CHAPTER 4

Reliability

Achieving consistency in research is as complicated as it is in everyday life. We may often have the expectation that most things we plan for on a daily basis are actually going to happen. Whether you are in the working world or a college student, you are faced with the daily task of getting from where you live to either work or college. Regardless if you get to your expected destination by train, car, bicycle, or whatever mode of transportation you take, you have the expectation that it will consistently get you where you need to be. What would you do if every time you get in your car, you are faced with never knowing if the car will start or not? What would you do if sometimes the train does not arrive at the time it is supposed to or if it stops running during the commute? We all have the expectation that there is a level of consistency with everything we do.

With that being said, research is no different. We expect some level of consistency when conducting research. The process of consistency in research is referred to as reliability. Prior to beginning a discussion on reliability, it is logical to ask, "What is **reliability**?" Reliability has a variety of different definitions such as the extent to which a measure is dependable or consistent (Gatewood & Field, 2001), the consistency of a measure across subsequent tests or over time, the stability of results on a measure, the preciseness of a measure, systematic or consist scores (Schwab, 2005), consistency (Shadish, Cook, & Campbell, 2002), or the degree to which the results can be replicated under similar conditions (McBride, 2010). Regardless of the definition, the common theme among the various definitions is that when a measure is reliable, then the results are consistent, dependable, precise, or stable. Reliability is based on probability with a reliability coefficient ranging from 0 to 1. A reliability coefficient of 1 would mean that there is 100% reliability in the measure, and a reliability coefficient of 0 would mean that there is 0% reliability in the measure.

In addition to reliability, another important concept is discussed in Chapters 5 and 6, which is **validity**. Validity is related to the accuracy of the results or process in a study. Not only are we concerned about how consistent or reliable the measures used in an experiment are, but we also need to ensure that these results are accurate or valid.

reliability: The extent to which a measure or process is consistent, dependable, precise or stable

validity: The extent to which a measure or process is accurate

The problem with measuring variables within an organization is that human behavior or any process relying on human interaction is not always 100% predictable. There will always be some variation within the measurement of any variable.

Regardless of this variability, reliability is important for two reasons:

1. Reliability is a necessary, but not sufficient condition for validity.
2. Reliability is the upper limit to validity.

The first statement implies that having a reliable measure does not mean that it will always be valid. For example, you may have a scale that consistently measures an individual's weight. However, if the scale is set back five pounds without anyone knowing, then the weight is not valid. The second statement implies that the validity coefficient will never be higher than the reliability coefficient. This means that if the reliability coefficient that is calculated from your measure is 0.6, then the validity coefficient cannot be higher than 0.6. This is critical because when studying human behavior, no one test is perfect.

RELIABILITY THEORIES

As previously mentioned, reliability is concerned with the consistency of a measure with the goal of reducing errors in measurement. Almost every measure of human behavior has some degree of error associated with the tool. Reliability errors are referred to as **random errors** and **systematic errors**, but the terms *random errors* and *nonrandom errors* may be used respectively.

While the purpose of this book is to provide a researcher with the tools to conduct well- developed applied research or evaluate existing research, we must still refer to some critical theoretical concepts. Two such theories for errors in measurement are as follows:

1. classical test theory or true score theory
2. generalizability theory

The purpose of **classical test theory** or **true score theory** is based on an assumption that measurement error exists. This theory is derived from the thought that a raw score (X) of a measure is comprised of a true component (T) and a random error (E) component, such that the formula for a raw score is $X = T + E$. The true component portion of the formula represents the score that the participant received on a measure. The random error component represents the amount that the participant's score was influenced by other factors unrelated to the construct at the time the measurement was observed. The combination of the true component and random

random error: A type of error in measurement where any factor or variable randomly has an impact of the measured variable

systematic error: A type of error in measurement where any factor or variable that is not random has an impact on the measured variable

classical test theory: Also referred to as True Score Theory. A measurement error theory derived from the thought has a raw score consists on a true and random component.

true score theory: Also referred to as classical test theory. A measurement error theory derived from the thought that a raw score consists of a true and random component.

error component is equal to the raw score or the actual score that was obtained from the measure used. This relationship implies that every measure used in an experiment has a portion of the result that truly represents the intended construct and that there is also some degree of error associated with the measurement. Based on this formula, it can be derived that when random error is reduced, then the true component is increased. In other words, removing the error that does not occur by chance, but is associated with a measure, causes the end result to be a more reliable measure.

generalizability theory: A measurement error theory extending the principles of classical test theory with the exception of not assuming a raw score is combined of a true and random error component but rather the distinction is between the reliability and validity of a measure

In addition to true score theory, a similar theory exists and is referred to as **generalizability theory**. This theory was first introduced by Cronbach, Gleser, Nanda, and Rajaratnam (1972) and has been said to extend the principles in classical test theory with the exception of not assuming a raw score is combined of a true score and random error component. The thought process behind this theory is that the distinction made between reliability and validity could be overcome through developing a set of observations that can be generalized from the sample collected to the population it was sampled from. In other words, the measures developed that utilize generalizability theory are referred to as dependable and generalizable rather than creating a distinction between validity and reliability. This line between validity and reliability is removed because measures are defined through maximizing construct validity. This means that before a measure is used in an experiment, it is properly operationalized or defined, and therefore, the measure will represent the intended construct. For a further discussion on generalizability theory, refer to Cronbach et al. (1972).

The bottom line to the discussion of these two theories is that there is always going to be some degree in measurement error. Whether this error is associated as both a true component and random error component or if the measurement is well developed to incorporate both validity and reliability is not the focus of this chapter. Measurement error does exist, and there are many theories on how to reduce this error. More importantly, there are many types of reliability that are critical to the developing of not only measures of constructs but also to the design of a research study. Knowing that measurement error exists, the next phase of research design is to determine what type of reliability is the most efficient with regards to research methodology. Similar to research design and variables, no one design or variable is better than the other. Each one has its own advantages and disadvantages, and they are all used with specific purposes.

GOALS OF RELIABILITY

When looking at reliability, there are five main goals, purposes, or types of reliability.

Test-Retest Reliability – Researchers may want to know if results are consistent when the same instrument is administered multiple times. For example, practitioners may want to assess job performance ratings over time to determine if the measures are consistent with the passage of time (Salgado, Moscoso, & Lado, 2003; Viswesvaran, Ones, & Schmidt, 1996).

Interrater Reliability or **Intrarater Reliability** – A researcher may desire to ensure that multiple items on a given survey or questionnaire produce a similar participant response to all the items on the survey or questionnaire: for example, a measure of job performance based on the same person rating performance versus different people rating performance (Viswesvaran et al., 1996).

Parallel Forms Reliability or **Equivalent Forms Reliability** – If multiple tests are developed to measure the same construct or variable, then these subsequent tests should measure similar items. For example, standardized tests, such as the graduate record exam (GRE), scholastic assessment test (SAT), or graduate management admission test (GMAT), etc. all have multiple versions that are designed to measure the same constructs. Bing, Stewart, and Davison (2009) examined the multiple forms of the Personnel Test for Industry-Numerical Test and an Employee Aptitude Survey Numerical Ability Test examining results based on using a calculator versus not using a calculator. To test this comparison, they utilized Forms A and B, which are two different, but identical tests for assessing ability and found support for both tests reliably assessing the same constructs.

Split Half Reliability – A test may be divided into multiple parts and compared to ensure they are measuring the same constructs. For example, Damitz, Manzey, Kleinmann, and Severin (2003) conducted a study examining the validity of using an assessment center to select pilots. The assessment center consisted of data on nine cognitive ability tests, four different assessment exercises measuring nine behavioral dimensions, and nine behaviorally anchored peer ratings on training performance. Split half reliability was calculated by using the peer ratings because each group of peers rated the same student. Therefore, they were randomly divided into two groups to calculate a mean rating for each group and then used a Spearman-Brown correction to estimate reliability.

Internal Consistency/Coefficient Alpha – Items that measure similar constructs that appear throughout a test should be related to each other. For example, Cheng, Huang, Li, and Hsu (2011) conducted a self-administered questionnaire to Taiwanese workers to examine the extent to which burnout had an impact on employment insecurity and workplace justice. In total, there were six items on employment insecurity and nine items on workplace justice. The purpose was to determine if these items measure the variables they were developed for.

Test-Retest Reliability

The first type of reliability is intuitive from the name of it. Test-retest reliability is when a researcher provides a participant with the same test at two different points in time. The purpose of this type of reliability is to show that scores are consistent on multiple administrations of the same test over time. When a test is found to have test-retest reliability, it is expected that a participant's scores on multiple administrations would be similar.

Salgado et al. (2003) examined the test-retest reliability with measures of job performance. Since supervisory ratings are frequently used for validation purposes within selection research (Barrick, Mount, & Judge, 2001), Salgado et al. (2003) conducted a study assessing the reliability of supervisory ratings on several dimensions of job performance and overall job performance. They found support that the test-retest reliability of overall job performance was 0.79 and other measures of performance ranged from 0.40 to 0.67: thus, providing support that there is test-retest reliability on ratings of performance.

Similarly, Sturman, Cheramie, and Cashen (2005) conducted a meta-analysis using 22 studies on the test-retest reliability of job performance ratings over time. They found test-retest reliability coefficients over time for low complexity jobs to be 0.83 and 0.50 for high complexity jobs. Despite these findings, they state that it is impossible to estimate the true stability of job performance because time is an important factor with impacting the job performance ratings.

The main issue with test-retest reliability as Sturman et al. (2005) point out is that the difference in measures between the first and second administration could impact the reliability due to the following factors:

1. Time interval between test administrations
2. The test or other factors associated with the participant

With respect to the time interval, a researcher measured job dissatisfaction through negative affectivity and hypothesized that this measure is stable over time. As a result, the researcher measured the negative affectivity 1 year later. In this case, the time lapse between both administrations of the same test may influence the reliability of the measurement. The result was that the true value of this measure a year later may have been underestimated. The error in measurement was associated to transient factors, such as the participant's mood, emotion, or feeling at the time. These measures could be different a week or day later and

test-retest reliability: The consistency to which the test scores are similar when participants are given the same test more than once

interrater reliability: The consistency to which the test scores are similar when participants are given the same test more than once

intrarater reliability: The extent to which measurement ratings are consistent among the same raters

parallel forms reliability: Also referred to as equivalent forms reliability. The extent to which two tests are developed to measure the same construct of interest.

equivalent forms reliability: Or also referred to as parallel forms reliability. The extent to which two tests are developed to measure the same construct of interest.

split half reliability: Measures the internal consistency of items on a test when different items assessing the same construct throughout the test are compared

internal consistency: Also referred to as coefficient alpha. Measures the consistency of the same items on a test that measure the same construct.

coefficient alpha: Also referred to as internal consistency reliability. Measures the consistency of the same items on a test that measure the same construct.

are not actual measures that impact the negative affectivity the researcher intended to measure (Schmidt & Hunter, 1996).

In addition to the time interval between administrations, a possibility exists that the scores from the first administration can influence the results of the second administration. This means that the participant taking a test the second time may learn from his or her mistakes during the first administration or review the items being asked after the test is completed. This would be a greater possibility when the time between administrations of the same test is a short time interval. For example, a researcher administers a vocabulary test once and then a month later the same test is administered. In this case, the test-retest reliability may be overestimated due to the lack of controlling for specific factors, such as personality factors, with each administration of the test (Schmidt & Hunter, 1996).

Interrater/Intrarater Reliability

The second type of reliability is interrater and intrarater reliability. The concept behind this reliability is that either the same (interrater) or different (intrarater) raters assess an individual rating on a specific variable. The challenge is that each rater, regardless if they are the same or not, must consistently assess the same behavior in the same consistent manner.

Interrater reliability is defined as measuring the consistency of ratings across different raters. In a quasi-experiment, Lievens and Sanchez (2007) examined the impact that a frame-of-reference training would have on the interrater reliability of competency ratings completed by human resources consultants in either a training group or a control group. The purpose of the program was to have consultants determine the competencies that were required for a specific job. They found that the consultants that received the training resulted in an interrater reliability coefficient of 0.65 compared to the 0.44 coefficient found for the control group not receiving the training.

Intrarater reliability, on the other hand, is when a researcher examines the consistency of one particular individual's ratings at multiple points in time. Within the realm of applied research, intrarater reliability is assessed in conjunction with job analysis research (Dierdorff & Wilson, 2003). The purpose of intrarater reliability is to determine the sustainability of an individual's ratings at two different points in time. Dierdorff and Wilson (2003) conducted a meta-analysis of 46 studies to explore the reliability of job analysis data. They found in cases of measuring task data, intrarater reliability results were higher than interrater reliability results and in cases of measuring general work ability, interrater reliability results were higher than intrarater reliability.

Similarly, Viswesvaran et al. (1996) conducted a meta-analysis on overall job performance to compare and contrast the impact of interrater and intrarater reliability. In the meta-analysis, Viswesvaran et al. (1996) utilized 10 measures of job performance to assess the differences in reliability using both peer and supervisory performance ratings. They found that interrater reliability measures of overall job performance were lower than that of intrarater reliability and that supervisory ratings have higher interrater reliability than peer ratings.

Parallel or Equivalent Forms Reliability

The next type of reliability is parallel or equivalent forms reliability. By definition, this type of reliability is where a researcher creates two different but similar tests that measure the same construct. One of the more well-known tests that are parallel or equivalent forms is standardized tests. The process to developing a standardized test is extremely arduous and requires an extreme precision to ensure the psychometric properties of the items are similar. In practice, it is possible to create parallel or equivalent forms of a test, but it may not be widely used due to the process of developing multiple tests.

The idea behind parallel or equivalent forms reliability is to have two conceptually identical tests that utilize separate questions to measure the same construct of interest. The number of items used to measure a particular construct of interest can be unlimited. Therefore, it is not possible for a test or measure to include every possible item to measure the constructs of interest. This has an important implication on reliability because creating a test to measure human behavior with a reliability coefficient of 1.0 is unlikely. On the other hand, having multiple items to measure the same construct could be a benefit for using parallel or equivalent forms reliability to create multiple similar but different tests. The challenge you face when multiple tests are created to measure the same construct is that the items on both versions of the same test may not actually measure the same construct.

From an applied perspective, Chan, Schmitt, Deshon, Clause, and Delbridge (1997) were interested in the relationships that factors such as race, test-taking motivations, and performance had on a cognitive ability test. To do this, a parallel form cognitive ability test battery was created that was used in an actual employment testing project. They found that the correlation between the first test and the parallel test was 0.84 ($p < 0.05$). This indicates that the two forms of the cognitive ability test were adequate in regards to parallel form reliability.

Similarly, Bing et al. (2009) conducted a study that utilized multiple forms (Form A and B) of a Personal Test for Industry-Numerical Test and an Employee Aptitude Survey Numerical Ability Test that involved the comparison of results for participants using a calculator compared to participants not using a calculator. Multiple comparisons were conducted to examine the reliability of these different conditions, and they found support that the results of both the calculator and noncalculator condition were similar on both forms of the test.

Split Half Reliability

The next type of reliability is split half reliability. The purpose of the split half reliability is to divide the test or measure into two halves and test the internal consistency of the items used. Split half reliability is similar but different to the parallel or equivalent forms reliability with a couple exceptions. Parallel or equivalent forms reliability requires two versions of a test. With split half reliability, a researcher only conducts one administration of the test or measure and splits the test in half (i.e., even vs. odd questions or the first half vs. the second half). This is different from parallel or equivalent forms because multiple versions of a test are not necessary.

One common criticism of this technique is determining where to split the test because of how the items are divided within the test or measure. A few techniques to split the test or

Box 4.1 In-Basket Test Reliability

When looking to select the most appropriate employee for a position, it is important to properly evaluate their ability to succeed because the cost to replace an employee can be extremely costly. In the day and age of cutting budgets to conserve money, an employer must be able to consistently (reliability) and accurately (validity) select one qualified candidate from a pool of applicants. One methodology to select employees is through the use of an assessment center. Within an assessment center, an applicant is given a variety of test batteries that may include simulations, tests, exercises, etc. in which they are designed to perform in a simulated work environment (Berry, 2002).

One such test battery in an assessment center is an in-basket exercise. The purpose of this is to provide an applicant with the ability to manage a variety of issues that could be accumulated in a day such as letters, memos, telephone messages, reports, or other items that may come up throughout the course of a day (Berry, 2002). In an effort to assess the reliability of the in-basket test, Schippmann, Prien, and Katz (1990) reviewed the existing literature on various components of the reliability of an in-basket test. In terms of reliability, the psychometric properties of an in-basket test reliability was examined through interrater reliability, parallel forms reliability, and split-half reliability.

While none of these three reliability techniques proved superior in assessing the reliability of the in-basket test, a lot of useful information was learned. In terms of interrater reliability, it was found that the range of reliability coefficients for this technique suggests that some other variable may create the rating patterns that may be a function of rater training. Parallel form reliability differences in coefficients may potentially be a result of being confounded with performance on the test. Lastly, for split-half reliabilities of odd and even numbers, Schippmann et al. (1990) suggest that there may be a need in further developing the test content or a more systematic and objective approach to scoring the test may yield more encouraging reliability coefficients. In summary, the in-basket test for reliability and validity provides only modest support for the usefulness.

measure include using odd and even numbered items, randomly selecting the items, or using the first half and second half of the test. The most commonly used method of split-half reliability within research is through odd and even items (Aamodt, 2007). As an example, Damitz et al. (2003) examined the validity of an assessment center used to select pilots. As a part of the assessment center, each group of peers had rated the same students and therefore, they randomly divided the group into two equal-sized subgroups. This grouping allowed for calculation of split half reliability utilizing the Spearman-Brown correction.

Internal Consistency/Coefficient Alpha

The last reliability technique is internal consistency reliability and is also referred to as coefficient alpha or Cronbach's alpha. This is most common and widely used reliability technique for purposes of reporting the reliability of a test or measure in experiments in applied settings (Edwards, Scott, & Raju, 2003) and is also by far the most commonly reported measure of reliability (Hogan, Benjamin, & Brezinski, 2000; Rogers, Schmitt, & Mullins, 2002). It is similar to split half reliability because this technique also measures the internal consistency or correlation between the items on a test. The main difference between split half and internal consistency/coefficient alpha is the entire test is used to estimate the correlation between the items without splitting the test or measure in half. The correlation utilized to calculate internal consistency is similar to the correlation used with inter/intrarater reliability.

Internal consistency is calculated by examining the pairwise correlations between the items on the measure. Cronbach (1951) outlined that a coefficient alpha of greater than or equal to 0.7 is generally acceptable. Despite the widespread use of Cronbach's alpha, there are a couple of caveats to its use. First, alpha is strongly influenced by the number of items on a measure, so the calculated alpha could be higher by increasing the number of items. The other problem with alpha is when it is too high, because a very high Cronbach alpha could indicate redundancy in the items.

In addition to examining parallel or equivalent forms reliability, Bing et al. (2009) also assessed internal consistency reliability for the 30 measures on mathematical reasoning and the 75 items on computational skill. Coefficient alpha for the measures were above 0.7, which by Cronbach's (1951) standards is acceptable. This means that the 30 measures included within the test for mathematical reasoning and the 75 items on computational skill reliably assessed their respective constructs.

In another study, Dirani and Kuchinke (2011) conducted a survey using a convenient sample of Lebanese banks to assess the validity and reliability of two measures of organizational commitment and job satisfaction. The survey consisted of 38 items comprised of three sections. Job satisfaction consisted of 20 items, nine items consisted of organizational commitment, and nine items assessed demographic questions. The results from the current study replicated previous reliability results with a coefficient alpha of greater than 0.84, thus indicating that the measures used to assess job satisfaction and organizational commitment were reliable.

RELIABILITY SUMMARY

While all types of reliability are important, internal consistency/coefficient alpha is the most widely used in applied settings (Edwards et al., 2003). The main reason why this type of reliability is the most widely used is because it only requires one administration of a test to determine the relationship between the items on the test. Other types of reliabilities need multiple versions of a test, many different raters, or multiple administrations of a test to generate a reliability coefficient. In applied settings, a researcher may not have the time, resources, or availability to conduct multiple administrations of a test.

As we know, a measure can be reliable and not valid, but it cannot be valid and not reliable. Additionally, reliability is a necessary, but not sufficient, condition for validity. Therefore, it is important to take a look at the reliability coefficient from the perspective of validity to better understand the relationship between reliability and validity and ensure the measure is both reliable and valid. Keep in mind that regardless of the type of reliability, the goal is to have a reliability coefficient close to 1, which would indicate a high degree of consistency and a low degree of measurement error with the measure. Simply having a high reliability coefficient in your study does not necessarily equate to the assumption that the measure on your test is valid. The reason is because there may be potential threats to validity that can provide an explanation to a high reliability coefficient.

When discussing the different reliability methods, the main conclusion drawn between the results of the different types of reliability is the explanation of the results found within the research. For example, when evaluating research that states the best predictor of future performance is past behavior, you, as a consumer of information, have to know that this result is true and that there is no other explanation that can justify this result. Whenever a measure of human behavior exists, there is some level of measurement error that occurs. While the reliability coefficient ranges from 0 to 1, with 1 being perfectly reliable, we know that no measure of human behavior is capable of achieving perfect reliability. There is bound to be some error associated with any measurement. Even classical test theory posits that the raw score of a measure is comprised of a true component and a random error component.

Therefore, the goal of being a consumer of information is to know and understand the various aspects of reliability techniques as well as understand the relationship between reliability and validity. Whenever a possibility exists that the relationship within an experiment can be explained by alternative explanations this means that there is a threat to the validity of the experiment and the reliability of the results. **Validity** is discussed in detail in Chapters 5 and 6. An alternative explanation for a result means that some other variable can explain the relationship between the cause and effect relationship.

validity: The accuracy of the results of a research study

CHAPTER SUMMARY

- Reliability is the consistency of a measure with a coefficient between 0 and 1 and is important for two reasons: (1) Reliability is a necessary, but not sufficient, condition for validity and (2) reliability is the upper limit to validity. This implies that a reliable measure may not always be valid and a validity coefficient can never be higher than the reliability coefficient.
- When developing tests to measure a construct, it may not be perfect, and there could be a degree of error associated with the test, but techniques can be utilized to improve the reliability of a test. Error in measurement can be categorized as random or systematic (nonrandom) errors.
- All tests or measures have some degree of error associated with the measurement, and there are two theories aimed at understanding these errors in measurement. Classical test theory or true score theory is based on an assumption that every raw score observation is comprised of two components, which are a true measurement and an error measurement. Generalizability theory extends the principles of classical test theory/true score theory by the premise of developing a set of observations that can be generalized from the sample collected to the population it was sampled from.

- There are five main goals of reliability and they relate to the different types of reliability, which are as follows: test-retest, interrater/intrarater, parallel or equivalent forms, split half, and internal consistency/coefficient alpha reliability. These different types of reliability are aimed at ensuring that a test consistently measures a construct of interest.

DISCUSSION QUESTIONS

- What are some ways that a researcher or practitioner can reduce systematic errors within a study design?
- How does classical test theory/true score theory or generalizability theory apply to research design?
- Given internal consistency reliability is the most commonly reported reliability technique, how might you use split half, parallel forms, intra/inter or test-retest reliability to demonstrate the consistency of your measures?

CHAPTER KEY TERMS

Classical Test Theory
Generalizability Theory
Random Error
Reliability
Reliability, Coefficient Alpha
Reliability, Equivalent Forms
Reliability, Internal Consistency
Reliability, Interrater
Reliability, Intrarater
Reliability, Parallel Forms
Reliability, Split Half
Reliability, Test-Retest
Systematic Error
True Score Theory
Validity