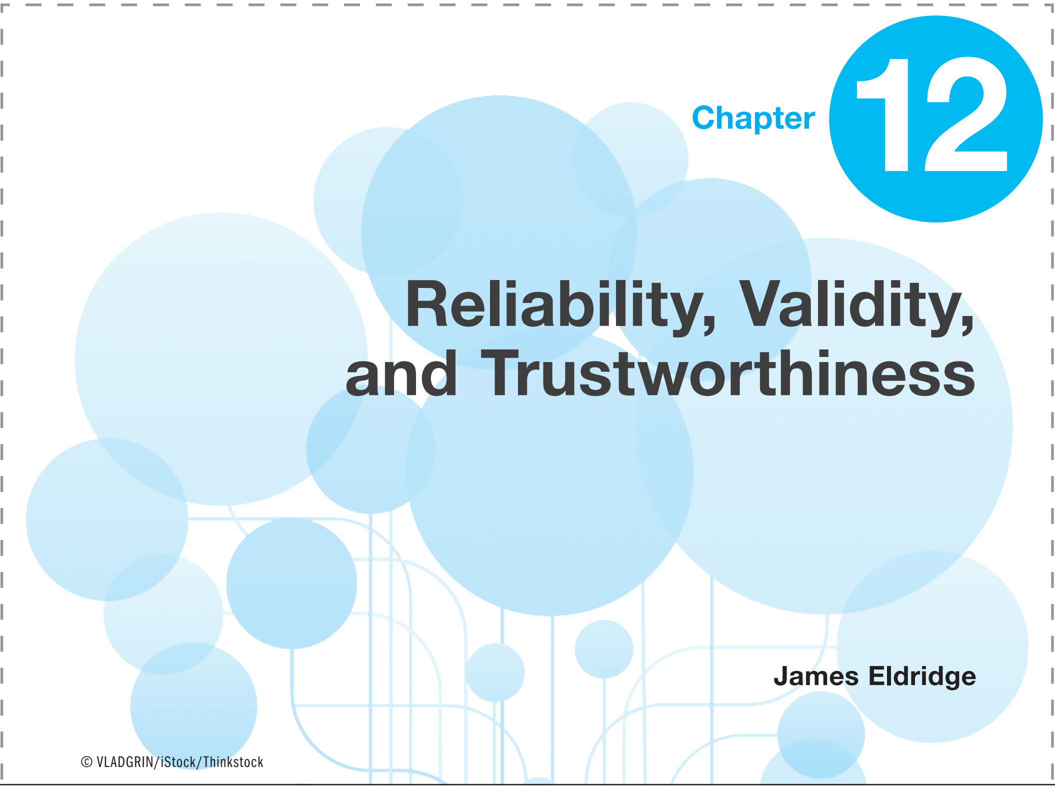


Chapter 12

Reliability, Validity, and Trustworthiness

James Eldridge

© VLADGRIN/iStock/Thinkstock

Chapter Objectives

At the conclusion of this chapter, the learner will be able to:

1. Identify the need for reliability and validity of instruments used in evidence-based practice.
2. Define reliability and validity.
3. Discuss how reliability and validity affect outcome measures and conclusions of evidence-based research.
4. Develop reliability and validity coefficients for appropriate data.
5. Interpret reliability and validity coefficients of instruments used in evidence-based practice.
6. Describe sensitivity and specificity as related to data analysis.
7. Interpret receiver operand characteristics (ROC) to describe validity.

Key Terms

Accuracy

Concurrent validity

Consistency

Construct validity

Content-related validity

Correlation coefficient

Criterion-related validity

Cross-validation

Equivalency reliability

Interclass reliability

Intraclass reliability

Objectivity

Observed score

Predictive validity

Receiver operand characteristics (ROC)

Reliability

Sensitivity

Specificity

Stability

Standard error of measurement (SEM)

Trustworthiness

Validity

Introduction

The foundation of good research and of good decision making in evidence-based practice (EBP) is the **trustworthiness** of the data used to make decisions. When data cannot be trusted, an informed decision cannot be made. Trustworthiness of the data can only be as good as the instruments or tests used to collect the data. Regardless of the specialization of the healthcare provider, nurses make daily decisions on the diagnosis and treatment of a patient based on the results from different tests to which the patient is subjected. To ensure that the individual makes the proper diagnosis and gives the proper treatment, the nurse must first be sure that the test results used to make the decisions are trustworthy and correct.

Working in an EBP setting requires the nurse to have the best data available to aid in the decision-making process. How can an individual make a decision if the results being used as the foundation of that process cannot be trusted? Put simply, a person cannot make a decision unless the results are trustworthy and correct.

This chapter presents five concepts to help the nurse determine whether the data upon which decisions are based are trustworthy: reliability, validity, accuracy, sensitivity, and specificity. Each defines a portion of the trustworthiness of the data collection instruments, which in turn defines the trustworthiness of the data, ensuring a proper diagnosis or treatment.

Reliability and validity are the most important qualities in the decision-making process.

- Reliability = the instrument consistently measures the same thing.
- Validity = the instrument measures what it is intended to measure.

If either of these qualities is lacking in the data, the nurse cannot make an informed decision and, therefore, is more likely to make an incorrect decision. An incorrect decision in the medical field can have catastrophic consequences for the patient. Thus, one can see why reliability and validity are so important. What do these concepts mean? What would happen if the same test was run on a person several times but the results were different each time? In the case of varying results, a decision becomes ambiguous because the results are unclear.

Reliability is defined as the consistency or repeatability of test results. Other descriptors used to indicate reliability include "consistency," "repeatability," "objectivity," "dependability," and "precision." Accuracy is a function of reliability: The better the reliability, the more accurate the results. Conversely, the poorer the reliability, the more inaccurate are the results, which increases the chance of making an incorrect decision. Furthermore, accuracy is affected by the sensitivity and specificity of the test. **Sensitivity** can be defined as how often a test measures a "true" positive result, while **specificity** determines the capability of the test for determining "true" negative results. The greater the sensitivity and specificity is for a test, the more accurate the test results. The concepts of sensitivity and specificity are discussed in more detail later in this chapter.

Validity is defined as the degree to which the results are truthful. It depends on the reliability and relevance of the test in question (**Figure 12-1**). Relevance is simply the degree of the relationship between the test and its objective, meaning that the test reflects what was reported to be tested.

An example of relevance is the measurement of the height of a patient. A nurse uses a stadiometer (a ruler used to measure vertical distance) to establish a patient's height. Is the stadiometer a relevant height measurement device? Height is the vertical distance from the floor to the top of the head, and a stadiometer measures vertical distance from the floor to any point above the floor; thus, the stadiometer is a relevant measure of height.

Validity cannot exist without reliability and relevance, but reliability and relevance can exist independently of validity. **Figure 12-2a** depicts the case in which there is a high degree of reliability and a low degree

Figure 12-1

Relationship of validity to reliability and relevance.

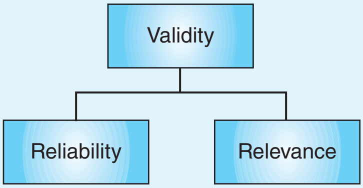
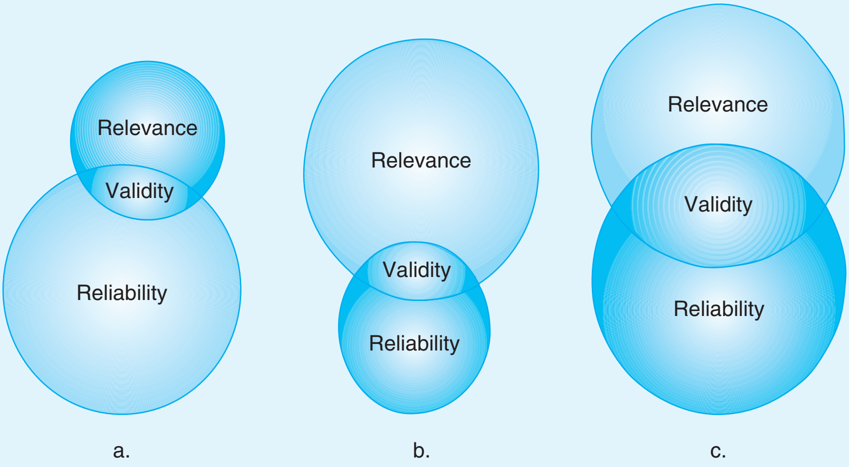


Figure 12-2

Relationship among reliability, relevance, and validity.



of relevance. In this representation, even when reliability is high, validity is low due to the lack of relevance. This figure shows that under the most reliable test, a low degree of relevance decreases the validity of the test.

Figure 12-2b depicts the situation in which there is a high degree of relevance and a low degree of reliability. In this representation, even when relevance is high, validity is low due to the lack of reliability. Even when a nurse uses what might be considered the most relevant test for the situation, if the instrument has a low degree of reliability, it will also have a low degree of validity.

Figure 12-2c shows the desired capacity for an instrument—to have both a high degree of reliability and a high degree of relevance, thereby creating a high degree of validity. Whereas the other examples show that a test can be reliable but not relevant, or relevant but not reliable, a valid test will always have some degree of reliability and relevance. When validity is absent, the results of the testing are not truthful and making an informed or evidence-based decision is impossible. However, when validity is present, a nurse can be assured that the decision is based on truthful evidence.

Reliability as a Concept

As previously described, reliability focuses on the repeatability or consistency of data. To understand the theoretical constructs of reliability, one must understand the concept of the **observed score**. By definition, the observed score is the score that is seen; stated in other terms, the observed score is the actual score printed on the readout of an instrument.

An example of an observed score is the measurement of a patient's blood pressure. The systolic and diastolic pressures are determined based on the aneroid dial or digital liquid crystal display (LCD) readings associated with the first sound (systolic) and the last sound (diastolic) heard in the brachial artery. If the first sound occurs at a reading of 130 mmHg, this is the systolic observed score. If the last sound occurs at 85 mmHg, this is the diastolic observed score. These observed scores for blood pressure are not the true blood pressure scores for the patient, as those scores ultimately depend on factors such as the amount of error incorporated in the type of sphygmomanometer, the quality of the stethoscope, the quality of hearing of the person taking the blood pressure, the experience of the person taking the measurements, and placement of the cuff over the artery.

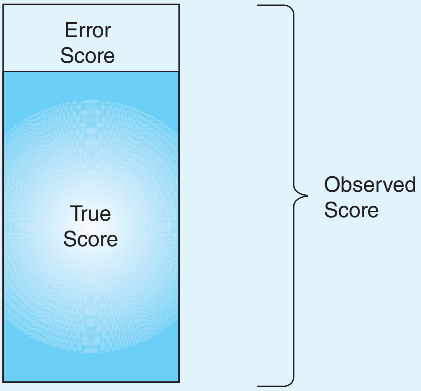
A second example that may help in understanding the reliability of an observed score is the measure of quality improvement of a specific program. For instance, consider a hospital that wants to determine whether a specific pain management protocol helped reduce hospital days for patients. It used a pain scale that patients completed every 6 hours, and patient release was determined by the patient achieving an observed score of 3 on a 10-point pain scale. In this case, the scale would need very little error, because a change of 1 point on the scale might determine the release or premature release of a patient. If the scale had a high error rate—for example, 2 points of error—and the patient scored a 3 on the scale, then it would not be possible to know if the score was a 3, as high as 5, or as low as 1. The scale number may be affected by time of day the question was asked, the way the question was asked, the type of pharmaceuticals the patient is receiving, the patient's language skills, the patient's tolerance, and the severity of the initial injury causing the pain.

Each of these nuances can add or subtract error from the true score, which increases the variability between the observed score and the true score. This variability can be described as the error score, thereby defining the observed score as the sum of the true score and the error score. As shown in **Figure 12-3**, any error within the measurement decreases the degree to which the observed score reflects the true score. Note that the net effect of an error score can be positive or negative, depending on the nature of the error.

The true score exists only in theory, because all data collected are observed score data. A nurse can think of the true score as the perfect score of a test—that is, a score without any error and void of any misinterpretation. Of course, the world is not perfect and, therefore, neither are any data that might be collected. Thus, a true score exists and never changes for a given period of time; changes occur only in the error score, which then determines the observed score.

Figure 12-3

Observed score, true score, and error score.



THINK OUTSIDE THE BOX

Discuss the elements of trustworthiness as related to making decisions about the data found in the study by Iverson and colleagues (2014). Where might error occur within the screening process and questionnaire used in this study?

Reliability is the degree to which the observed score of a measure reflects the true score of that measure. Therefore, reliability could theoretically be calculated as the proportion of observed score variance that consists of true score variance (**Figure 12-4**). In this equation, if no error exists, then the observed score variance and the true score variance are equal, and the reliability coefficient is 1.0. Conversely, when the observed score variance and the error score variance are equal, the reliability coefficient is 0. Thus, reliability always falls within the range of 0–1.0, with perfect reliability equaling 1.0 and no reliability equaling 0. For research

Figure 12-4

Theoretical calculation of reliability.

$$\text{Reliability} = \frac{S^2\,\text{true}}{S^2\,\text{observed}} = \frac{S^2\,\text{observed} - S^2\,\text{error}}{S^2\,\text{observed}}$$

purposes, high reliability measures are desired if at all possible. The general rule is that reliability coefficients greater than 0.80 are considered to be high. Note that if the reliability coefficient is calculated to be greater than 1 (e.g., 1.15), a calculation error has been made, because the range of reliability is always between 0 and 1.0.

Forms of Reliability

Although the purpose of the theoretic concept of reliability is to determine the relationship between the true and observed scores of a measurement, practical use of this concept allows a nurse to determine the relationship only between two or more observed scores. The relationship between these observed scores allows an individual to estimate reliability and to determine a range for the true score. The outcome of the calculation of the relationship between two or more observed scores is known as the correlation coefficient. The **correlation coefficient** is the practical calculation of the theoretic expression of the proportion of observed score variance that consists of true score variance, as described previously.

Given this basic understanding of reliability as a concept, it is now time to learn about the forms of reliability. Globally, reliability can be described as either **interclass reliability** or **intraclass reliability**. The most basic description of interclass reliability is the reliability between two and only two variables or trials, whereas intraclass reliability is the reliability between more than two variables or trials. The limiting factor that separates the two forms of reliability is the number of variables or trials that can be used in the calculation of the correlation coefficient. The number of variables also determines which statistical equation is used to develop the correlation coefficient. Each of these considerations has its place in EBP depending on the number of variables a nurse uses to calculate the reliability coefficient.

Interclass Reliability

Interclass reliability is the reliability between two measures that are presented in the data as either variables or trials. Four types of interclass reliability are distinguished:

- Consistency
- Stability
- Equivalency
- Internal consistency

Each of these reliability coefficients is developed using a Pearson Product Moment (PPM) correlation. Most statistical packages or spreadsheet software can calculate PPM correlations; therefore, the actual equation is not included in this text. Although each interclass reliability coefficient uses the same formula, the calculated reliability coefficient is defined by the type of variables to be compared and the methods used for interpretation of the results. This concept becomes more evident as the types of interclass reliability are further defined.

Consistency

One type of interclass reliability to report is the consistency of a measure. **Consistency** simply describes the degree to which you can expect to get the same results when measuring a variable more than once on a single day. Consistency reliability is sometimes described as test–retest reliability, because it compares two trials of a single measure. An example of testing for consistency would be running two tests on a single blood sample from each subject to measure hemoglobin using a single hemoglobin analyzer. The question is whether the results from the hemoglobin analyzer are consistent within a single day. In **Table 12-1**, the subjects' hemoglobin from a single sample of blood was measured twice, and the reliability coefficient was calculated to be 0.996.

This coefficient simply means that 99.6% of the observed score variance is true score variance. Because the reliability coefficient is close to 1.0, the reliability of the instrument is high. The initial question with this data was whether or not the machine was consistent. The results demonstrate that it was consistent, with a consistency reliability coefficient of $r_{xx}' = 0.996$.

Table 12-1

Consistency and Stability of the Ac•T diff Analyzer

Subject Number	Test 1 (g/dL)	Test 2 (g/dL)
1	14.10	14.00
2	12.20	12.10
3	11.90	11.90
4	14.50	14.40
5	13.80	13.90
6	13.20	13.10
7	13.50	13.60
8	14.00	14.10
9	11.10	11.00
10	9.60	9.90
		r = 0.996

Stability

When results of trials or tests are collected over 2 or more days, consistency becomes **stability**. Suppose we take the same data from Table 12-1, this time imagining that the samples were tested over a 2-day period. The question now becomes whether a blood sample is stable over a 2-day period. Notice that the results remain constant, because nothing has changed except the theoretical timing of the tests. The reliability coefficient is still 0.996, but this time a nurse would interpret the results as the samples being stable over a 2-day period, with a stability reliability coefficient of $r_{xx}' = 0.996$.

Both consistency and stability have their place in EBP. In the current example of hemoglobin testing, the consistency of the measures is described by determining that, for any time during a single day, the data would be repeatable. A nurse can expect the same results as long as no other factors have occurred in the interim, such as acute onset of anemia. In other words, the nurse is sure that the hemoglobin analyzer will give the same measure of hemoglobin for the same sample within the same day. Notice that nowhere in this example of consistency do we assume that the measurement gives the correct amount of hemoglobin, only that it indicates the presence of the same amount of hemoglobin. To determine if this is the correct amount of hemoglobin, the relevance and the validity of the instrument would have to be known.

When discussing this example in terms of stability, the key determination relates to the length of time that the blood samples remain stable. Hemoglobin analyzers usually have instructions that indicate the timeframe for running samples before differing results would be seen. In most instances, the timeframe is usually 24 hours. A question might arise concerning how the manufacturer determined this timeframe. The answer simply is that the manufacturer developed a stability coefficient using the same techniques described previously.

Again, notice that nowhere in the example of stability is there any mention of the correctness of the amount of hemoglobin over a 24-hour period; the only consideration is that it is the same amount of hemoglobin measured for a 24-hour period. To determine whether this is the correct amount of hemoglobin over the 24-hour period, the relevance and the validity of the instrument and the measures would need to be determined.

Equivalency

Another type of interclass reliability to report is equivalency. This kind of reliability allows a person to report whether one type of test is equivalent to another. **Equivalency reliability** is calculated in the same manner as the consistency and stability coefficients described previously, except that a PPM correlation between two forms of a single test is calculated, rather than a single variable over two trials.

An example of testing for equivalency reliability would be comparing two methods of blood pressure measurement to determine if they are equivalent. In this case, the question is whether the systolic blood pressure results determined by an automatic blood pressure cuff are equivalent to those recorded from manual blood pressure measures using a stethoscope and sphygmomanometer. As shown in **Table 12-2**, subjects' systolic pressure was measured once with an automatic cuff and once using manual methods. The reliability coefficient was calculated to be 0.959.

This coefficient simply means that 95.9% of the observed score variance consists of true score variance. The reliability coefficient is close to 1.0 reflecting the reliability between the instruments is high. The initial question with these data was whether automatic cuff readings are equivalent to manual readings of systolic blood pressure. A person can now report that the two methods are equivalent, with an equivalency reliability coefficient of $r_{xx}' = 0.959$. These results indicate that either an automatic cuff or manual methods are acceptable for measuring systolic blood pressure, because they are equivalent. No matter which method is used, a nurse can expect to get similar measures from a single individual. Notice again that there is no mention of the correctness of the data, only the similarity of the data. To determine if the blood pressure measures are correct, the relevance and the validity of the measures would need to be determined.

Table 12-2

Equivalency of Automatic Versus Manual Systolic Pressure Readings

Subject	Automatic Cuff Systolic (mmHg)	Manual Method Number Systolic (mmHg)
1	150.00	155.00
2	130.00	128.00
3	125.00	129.00
4	124.00	120.00
5	122.00	125.00
6	148.00	144.00
7	133.00	135.00
8	146.00	143.00
9	117.00	120.00
10	121.00	120.00
		r = 0.959

Internal Consistency

The final type of interclass reliability discussed here is the internal consistency of written tests. Internal consistency reliability is sometimes described as split-halves reliability, because it entails comparing two halves of a written test. To calculate the internal consistency of a written instrument, the instrument responses are divided into two equal halves. The sum of each half is calculated to make the comparison.

The simplest means for dividing a test in half is to compare the sum of the odd-numbered question responses with the sum of the even-numbered question responses. If possible, the questions should be matched between each half, based on their content and difficulty. Another possible method is to make the *a priori* assumption that both halves are equal because the questions were randomly placed in order during the development of the written test. As with the other types of interclass reliability, the PPM correlation is used to develop the reliability coefficient.

Data for a 10-item pain questionnaire are presented in **Table 12-3** to demonstrate the principle of internal consistency. Each item of the pain questionnaire is scored from 0 (strongly disagree) to 5 (strongly agree). The questionnaire is then divided into odd and even scores, with the sum of the scores for the odd-numbered items and the sum of the scores for the even-numbered items presented in the table. The question under

Table 12-3

Internal Consistency of a 10-Item Pain Questionnaire

Subject	Odd-Numbered Item Scores	Even-Numbered Item Scores
1	25.00	21.00
2	18.00	14.00
3	16.00	18.00
4	12.00	14.00
5	10.00	10.00
6	18.00	19.00
7	15.00	18.00
8	12.00	9.00
9	14.00	15.00
10	17.00	13.00
		$r = 0.761$

THINK OUTSIDE THE BOX

Look around your clinical setting. Which tools or instruments are present, and how are they typically used for data collection? Do they include surveys of employees, patients, or consumers? Are the tools or instruments used appropriately?

consideration is whether this questionnaire has internal consistency. As with the previous types of reliability, the reliability coefficient is reported; here, it is 0.761. This coefficient simply means that 76.1% of the observed score variance consists of true score variance. Notice that the internal consistency is lower than in previous examples. The fact that the reliability coefficient is lower does not mean that the questionnaire is not reliable—just that it is less reliable than it could be.

The initial question for these data was whether the pain questionnaire was internally consistent. We can now report that it has some internal consistency, with a reliability coefficient of $r_{xx}{}' = 0.761$, but there is at least some error present in the questionnaire. In other words, the questionnaire is not perfectly consistent internally, so the results from using the questionnaire will not be an accurate reflection of the true score. This does not mean that this questionnaire should not be used, but rather that a person needs to be careful in the interpretation and use of the results of the questionnaire. When using written item tests, individuals can actually estimate how reliability will change as a result of adding items to the questionnaire. To estimate a new reliability for a written questionnaire with added items, the Spearman–Brown prophecy formula (**Figure 12-5**) could be used.

Where $r_{kk}{}'$ is the new reliability coefficient, $r_{xx}{}'$ is the original reliability coefficient, and *k* is the total items on the new questionnaire divided by the number of items on the original questionnaire, the Spearman–Brown prophecy can be determined. In the example given in Table 12-3, the reliability coefficient was 0.761. To calculate the reliability of the questionnaire if 10 questions were added, a person would solve for $r_{kk}{}'$ using the information shown in **Figure 12-6**.

The original reliability coefficient is 0.760 and the number of total items on the new questionnaire divided by the total items on the original

Figure 12-5

Spearman-Brown prophecy.

$$r_{xx'} = \frac{k \times r_{xx'}}{1 + r_{xx'}\,(k - 1)}$$

Figure 12-6

Spearman-Brown prophecy example.

$$r_{xx'} = \frac{2 \times 0.761}{1 + 0.0761(2 - 1)}$$

test is 2. Notice that by increasing the number of items on the questionnaire to 20, the new reliability coefficient for the questionnaire becomes 0.864. This coefficient is higher than the original value. Thus, adding items to the questionnaire improves this tool's internal consistency and strengthens the interpretation of its results. As discussed earlier, as reliability and relevance increase, so does validity. If the questionnaire being used has a high degree of relevance, the addition of more questions to the questionnaire (assuming they are relevant) would increase the reliability of the questionnaire, thereby improving the validity of its results.

In the article by Hanna, Weaver, Slaven, Fortenberry, and DiMeglio (2014), Cronbach's alpha coefficients were provided for both of the instruments used within the study. The diabetes-related quality of life (DQOL) tool demonstrated Cronbach's alpha coefficient scores of the subscales of 0.84, 0.83, and 0.90 during T1 and 0.85, 0.84, and 0.90 during T2. For the second tool, the Emerging Adult Diabetes Management Self-Report, the Cronbach's alpha coefficient was 0.81 at T1 and 0.85 at T2. As can be seen, each of these scores are close to the 1.0 level which implies a high internal consistency. As was stated earlier, a common interpretation of the reliability coefficient scores reflects that any level greater than 0.80 is considered to be high. All eight of these Cronbach's alpha coefficient scores exceed this level. The article does discuss the covariates for the depressive symptoms as measure by the Beck Depression Inventory (BDI-II). In-depth discussion related to the internal consistency was not provided within the article. The discussion centered on providing the statistical levels, which were found for the different tools.

Intraclass Reliability

Now that we have an understanding of interclass reliability, it is time to move on to intraclass reliability. As discussed earlier, the basic difference between interclass reliability and intraclass reliability is the number of variables that can be analyzed. Interclass reliability testing allows for the reliability analysis of only two variables, whereas intraclass reliability testing allows a researcher to develop a reliability coefficient for more than two variables.

Suppose we wanted to measure the reliability of three different pain scales. One of the scales requires only 2 minutes for completion, the second scale requires 10 minutes for completion, and the third scale requires 30 minutes for completion. The nurse would prefer to use either the 2-minute or 10-minute scale for efficiency, but the 30-minute scale is currently being used. Although the data could be analyzed using three PPM correlations to determine the equivalency reliability coefficients for these tools, this kind of analysis would miss a very important portion of the error: In the PPM interclass analysis, the statistic estimates only the error between the items, but it ignores the error within the item that reflects the differences in individuals taking the test.

In contrast, the intraclass reliability coefficient uses analysis of variance (ANOVA) to determine not only the error between the tests, but also the error within the tests. Using ANOVA allows for construction of a better estimate of the overall reliability of the scales and the errors that reduce the observed score variance, which is the true score variance. Thus, whereas PPM analysis allows for only a two-dimensional view of reliability, ANOVA supports a three-dimensional view of reliability. Notice that the basic terms of reliability remain the same. In the current example, a nurse is still estimating the equivalency of the scales, but now an error that might exist within each individual scale is included.

Figure 12-7 shows the equation used in determining a reliability coefficient using ANOVA. In this equation, a reliability coefficient is developed using the mean square between scales and the mean square within scale data from the ANOVA table.

Table 12-4 presents data for the example of the three pain scales. These ANOVA data include the between-cells mean square of 2908.233 and the within-cells mean square of 35.100. As shown in **Figure 12-8**, the reliability coefficient is determined by substituting the numbers represented in the table into the ANOVA equation for reliability (Figure 12-7).

In this example, the equivalency reliability is 0.988 for the three scales. We can now state that the 2-minute pain scale is equivalent to the 10-minute pain scale and the 30-minute pain scale. The evidence for replacing the longer 30-minute test with the more efficient 2-minute test is now documented, because the tests are equivalent. The same ANOVA reliability equation can be used to determine consistency, stability, and equivalency, depending on the intended use of the data.

Figure 12-7

Intraclass reliability coefficient using ANOVA.

$$r_{xx'} = \frac{MS_{\text{between}} - MS_{\text{within}}}{MS_{\text{between}}}$$

Table 12-4

Intraclass Reliability Using ANOVA

		Scale		
Subject Number	**2-Minute Scale**	**10-Minute Scale**	**30-Minute Scale**	
1	15.00	35.00	60.00	
2	12.00	30.00	51.00	
3	9.00	22.00	40.00	
4	10.00	25.00	42.00	
5	11.00	19.00	43.00	
6	14.00	31.00	45.00	
7	6.00	20.00	38.00	
8	3.00	15.00	33.00	
9	12.00	22.00	45.00	
10	11.00	21.00	45.00	
Source of Variation	**SSq**	**DF**	**MSq**	**F**
Between cells	5816.467	2	2908.233	82.86
Within cells	947.700	27	35.100	
Total	6764.167	29		

Note: DF = degrees of freedom; F = *F*-distribution; MSq = mean square; SSq = sum of squares.

Objectivity

An area of intraclass reliability that is often overlooked is the measure of **objectivity**. Objectivity is the reliability of scores assigned by judges, multiple observers, or reviewers. In theory, if three individuals see the same performance, they should score the performance based on the merits of the performance, such that their scores are not affected by internal biases that each may possess. When no bias is evident, the scores should be similar among the judges.

A good example of objectivity (or lack of objectivity) comes from the 2002 Winter Olympics figure skating competition, in which three

Figure 12-8

Intraclass reliability coefficient using ANOVA.

$$r_{xx'} = \frac{2908.233 - 35.10}{2908.233} = 0.988$$

judges rated the performance of the Canadian skating pair. Two of the judges assigned scores of 9.9 and 9.8 for the pair's performance, but a third judge scored the pair at 7.8. If no biases were associated with the scoring method, then the third judge should have been expected to score the performance in the 9.7–9.9 range.

Objectivity also has relevance for EBP. The Apgar score—a tool for assessing the health of newborn infants—offers an example of objectivity in healthcare practice. If three medical professionals are in the delivery room, the Apgar scores each assigns to the newborn should be equivalent. This factor can be tested using the same ANOVA techniques described in the previously given pain scale example, albeit with scores for each observer, rather than each scale, being used. A researcher could determine if the Apgar scores are objective. If they are not, the researcher could meet with the observers to determine where differences occurred.

By now, it should be clear that the same formula (either PPM or ANOVA, depending on the number of trials) is used to determine the reliability of any measure. The only difference in the results relates to the interpretation based on the intended use of the data.

Accuracy

Another item that is important when determining the intraclass reliability of a test is the test's **accuracy**. The measure of the accuracy of a test is known as the **standard error of measurement (SEM)**. The SEM reflects the fluctuation of the observed score attributable to the error score. Computing the SEM allows a researcher to determine confidence intervals for the observed score based on the standard deviation of the test and its reliability. The relationship between the true score and the observed score was discussed earlier in this chapter. The SEM allows a researcher to provide a range for which the true score is present.

The equation shown in **Figure 12-9** is used to calculate the SEM. Notice that in this equation, the reliability coefficient of the test and the standard deviation of the sample are used.

The SEM can be determined for any of the prior examples. For Table 12-1, the standard deviation of the sample is 1.496, and the reliability coefficient is 0.996. Using the equation in Figure 12-9, we can compute the SEM as ± 0.0946 mg/dL (**Figure 12-10**).

THINK OUTSIDE THE BOX

On most clinical units, many different tools are regularly used, such as thermometers, glucometers, sphygmomanometers, and weight scales. Are these tools accurate? How can you be sure that they are reliable and valid for what they are being used to evaluate? Are they valid and reliable tools?

Figure 12-9

Standard error of measurement.

$$\text{SEM} = s\sqrt{1 - r_{xx'}}$$

In a normal distribution, 68% of the sample scores fall between ± 1 standard deviation of the mean. Thus, for this example, we have 68% confidence that the hemoglobin scores will fall between ± 0.0946 mg/dL of the measured score. If a ± 2 standard deviation from the mean is used, a 95% confidence interval for the scores is expected. To find the SEM for ± 2 standard deviations from the mean, we multiply the SEM by 2 (the number of standard deviation units). In our example, we have 95% confidence that the true hemoglobin score will fall between ± 0.1892 mg/dL of the measured score. If a blood sample is run in the analyzer and the hemoglobin level is found to be 14.0 mg/dL, we would therefore have 95% confidence that the true score is between 13.1080 mg/dL and 14.1892 mg/dL. Notice that as the standard deviation increases for a set of scores, the SEM increases. Also, as the reliability of a set of scores decreases, the SEM increases. To proclaim a tool as giving an accurate measure, test scores need a relatively low standard deviation and a high reliability coefficient.

Up to this point, we have examined accuracy as it relates to continuous data. But what happens when a test uses nominal data—how do we determine its accuracy? In the case of nominal data, we use the χ^2 (chi-square) statistic and its corresponding phi coefficient as a measure of accuracy. Think of the phi coefficient as a correlation or reliability coefficient for nominal data. A χ^2 statistic and its corresponding phi coefficient would most likely be used when you are trying to determine whether a new test is equivalent to a "gold standard" test. All of the same rules apply just as they have in the previous discussion of reliability for continuous data; however, now you are simply determining the accuracy of the new test based on its "pass or fail" performance compared to the "gold standard" test.

Be aware that reliability and accuracy can be sensitive to situational changes; although a test is reliable in one situation or within one group,

Figure 12-10

SEM for consistency of a hemoglobin analyzer.

$$\begin{aligned}\text{SEM} &= 1.496\sqrt{1 - 0.996}\\ &= \pm\ 0.0946\ \text{mg/dL}\end{aligned}$$

it may not always be reliable when the situation or group changes. This consideration is especially important concerning written items. Factors that can affect reliability and accuracy include the following issues:

- *Fatigue*. Fatigue of the person taking the test or collecting the data can decrease reliability.
- *Practice*. The more practiced a person becomes at taking a test or in collecting data, the more reliability is improved.
- *Timing*. The more time that passes between test administrations, the more the reliability of the test is decreased.
- *Homogeneity of the testing conditions*. The more homogeneous the testing conditions (e.g., same room, same time taken to collect data, same time of day), the better the reliability.
- *Level of difficulty*. The more difficult a test or data collection procedure, the lower the reliability.
- *Precision*. The more precise the measurement (e.g., 1/100 vs. 1/1000 decimal), the better the accuracy.
- *Environment*. Environmental changes such as ambient pressure or temperature variations can decrease reliability.

The more control maintained over these factors, the better the reliability and accuracy of the resulting data. Accuracy and reliability improve the decision-making process in EBP.

Receiver Operand Characteristics (ROC) and Accuracy

When discussing nominal data, historically the use of the χ^2 statistic determines accuracy; however, newer statistical methods such as **receiver operand characteristics (ROC)** curves are being implemented in the field of nursing to determine accuracy of test results (Zou, O'Malley, & Mauri, 2007). ROC analyses were first developed for the armed services during World War II as a method for determining the accuracy of radar signals. More recently, this statistical method is being adapted to the medical field for defining the accuracy of diagnostic tests. ROC analysis determines the sensitivity and specificity (accuracy) of a diagnostic test to predict a specific outcome of disease the test is reported to measure. Most of the time, ROC analysis uses dichotomous variables much like a 2 × 2 χ^2 statistic; however, the analysis can also be used when an ordinal grading system is available for determining disease severity. The most basic form of ROC analysis uses a 2 × 2 method for determining accuracy of positive and negative results from a specific diagnostic test compared to whether or not the patient actually possesses the disease. **Table 12-5** represents the conceptual nature of a dichotomous diagnostic test comparing the positive and negative test results to actual disease state of a patient (nondiseased or diseased).

In Table 12-5, a perfectly accurate test would indicate only true negative results and true positive results; however, as discussed previously, there is always some inherent measurement error in diagnostic tests.

Table 12-5

Conceptual Nature of a Dichotomous Test

	Disease State	
Test Result	**No Disease**	**Disease**
Negative test result	True negative	False negative
Positive test result	False positive	True positive

The ROC analysis allows the medical provider a means to quantify this error and determine in which area of the figure the error is greatest. Unlike the SEM, which gives the researcher a global characterization of the measurement error, ROC analysis allows the researcher to determine the sensitivity (rate of true positive results) and the specificity (rate of true negative results) for any diagnostic test.

Calculations for sensitivity and specificity are fairly simple to develop. The researcher needs to know the rates or number of individuals within each of the four groups (true negative, false negative, true positive, and false positive). **Table 12-6** simplifies the variables necessary to calculate sensitivity and specificity.

In Table 12-6, TN represents the number of individuals who do not have the disease and have negative test results on the diagnostic test (true negatives). FP represents the number of individuals who do not have the disease but have positive results on the diagnostic test (false positives). FN represents the number of individuals who have the disease but have negative test results on the diagnostic test (false negative). TP represents the number of individuals who have the disease and have positive results on the diagnostic test (true positives). To calculate sensitivity (the probability of the test to correctly predict true positive

Table 12-6

Variables for Calculating Sensitivity and Specificity

	Disease State		
Test	**No Disease**	**Disease**	**Total**
Negative	TN	FN	TN + FN
Positive	FP	TP	FP + TP
Total	TN + FP	FN + TP	*n*

Note: FN = false negative; FP = false positive; TN = true negative; TP = true positive.

scores), the formula used is TP/(TP + FN). To calculate specificity (the probability of the test to correctly predict true negative scores), the formula used is TN/(TN + FP). Many reasons might be discussed for why ROC analyses might be used, but one of the most common reasons is to determine if a less invasive and less expensive diagnostic test will provide as good or better results when compared to the "gold standard" diagnostic test for a given disease.

For an example of calculating sensitivity and specificity of a diagnostic test, let's assume a new diagnostic test was developed for assessing the presence of carpal tunnel syndrome. The test uses a tactile response of the index fingers by touching the fingers with a thin monofilament line while conducting a modified Phalen's test (MPT) for carpal tunnel syndrome. The response from the patient is simply YES, they feel the thread (positive MPT) or NO, they do not feel the thread (negative MPT). Previously each patient was diagnosed for the presence (positive electrodiagnostic neural conduction study [EDS]) or absence (negative EDS) of carpal tunnel syndrome via an EDS. The data for the test are provided in **Table 12-7**.

To determine the sensitivity of the modified Phalen's test, the equation would be 39/46, where 39 represents the number of individuals who reported a positive MPT and a positive EDS score, and 46 represents the total positive EDS scores. The sensitivity of the modified Phalen's test is 0.848, or 84.8% probability of predicting true positive tests. To determine the specificity of the modified Phalen's test, the equation would be 20/21, where 20 represents the number of individuals who reported a positive MPT and a negative EDS score, and 21 represents the total negative EDS scores. The specificity of the modified Phalen's test is 0.952, or 95.2% probability of predicting true negative tests. The conclusion from these data is that the modified Phalen's test can accurately predict both true positive and true negative tests.

The philosophical discussion that occurs when using ROC analysis in the healthcare field is what should be considered acceptable values for sensitivity, specificity, and overall accuracy. Acceptable values are often

Table 12-7

Sensitivity and Specificity of the Modified Phalen's Test

Neg	Negative EDS	Positive EDS	Total
Negative MPT	20	7	27
Positive MPT	1	39	39
Total	21	46	66

Note: EDS = electrodiagnostic neural conduction study; MPT = modified Phalen's test.

dependent on the severity of the disease state. If the disease is life threatening, then sensitivity values should be above 85%, while specificity values may be somewhat lower. If the disease or diagnosis is mundane, then sensitivity values may be lower, but specificity values should be higher. The overall accuracy of a test should still follow the general rules of reliability and exceed 80%.

Validity

To this point in the chapter, the knowledge necessary to understand the reliability and accuracy of the data collected has been provided. The fact that a test has accuracy and reliability does not mean that the test is valid, however. A valid test is defined as a test that truthfully measures what it purports to measure. Validity can be classified as either logical or statistical in nature. Logical validity requires inference and understanding of the subject being measured. Statistical validity uses statistical formulas to compare the test in question with a specific criterion or known valid measure. In EBP, validity is further delineated into three types: content-related validity, criterion-related validity, and construct-related validity. Depending on the measure, either one type or several types of validity can be used to determine if a measure is valid.

Content-Related Validity

Content-related validity is based on the logical thought process and interpretation of the measure. Many people refer to this quality as face or logical validity. The American Psychological Association (APA, 1985) defines content-related validity as "demonstrating the degree to which the sample of items, tasks, or questions on a test is representative of some defined content" (p. 10). A humorous restating of this concept is the cliché, "If it looks like a duck and quacks like a duck, then it must be a duck." A valid test using content-related validity should logically measure the content being reported.

Consider the pain scale example introduced earlier in this chapter. Content-related validity would assume that if it logically asks questions concerning the specific nature and degree of pain for a patient, then it must be measuring the pain of the individual. Another example arises with the stadiometer: If the stadiometer is a ruler, and a ruler measures distance, then it must logically be able to measure height. Both of these examples show the use of a logical thought process to validate the measure as a truthful representation of what the instrument purports to measure.

The fact that a test has content validity does not always mean that the test is valid. Other nuances may add error to the test and negate the test's content validity. Consider the practice of measuring blood pressure at the arm, which is an accepted, valid method for measuring blood pressure. But what happens when the person obtaining the measurement is inexperienced or does not place the cuff in the proper position? The result will be an invalid measurement owing to the use of an improper measurement procedure. Any deviations in measurement procedures decrease the reliability of the test, thereby invalidating the data collected with the instrument.

The criteria for content-related validity can be traced back to the process used in developing the test, the interpretation of the results, and a well-defined protocol for collection of the data. In developing content-related validity, the researcher needs to be aware of extraneous factors that can affect the outcome of the test and render the test invalid. Whenever content-related validity for an instrument is relied upon, a set of strict guidelines concerning the use and collection methods of the instrument need to be in place to ensure that the validity of the instrument is not rendered useless by these factors.

Criterion-Related Validity

Criterion-related validity is based on a comparison between the test being used and some known criterion. According to the APA (1985), criterion-related validity involves "demonstrating test scores are systematically related to one or more known criteria" (p. 11). Criterion-related validity is the statistical validity identified earlier in this chapter. (Terms such as *statistical validity* and *correlational validity* are sometimes used as synonyms for *criterion-related validity.*) The same statistical technique used to determine reliability (i.e., PPM) is used to develop a validity coefficient.

Consider the following example: measurement of oxygen saturation of arterial blood in patients. The criterion for arterial saturation would be blood gas analysis from an arterial line; however, this type of measurement brings the risk of complications and should not be used during a routine office visit. An alternative method for measuring oxygen saturation is via an infrared monitoring device that attaches to the fingertip. The infrared monitor is minimally invasive, can be used with the general population without risk, and is supposedly valid for estimating arterial oxygen saturation. To verify that the alternative method of infrared monitoring is valid, a researcher would identify a small sample of patients, subject those patients to both tests, and compare their actual blood gas results with the infrared monitoring scores. The PPM would be calculated to quantify the comparison, which would be between the alternative test to be used and the known criterion. The results would have a validity coefficient associated with the infrared monitoring model instead of a reliability coefficient. Interpretation would be done in the same manner used to interpret the reliability coefficient.

Criterion-related validity can be subdivided into **concurrent validity** and **predictive validity**, based on the time between the collection of data using the alternative method test to be validated and the criterion measurement. Concurrent validity can use the PPM statistic for validity coefficient development. With predictive validity, however, the researcher is not limited to using the PPM correlation; a linear or logistic regression can be used to develop a validity coefficient. Concurrent validity coefficients are developed simultaneously for the criterion and the alternative method test, whereas predictive validity is not limited by time.

The arterial blood oxygen saturation testing described previously is an example of concurrent validity. In this example, both criterion and alternative method measures are collected at the same time to develop the validity coefficient.

The criterion in predictive validity can be measured years after the collection of alternative method test data. Testing for the occurrence of heart disease is an example of predictive validity. A patient's total cholesterol, high-density lipoprotein (HDL) cholesterol, and low-density lipoprotein (LDL) cholesterol levels, along with other measures, are used to predict the future occurrence of atherosclerosis. Atherosclerosis—the criterion in this example—does not occur until later in life, whereas the lipid profiles, which are the alternative method test, are collected years earlier. In the predictive validity example, if a PPM correlation is used, the validity coefficient might be low because the criterion measure is a nominal value. In this case, a researcher might use logistic regression techniques to predict the probability of occurrence and develop the validity coefficient from the probability of occurrence, rather than simply from the dichotomous variable (i.e., either a person does or does not have heart disease). A good point to remember is that whenever the criterion is a continuous variable, there is a better chance of having a high validity coefficient due to the possibility of improved true score variance and lower error score variance.

When a dichotomous or nominal variable is used as the criterion, a researcher should expect to have a lower validity coefficient due to a decline in true score variance and an increase in error score variance. An example of this mystery is presented in **Table 12-8**.

? THINK OUTSIDE THE BOX

Discuss how you could make sure that each person who collects data as part of a research project does the collection in the same manner to ensure reliability of the study results.

Table 12-8

Effects of Variable Scale on Validity Coefficient

Subject Number	Heart Disease (Yes or No)	Probability of Heart Disease	Total Cholesterol Level
1	0	40%	145
2	1	75%	200
3	1	89%	225
4	0	45%	170
5	0	30%	160
6	1	65%	195
7	0	40%	165
8	1	85%	250
9	1	88%	300
10	0	50%	180

Heart disease and total cholesterol $r = 0.777$

Probability of heart disease and total cholesterol $r = 0.879$

In this example, the criterion measure of atherosclerosis is presented both as a dichotomous variable and as a probability of occurrence based on a logistic regression formula. The alternative test for the validity coefficient is the total cholesterol levels of the subjects collected when they were 40 years of age. Notice that when a continuous variable is used as the criterion in this example, the validity coefficient is 10% higher compared with use of a dichotomous criterion. When using dichotomous variables as measures of validity, a researcher can expect to have lower validity coefficients than when using continuous variables. This decline in the validity coefficient reflects the lack of variability within the dichotomous measure—the lack of variability decreases the effectiveness of determining the true score of the measure. If the true score measure is decreased, then the error score measure is increased, which also affects reliability.

Many times, **cross-validation** techniques are used to develop a validity coefficient from a predictive validity criterion. Cross-validation simply implies that the researcher uses one group of subjects to develop the regression equation to predict the criterion and then gathers data from a second separate, but similar, group to develop the actual validity coefficient. Cross-validation techniques are generally used in developing new prediction models for a criterion.

ROC Analysis for Determining the Criterion-Related Validity of Diagnostic Exams

Because predictive analysis as a subsidiary of criterion-related validity compares an alternative method test to a criterion measurement for developing a validity coefficient, one can logically infer that ROC analysis may be used not only to describe the accuracy of a diagnostic test, but also as a measure of the validity of a diagnostic test. When using ROC analysis for validation of testing, the evidence-based practitioner can develop an inherent validity coefficient that quantifies the predictive quality of the alternative test to predict the presence or absence of the disease. An inherent validity coefficient quantifies the ability of the alternative diagnostic test to identify true positive and true negative results. The inherent validity coefficient is calculated using the following equation:

$$(\text{True positive tests} + \text{True negative tests}) / n$$

In the case of the modified Phalen's test from Table 12-7, the inherent validity is (20 + 39)/66 = 0.893, or 89.3% probability of correctly identifying true positive and true negative disease states (Bilkis et al., 2012).

Construct Validity

The most abstract of validity procedures is construct validity. **Construct validity** refers to the concept of "focusing on test scores that are associated with a psychological characteristic" (APA, 1985, p. 9). In practice, construct validity attempts to develop validity for measures that exist in theory but are unobservable.

The best example of this type of validity in EBP is the measure of pain perceived by a patient. Although we know pain exists, direct measurement of pain is somewhat convoluted and is affected by the psychological traits, tolerance levels, and perceptions of the patient. The tool most commonly used to measure pain today is the analog pain scale, which measures pain on a one-dimensional scale of 1 to 10. To develop a more precise pain scale that measures several dimensions of pain and has a high validity coefficient, constructs must be developed that can measure these traits associated with pain. Thus, we can think of construct validity as the combination of content validity and statistical validity to develop a validity coefficient for an abstract variable such as pain.

To develop construct validity of a variable, the variable must first be defined as specifically as possible. The researcher would then need to identify all of the constructs associated with the variable and to define them as specifically as possible. These definitions would prove helpful in developing the measurement scales and tools to quantify the variable. In the case of the pain example, pain might be defined as the degree to which a physical symptom causes discomfort at greater than normal levels for a patient. In using this definition, the constructs associated with this variable need to be identified and defined. Notice in the definition of pain that the term *degree* is used, which assumes that some type of quantifiable scale with specific unit differences is available to quantify the intensity and severity of the variable. Also, the term *discomfort* is used in the definition, which assumes that some type of non-wellbeing exists. In this case, intensity is one construct, severity is another construct, and discomfort is the final construct that needs to be defined and measured.

To start the process of developing a pain scale, think about the physical pain that you have experienced previously in relation to the constructs of intensity, severity, and discomfort. If your experience with pain is limited, you might seek the help of others who have more experience with pain or investigate current publications in pain research to help you with the definition and development of these constructs. For the current example, assume the definitions for your constructs are as follows:

- Intensity is the degree of pain.
- Severity is the degree of debilitation associated with pain.
- Discomfort is the degree of the measure associated with the patient's pain tolerance.

In this example, it is assumed that these three constructs are measurable and part of the content that defines the overall construct of pain.

Once you have defined the constructs, you need to determine the type of scale that can be used to measure each one. For intensity, you might decide to use a scale of 0 to 10, where 0 is defined as the absence of pain and 10 is defined as the most excruciating pain imaginable. For severity, you might have to develop a scale using terms that reflect a decline in functional capacity associated with debilitation. For discomfort, you might use a scale that reflects the type of pain, such as sharp, dull, or throbbing.

After developing the scales for the constructs that are included in the measurement of pain, you must determine how each scale should be weighted to reflect the absolute construct of pain. Again, you might want to rely on personal experience when developing your construct weights; alternatively, you might wish to seek expert opinions or explore previous research to help in developing your weighting system.

When you have accomplished this last step, you have a measure that logically measures pain (content validity). You are ready to test the merits of the measure by applying it to comparable groups to determine the statistical validity of the measure. In using statistical validation measures, you are attempting to prove the following hypothesis: Those individuals with diseases that are not associated with pain should score low on the new pain scale, and those individuals with diseases or disorders associated with a high level of pain should score high on the new pain scale.

By combining the logical validation of the pain scale with the statistical interpretation of the pain scale, you have developed construct validity for a measure of pain. As you become more comfortable with the process of developing construct validity for abstract or unobservable measures, you will find that the greater the number of definable constructs, the greater the validity gained by the measure.

Conclusion

This chapter focused on two key principles that determine trustworthiness of research data: reliability and validity. Whereas reliability and relevance can exist independently of each other, validity cannot exist without the presence of both reliability and relevance.

The two basic statistical techniques used to determine reliability and validity are the PPM correlation and the ANOVA test. As with most techniques, the selection of which to use is based on the number of variables being compared. When there are only two variables, a researcher would use PPM; when more than two variables are being compared, the ANOVA technique would be used. Both techniques generate a coefficient between an absolute value of 0 and 1.0, and the presence of a coefficient greater than 1.0 signifies an error in the calculations.

The interpretation of the coefficient is the only change that should occur regardless of the technique used. In the case of reliability, the coefficient can be used to interpret the consistency, stability, equivalency, or objectivity of the measure depending on which aspects were used to determine the estimate. A researcher can also use the reliability coefficient in conjunction with the standard deviation of the sample to determine the accuracy of the measure using the SEM equation. With reliability and accuracy determined, a nurse can be sure that comparable measures are similar and can be interpreted as consistent, stable, equivalent, or objective within a defined range of error. In the case of validity, these techniques can be used to develop a validity coefficient for concurrent validity or predictive validity based on the time between the collection using the alternative method test, or a validity coefficient for construct validity to improve the interpretation of the measure beyond simple content validation.

Summary Points

1. Trustworthiness of study data is only as good as the instruments or tests used to collect the data.
2. Reliability and validity are the most important concepts in the decision-making process when designing research studies.

3. Reliability is the determination that an instrument consistently measures the same thing.
4. Validity is the determination that an instrument measures what it is supposed to measure.
5. Validity cannot exist without reliability and relevance.
6. Reliability and relevance can exist independently of validity.
7. The correlation coefficient is the degree (positive or negative) of the relationship between the variables.
8. Interclass reliability is the consistency between two measures that are presented in the data as either variables or trials.
9. The three types of interclass reliability are consistency, equivalency, and internal consistency.
10. Intraclass reliability allows for the development of a reliability coefficient for more than two variables.
11. Within intraclass reliability, objectivity and accuracy need to be considered.
12. The three types of validity are content-related validity, criterion-related validity, and construct-related validity.
13. Content-related validity is the level at which a sample of items, tasks, or questions represent the defined content.
14. Criterion-related validity reflects the demonstration that test scores are systematically related to one or more identified measures.
15. Criterion-related validity is subdivided into concurrent validity and predictive validity.
16. Construct-related validity concentrates on the test scores that are associated with a psychological characteristic.
17. A receiver operand characteristics (ROC) analysis can be used for determining the criterion-related validity of diagnostic exams.

RED FLAGS

- If reliability and validity are missing from the data, an informed decision concerning the trustworthiness of the results of a research study cannot be made.
- If validity is documented in a study without any indication of reliability and relevance, concerns about the trustworthiness of the results should be raised.
- If a tool is documented as being used within a study, the report should provide information concerning the validity and reliability indices for the tool.

Multiple-Choice Questions

1. When making good decisions in evidence-based practice, the ______ of the data is necessary.

   A. Confirmability
   B. Trustworthiness
   C. Independence
   D. Timing

2. Reliability is defined as the case in which an instrument:

   A. Consistently measures the same thing.
   B. Measures what it is supposed to measure.
   C. Measures demographic data.
   D. Consistently measures the same sample.

3. Reliability and relevance may exist:

   A. With dependence on validity.
   B. With only independence of validity.
   C. Independently of validity.
   D. None of the above.

4. A valid test will ______ have some degree of reliability and relevance.

   A. Never
   B. Sometimes
   C. Frequently
   D. Always

5. When measuring blood pressure, the actual score is the:

   A. Observed score on the instrument.
   B. Estimated score determined by the nurse.
   C. Perfect score without error.
   D. First sound heard by the nurse.

6. Reliability coefficients greater than ______ are considered to be high.

   A. 0.50
   B. 0.60
   C. 0.70
   D. 0.80

7. As an example of consistency and stability in EBP, when a urinalysis is done four times in a 24-hour period, the urine sample needs to be the ______ amount.

A. Correct
B. Same
C. Smallest
D. Largest

8. Dividing scores on a pain questionnaire (with 0–5 items) into odd-numbered and even-numbered scores is a mechanism that can be used to determine:

A. External consistency.
B. Relevance.
C. Internal consistency.
D. Validity.

9. A research study was developed to consider the assessment of skin color. Nurses on a medical–surgical unit were asked to record their judgments of the skin color from four pictures of individuals with differing skin tones. This process is an example of which area of reliability measurement?

A. Accuracy
B. Objectivity
C. Feasibility
D. Equivalency

10. Which test is used to establish the measurement of the accuracy related to reliability?

A. ANOVA
B. Standard error of measure (SEM)
C. Pearson Product Moment (PPM) correlation
D. Reliability coefficient

11. To establish a test as an accurate measurement of reliability, the test scores need a relatively ______ standard deviation and a ______ reliability coefficient.

A. High; high
B. Low; low
C. Low; high
D. High; low

12. Factors that can affect the reliability, objectivity, and accuracy of a tool or test include:

A. Practice, timing, and environment.
B. Fatigue, subjects, and environment.
C. Precision, homogeneity of the test conditions, and the researcher.
D. Sequencing, practice, and level of ease.

13. Validity can be classified as:

A. Universal.
B. Concise.
C. General.
D. Logical.

14. A criterion for content-related validity determination is:

A. The inclusion of extraneous variables.
B. Establishment of brief guidelines for using the tool.
C. A well-defined protocol for data collection.
D. The clarification of nuances that might add errors.

15. A researcher was comparing alternative methods for establishing a child's core body temperature for a study. The testing included the measurement of anal, oral, and aural temperatures. This example reflects which type of validity determination?

A. Construct-related validity
B. Criterion-related validity
C. Content-related validity
D. Predictive validity

16. A study presented the results from the development of a new tool. This tool was established to measure the level of anxiety perceived by children. Which type of validity would this study need to document for the tool?

A. Content-related validity
B. Criterion-related validity
C. Construct-related validity
D. Concurrent validity

Discussion Questions

Use the following data to answer questions 1–4.

Patient #	Oral Temperature (°F)	Temperature (°F)
1	98.6	98.7
2	99.4	99.3
3	101.2	101.3
4	98.6	98.6
5	100.5	100.7
6	99.7	99.4
7	101.0	101.1
8	98.4	98.6
9	102.9	102.5
10	103.1	102.9

Patient #	Oral Temperature (°F)	Tympanic Temperature (°F)
1	98.6	98.7
2	99.4	99.3
3	101.2	101.3
4	98.6	98.6
5	100.5	100.7
6	99.7	99.4
7	101.0	101.1
8	98.4	98.6
9	102.9	102.5
10	103.1	102.9

1. Is tympanic temperature a similar measure of temperature?

2. Which type of reliability coefficient have you developed with these data?

3. What is the accuracy of tympanic temperature?

4. Is tympanic temperature a valid measure of patient temperature based on the information provided in the second table?

5. Using the example of the pain scale provided in the text, define and develop five additional constructs that might be used to measure pain.

Suggested Readings

Baumgartner, T., & Jackson, A. J. (1999). *Measurement for evaluation in physical education and exercise science* (6th ed.). Dubuque, IA: McGraw-Hill.

Cunningham, G. K. (1986). *Educational and psychological measurement*. New York, NY: Macmillan.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *Qualitative Report, 8*(4), 597–607.

MedicalBiostatistics.com. (n.d.). Sensitivity-specificity, Bayes' rule and predictivities. Retrieved from http://www.medicalbiostatistics.com/Sensitivity-specificity.pdf

Morrow, J. R., Jackson, A. W., Disch, J. G., & Mood, D. P. (2000). *Measurement and evaluation in human performance* (2nd ed.). Champaign, IL: Human Kinetics.

Thomas, J., & Nelson, J. (2005). *Research methods in physical activity* (5th ed.). Champaign, IL: Human Kinetics.

References

American Psychological Association (APA). (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

Bilkis, S., Loveman, D. M., Eldridge, J. A., Ali, S. A., Kadir, A., & McConathy, W. (2012). Modified Phalen's test as an aid in diagnosing carpal tunnel syndrome. *Arthritis Care & Research*, *64*(2), 287–289. doi:10.1002/acr.20664

Hanna, K. M., Weaver, M. T., Slaven, J. E., Fortenberry, J. D., & DiMeglio, L. A. (2014). Diabetes-related quality of life and the demands and burdens of diabetes care among emerging adults with type 1 diabetes in the year after high school graduation. *Research in Nursing & Health*, *37*, 339–408. doi:10.1002/nur.21620

Iverson, K. M., Huang, K., Wells, S. Y., Wright, J. D., Gerber, M. R., & Wiltsey-Stirman, S. (2014). Women veterans' preferences for intimate partner violence screening and response procedures within the Veterans Health Administration. *Research in Nursing & Health*, *37*, 302–311. doi:10.1002/nurs.21602.

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, *115*(5), 654–657.