logical argument for the choice of variables and the hypothesized relationship between them. The nurse using research for evidence-based practice must know how to evaluate such arguments and decide on the legitimacy of the approach used.

Understanding what variables are, how they are classified, and how they are related to one another is crucial in deciding what statistical method is appropriate for analyzing data. Practicing using statistics is an important part of becoming comfortable with evidence-based practice, and nurses in advance practice and leadership roles will have to work with many examples to become proficient.

LEVELS OF MEASUREMENTS

Wound healing is an important indicator of nursing care across a variety of settings. Let us consider an example where we are interested in implementing a wound healing intervention based on research evidence found in an article. The authors report that wounds healed 50% faster with the intervention than with another commonly used treatment. As part of our plan to implement this intervention (independent variable) we need to know how wound healing (the dependent

Table 3-1 Examples of Different Levels of Measurements

Nominal	Ordinal	Interval	Ratio
Gender	Pain scale (0–10)	Temperature	Age
Ethnicity	Age groups (18–25, 26–35, etc.)	IQ	Height
Zip code	Grade (A, B, C, D, and F)	SAT score	Weight
Religious affiliation	Histological opinion (−/±/+/++/+++)	Depression score	Blood pressure
Medical diagnosis	Patient satisfaction scale (poor, acceptable, good)	Time of day	Years of work experience
Names of medicines	Nurse performance (below average, average, above average)	Dates (years)	Time to complete a task

variable) was measured to determine if the new intervention is an improvement over other approaches to treatment.

After defining the variables of interest, the investigator must think about how to measure the variables. There are four common **levels of measurement**: nominal, ordinal, interval, and ratio (**TABLE 3-1**). It is important to understand the level of measurement because different levels of measurement require different statistical procedures. Measurement is also important for application of research to practice. Let us discuss each level of measurement one by one and see how they differ from each other.

In **nominal level of measurement**, data are classified into mutually exclusive categories (selection of only one category is allowed) where no ranking or ordering is imposed on categories. The word *nominal* simply means to name or categorize. Common examples in this level of measurement are gender and ethnicity; an investigator can classify the subjects as men, women, or transgender for gender and as different ethnic groups (e.g., Caucasian, African American, Asian, or Hispanic), respectively. However, no ranking or ordering can be imposed on any of those categories as we cannot say one gender or ethnic group is superior/inferior, or is more or less than the other groups. Other examples of nominal level of measurement include religious affiliation (e.g., Christian, Catholic, or Buddhist), political party affiliation (e.g., Democrat, Independent, or Republican), and hair color (e.g., black, brown, or blond). Nominal measurement is often used in health-related research to characterize a wide variety of variables such as treatment results (improvement or recurrence) and signs and symptoms (present or not present).

In **ordinal level of measurement**, data are also classified into mutually exclusive categories. However, ranking or ordering is imposed on categories. A common example in this level of measurement is grouped age. People can be categorized into one of the following groups: (a) 18 and under, (b) 19–30, (c) 31–49, and (d) 50 and above. Here, we not only have distinctive categories with no overlapping (mutually exclusive categories), but there is a clear ranking or ordering among categories. Category (b) has older people than category (a), but younger people than categories (c) or (d). Other examples of ordinal level of measurement include letter grade (A, B, C, D, F), Likert-type scales (strongly disagree, somewhat disagree, neutral, somewhat agree, strongly agree), ranking in a race (first, second, third, etc.), and histological ratings (−, ±, +, ++, + + +). A common pain scale, ranking from 0–10, is a good example of ordinal level of measurement in health care, where 0 is equal to no pain and 10 is severe pain. While these data can be ordered, we cannot determine accurately the distance between two categories. That is, we cannot say that the level of pain between 1 and 2 is exactly the same as between 3 and 4.

In **interval level of measurement**, data are classified into categories with rankings and are mutually exclusive as in ordinal level of measurement. In addition, specific meanings are applied to the distances *between* categories. These distances are assumed to be equal and can be measured. Temperature, for example, is measured on categories with equal distance, and any value is possible and can be measured; the distance between 35°F and 40°F is the same as the distance between 55°F and 60°F. However, in interval level of measurement there is no absolute value of "zero." Zero degrees Fahrenheit is not the same as zero degrees Celsius. Therefore, there is no absolute or unconditional meaning of zero. In addition, we cannot say 25°F is three times as cold as 75°F—that is, there is no concept of ratio, or equal proportion, in interval level of measurement. Other examples of interval level of measurement include standardized tests such as intelligence quotient or educational achievement tests. In health care, we often use interval level of measurement for clinical purposes. Screening tests for depression, falls, or risk for pressure ulcers are examples of interval level of measurement.

In **ratio level of measurement**, all characteristics of interval level of measurements are present, in addition there is a meaningful zero, and ratio or equal proportion is present. Therefore, it is possible to apply statistical tests with ratio level of measurement that are not possible with nominal, ordinal, or interval levels of measurement. For example, income is measured on scales with equal distance and a meaningful zero. The measurement of income for someone will be zero if they have no source of livelihood. We may also say that someone making $60,000 a year makes exactly twice as much as someone making $30,000 a year. Blood pressure is another example of ratio level of measurement, as it is possible to have a blood pressure of zero and a systolic pressure of 100 mmHg is twice that of 50 mmHg. Other examples of ratio level of measurement are age, height, and weight.

Level of measurement is important because it directs what statistical tests may be used to analyze the data sets collected by the investigator. Clinicians who understand the relationship between level of measurement and choice of statistical test are able to evaluate the strength of any given study. **TABLE 3-2** gives some examples of statistical tests per level of measurement.

There are times when an investigator may choose to reclassify or transform a variable's level of measurement. For example, blood pressure that is originally measured at the ratio level may be transformed to an ordinal level of measurement if the investigator categorizes the blood pressure in intervals of 40 (i.e., 0–40, 41–80, 81–120, and 121–140), or if transformed into categories of "high" and "low." There may be sound reasons for such transformations, such as wanting to make comparisons between people in those categories. However, transformation from a higher level of measurement to a lower level (ratio to ordinal, for example) limits analysis to those statistical tests for categorical measurements. Second, it will always result in the loss of information, as everyone with blood pressure between 41 and 80 is categorized into a single group. To return to our earlier discussion on variables, variables measured at the nominal and ordinal levels of measurements are discrete or **categorical variables**, and those variables measured at interval and ratio levels of measurements are **continuous variables**.

Table 3-2 Statistical Tests According to Level of Measurement

Independent Variable	Dependent Variable	Statistical Test to be Utilized
Nominal (control/patient)	Ratio (systolic blood pressure)	Independent sample *t*-test, paired sample *t*-test
Nominal or ordinal (low/middle/high systolic)	Ratio (liver function)	One-way analysis of variance (ANOVA)
More than one nominal or ordinal (systolic blood pressure group + gender)	Ratio (liver function)	Factorial analysis of variance
Nominal (control/patient)	Nominal (nondiabetic/diabetic)	Chi-square test of association
Nominal + ratio (control/patient + age)	Nominal (nondiabetic/diabetic)	Logistic regression
Ratio (weight)	Ratio (systolic blood pressure)	Correlation, simple linear regression, multiple linear regression (if more than one independent variable)
Nominal with ratio to control (control/patient with age to control)	Ratio (systolic blood pressure)	Analysis of covariance (ANCOVA)
One or more nominal or ordinal (systolic blood pressure group)	More than one ratio (liver function + depression)	Multivariate analysis of variance (MANOVA)

RELIABILITY AND VALIDITY

In addition to determining what level of measurement will be needed for any given variable, the investigator designing a study or evidence-based practice project needs to choose the best measurement tools. A **tool** or **instrument** is a device for measuring variables. Examples include paper and pencil surveys or tests, scales for measuring weight, and an eye chart for estimating visual acuity. There may be one or more measurement tools available for variables of interest, or there may be none and a tool will need to be created. Whether you use an existing tool or you create one, you should make sure the measurement tool is the best approach for measuring the variable of interest. Strong instruments will reduce the likelihood of **measurement error**. Measurement error is the difference between the measurement value and the true value. Measurement error is unavoidable in research and can be either **random** or systematic. Systematic errors occur consistently because of known causes while random errors by chance and the result of unknown causes. Systematic error may result from the incorrect use of tools or instruments. For example, suppose that you need to measure depression in older adults and you found an instrument, *Beck's Depression Inventory (BDI)* (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). Would you start measuring older adults' depression levels using the BDI right away? Probably not! First, you would want to make sure that the BDI is a good measurement tool to assess depression in older adults. The concepts of **reliability** and **validity** help us to make that decision.

Reliability

Reliability tells us whether or not a test can *consistently* measure a variable. If a patient scores 35 on the BDI over and over again, it means that BDI is reliable, since it is measuring depression consistently at different times. Whether you are engaged in research, evidence-based practice, quality or process improvement choosing a dependable measurement tool is important.

There are three commonly used measures of reliability and all of these are correlation coefficients: internal consistency, test-retest, and interrater reliability. **Internal consistency** is used to measure whether items within a tool, such as a depression scale, measure the same thing (i.e., are they consistent with one another?). Cronbach's alpha, the most commonly used coefficient, ranges from 0 to 1, with a higher coefficient indicating that the items are consistently measuring the same variable. Note that Cronbach's alpha is normally used when the level of measurement is interval or ratio but the Kuder–Richardson (K-20) coefficient is used when the level of measurement is ordinal or nominal. **Test–retest reliability** is used to address the consistency of the measurement from one time to another. If the tool is reliable, the subjects' scores should be similar at different times of measurement. Investigators commonly correlate measurements taken at different times to see if they are consistent. The higher the correlation coefficient, the stronger the test–retest reliability. **Interrater reliability** is used to determine the degree of agreement between individual's scores l on ratings (i.e., are they giving consistent ratings?). Cohen's kappa is commonly used and ranges from 0 to 1, with a coefficient of 1 signifying a perfect agreement. For example, pressure ulcers are often scored on a scale reflecting depth, area, color, and drainage. If two nurses are using a rating scale to score the seriousness of pressure ulcers, we would want to know how consistent the scores are between the nurses. Ideally, both nurses would score the same pressure ulcer very closely.

Instrument reliability is influenced by a number of factors. For surveys or inventories like the BDI, the length of the tool influences reliability. The shorter the tool is the less reliable it will be because it will be more difficult to include all aspects of the variable under study. The second factor is the clarity of expression of each question or item. Confusing questions/items introduce measurement error. The third factor is the time allowed for completing the measurement tool. If the investigator does not allow participants enough time, reliability will decline. The fourth factor is the condition of test takers on the measurement day. If a test taker is ill, tired, or distracted, these conditions can negatively affect the reliability of the measurement tool. The fifth factor is the difficulty of the measurement tool. If the tool is not designed at an appropriate level for the target audience, it can affect the reliability both positively and negatively; reliability will be inflated if the tool was too easy for the target audience and deflated if the tool was too difficult for the target audience. Lastly, the investigator must consider the homogeneity of the subjects, that is, how similar the participants in a study are to one another. If the subjects in a group are very alike they will tend to respond similarly to the instrument and produce similar scores resulting in high reliability. If the subjects in a group are heterogeneous (very different), their scores will range widely and reliability will be lower.

Reliability in its simplest form may be thought of as consistency or stability, and this is an important element to consider when choosing a measurement tool. However, consistency/reliability does not imply accuracy. For example, if we have a thermometer that consistently measures temperature two degrees above the actual temperature, it is a reliable thermometer, but not particularly accurate. In health care and nursing, accuracy of measurement is very important and instruments must be assessed for this element in reports of research.

Validity

Validity tells us whether or not a tool, an instrument, or a scale measures the variable that it is supposed to measure. There are three main types of validity: content validity, criterion-related validity, and construct validity. **Content validity** has to do with whether a measurement tool measures all aspects of the idea of interest. For example, the BDI would not be a valid measure of depression if it did not include somatic symptoms of depression. **Criterion-related validity** is how well a tool is related to a particular standard or benchmark. For example, suppose you wanted to validate the usefulness of a new depression scale, the Patient Health Questionnaire depression scale (PHQ-9) (Kroenke & Spitzer, 2002). An investigator might administer both the BDI, a known and an accurate measure of depression, and the new instrument, the PHQ-9, to a group of patients with depression and compute a correlation coefficient for two scales. A strong association between the BDI and the new PHQ-9 would establish the criterion-related validity of the PHQ9. **Construct validity** is the extent to which scores of a measurement tool are correlated with a **construct** we wish to measure. A construct may be thought of as an idea or concept. For example, an investigator can ask himself/herself this question: "Am I really measuring depression with the BDI or could it also be measuring anxiety?"

Internal and External Validity

The quality of measurement in a study has a direct influence on the strength of the findings or inferences that we can make. **Internal validity** is the extent to which we can say with any certainty that the independent and dependent variables are related to each other. The strength of the internal validity of a study is often evaluated based on whether there is any uncontrolled or **confounding variable** that may influence this relationship between independent and dependent variables. Such confounding variables may include outside events that happened during the study in addition to the variable under the study. These confounding events can actually cause a change in scores or measurement and result in less accurate findings. Changes in the participants due to aging or history may also introduce an element of inaccuracy. In longitudinal studies, for example, past experience with the measurement tool may also confound results, as merely having been exposed to the tool previously may influence the performance of the subjects on the later measurements. The choice of sampling, random or nonrandom, will also influence internal validity. Random sampling is designed to ensure that all participants are equal in every way, reducing the likelihood of confounding variables influencing the study results. Finally, human beings are prone to change their behavior when they know that they are being studied (i.e., Hawthorne effect) and this introduces bias that is difficult to quantify and to explain.

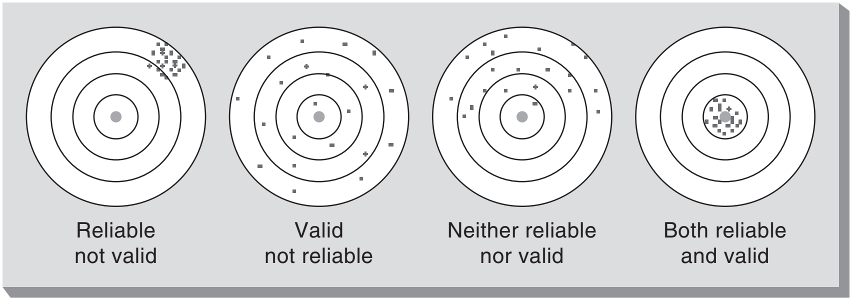
External validity is whether or not the results of a study can be generalized beyond the study itself. Can we make accurate inferences about the population from the sample we have selected? Can we verify with any confidence the hypothesis that we are testing? External validity is influenced by the quality of the sample. If the characteristics of the sample used in the study do not represent the population, the results from the study should not be generalized or inferred to the population. External validity is also influenced by measurement. If our measures are unreliable or inaccurate measures of the variables of interest then we cannot make useful inferences about those variables.

Evaluating Measures

- Are the measures congruent with variables identified in the study?
- Have the investigators reported on the reliability and validity of the measures and instruments?
- Have the investigators reported on limitations related to measurement?

Figure 3-3

Reliability and validity.



Reliability and validity are two important concepts to consider whenever using a measurement tool. You should make sure either that the tool used is both reliable and valid or that its limitations are discussed if the tool is not proven to be reliable or valid. One thing to note is that reliability always precedes validity (**FIGURE 3-3**). An instrument can be reliable but not valid, but an instrument cannot be valid without being reliable. Unless you ensure that the tool has appropriate reliability and validity for your sample, you cannot make inferences following statistical tests.

Case Study

Melnyk, B. M., Kelly, S., & Lusk, P., (2014). Outcomes and feasibility of a manualized cognitive-behavioral skills building intervention: Group COPE for depressed and anxious adolescents in school settings. *Journal of Child and Adolescent Psychiatric Nursing*, 27, 3–13.

In 2014, Melnyk, Kelly, and Lusk reported results from an experimental study designed to test the feasibility and determine preliminary effects of an intervention to reduce depression and anxiety among teenagers aged 14–17. The intervention based on cognitive-behavioral training was delivered to all 16 students over seven sessions. The investigators found that the teens reported statistically significant decreases in depression and anxiety on the Beck Youth Inventory from time one to time two.

The investigators used a known, valid, and reliable measure for the dependent variables of depression and anxiety—the Beck Youth Inventory—to ensure that the findings would be useful to the research and practice community.

SUMMARY

This chapter is intended to introduce the essentials of measurement, including the definition of variables and data, different levels of measurement, reliability, and validity of the measurements, and how these influence the quality of a study by promoting internal and external validity.

Variables are the characteristics or traits that are assumed to vary and data are the values of variables. There are four levels of data: nominal, ordinal, interval, and ratio. Nominal and ordinal data are both made up of categorical/discrete variables; however, only ordinal data has ranking or ordering between/among those categories. Interval and ratio data are both continuous variables with equal distance between intervals, but only ratio data have a meaningful zero and allow for proportionate understanding of the measure.

Reliability and validity may be thought of as the consistency and the accuracy of any given tool or instrument for measuring variables. Reliability and validity of measures influence the internal validity (strength of the relationship between independent and dependent variables) and the external validity (ability to generalize from sample to population) of a study.

Critical Thinking Questions

1. What is the difference between the independent variable and the dependent variable? Give examples of each.

2. Imagine you are to conduct a study on how weight and age group (18–35, 36–53, and ≥54) relate to systolic blood pressure. What are the variables in this study? Characterize each variable in terms of discrete vs. continuous, qualitative vs. quantitative, independent vs. dependent, and level of measurement.
3. What are some ways to improve reliability and validity?
4. Case study: From the following abstract, identify the population, probable sample type, independent and dependent variables, and level of measurement for each variable.

Case Study

Arslan, G. G., & Eşer, I. (2011). An examination of the effect of castor oil packs on constipation in the elderly. *Complementary Therapies in Clinical Practice, 17*(1), 58–62.

Abstract

This research, conducted at two rest homes in Manisa, Turkey, was undertaken to examine the effect of castor oil pack (COP) administrations on constipation in the elderly. Study participants were monitored for 7 days before, 3 days during, and 4 days after COP administration utilizing the Recall Bias and Visual Scale Analog (RB-VSAQ) as well as the Standard Diary developed by Pamuk et al. Eighty percent of study subjects had been constipated for 10 years or longer. COP administration did not have an effect on the number of bowel movements or amount of feces, but it decreased the feces consistency score, reduced straining during defecation, and improved the feeling of complete evacuation after a bowel movement, thus decreasing symptoms of constipation. We conclude that COP may be used for controlling symptoms of constipation.

Self-Quiz

1. True or false? The length of a time in minutes a patient waits to be called at a hospital is interval level of measurement.
2. True or false? An investigator asks a patient participant at a local hospital to rate the service as outstanding, good, fair, poor, and very poor. These data are measured on ordinal level of measurement.
3. Which of the following is the example of continuous variable?
   a. Zip code
   b. Gender
   c. Income
   d. Profit vs. nonprofit nursing home
4. True or false? An instrument can be valid without being reliable.

REFERENCES

Arslan, G. G., & Eşer, I. (2011). An examination of the effect of castor oil packs on constipation in the elderly. *Complementary Therapies in Clinical Practice, 17*(1), 58–62.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561–571.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals, 32*(2), 1–7.

Melnyk, B. M., Kelly, S., & Lusk, P., (2014). Outcomes and feasibility of a manualized cognitive-behavioral skills building intervention: Group COPE for depressed and anxious adolescents in school settings. *Journal of Child and Adolescent Psychiatric Nursing*, 27, 3–13.