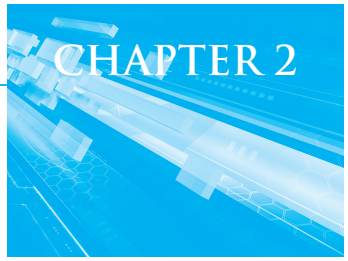
CHAPTER 2



Practical Aspects of Testing

CONTENTS

KEY WORDS

classical test theory
construct
discrimination
fairness
mass testability
measurement
measurement error
measurement instrument
practicality
pretest planning
reliability
test
validity

OBJECTIVES

In the first part of this chapter, we will discuss the nature of testing and present examples of testing used in the broad range of disciplines and constructs within kinesiology. We then present the order of stages involved in the test process, from selecting the test to the final evaluation of test scores and providing feedback to the test-takers, with particular attention paid to the practical elements of administering a test. The most important characteristics of tests are then described, and then we draw your attention to the ethical responsibilities of the test user.

After reading this chapter, you should be able to:

1. Describe the key elements of the five stages of testing.
2. Describe practical ways by which the tester can minimize measurement error during test administration.
3. Identify the important characteristics of a test.
4. Describe the ethical responsibilities of test users.

Introduction to Testing

Almost everyone has been tested at some point in their lives. Within minutes of being born, many of us were given a numeric score. The APGAR (Appearance, Pulse, Grimace, Activity, Respiration) assessment is used to provide a numeric score that is evaluated against a set of criterion-referenced standards for determining the health of newborns. Throughout our early months, measurements are taken of our height and weight and other body dimensions, such as head circumference, in order to monitor our growth and development. This is done by comparing our scores against norm-referenced standards from other children of a similar age and sex. We continue to be tested throughout our school years, most notably on our learning and knowledge, and testing continues throughout the rest of our lives. Information from tests is used to make important decisions about us; admission to college, job offers, promotions, and other important life events hang on the outcomes of testing situations. We therefore have plenty of experience with tests from the perspective of a *test-taker*, and of having decisions made about us by others, based on our test scores. In this chapter, we provide insight into testing from the perspective of the *test user* or *administrator*.

Testing has several functions, or purposes, and is important for societal reasons with regards to health, fitness, and education. Kinesiology covers a broad range of disciplines and professions, such as exercise science, physical education, physical therapy, coaching, and sport management. Within these subdisciplines of kinesiology there is an array of different types of tests to choose from, but the principles of testing remain broadly the same regardless of the type of test you are using. Testing situations can vary from individual testing that takes up to 1 or 2 hours to accomplish (e.g., a maximal treadmill exercise test given to a single client or clinical

patient) to mass testing, in which several people are tested simultaneously over a relatively short period (such as a PACER test given to a whole class). In some situations, the testing is performed by a professional, such as a physical education teacher or clinical exercise physiologist; in other situations, we may need to rely on the participant or co-participant to conduct the test. An example of the latter would be a participant submitting pedometer scores via an online survey or pairs of students in a class counting each other's push-ups. Within kinesiology, therefore, the nature of the testing situation varies considerably. In this chapter, we will guide you through the factors you should consider and help you to decide which factors are the most important in any given testing situation. Throughout, practical examples will be used to illustrate the concepts, so that when you meet terms such as *reliability*, *validity*, and *evaluation* in your studies and research you will understand them in the context of how you will apply them in practical testing situations in kinesiology.

The Nature and Purpose of Testing

When you hear the word *test*, what comes to mind? For most people, a test was something they took at school (also called an *exam*), requiring them to respond to a set of questions in order to assess their level of learning and knowledge. Much of measurement theory and practice in kinesiology evolved from educational testing, and it is sometimes useful to think of a test in this way. Drawing on personal experiences can be helpful in understanding the relevant principles of testing. However, tests, or *measurement instruments*, in kinesiology take many more shapes and forms than a paper-and-pencil exam and can pose some unique challenges to the tester. **Table 2.1** lists several examples of types of tests used in kinesiology. It lists the tests by category and describes the construct measured by each test.

The *construct* is the underlying characteristic we wish to assess. Although some constructs in kinesiology are very tangible and concrete, such

TABLE 2.1 Examples of Measurement Instruments Within Kinesiology

Type of Test	Example Test	Construct
Physical performance	Loughborough Soccer Passing Test (LSPT)	Soccer-playing ability
	PACER test	Aerobic fitness
Written	BREQ-2	Motivation
	Physical Activity Questionnaire for Adolescents (PAQ-A)	Physical activity
Technological	Body plethysmograph	Body fatness
	Heart rate monitor	Exercise intensity
Clinical	Fasting blood glucose	Cardiometabolic health
	Standardized Assessment of Concussion (SAC)	Concussion

as height or strength, others are more abstract and less directly observable. For psychological constructs such as motivation or attitude, we have to use an indirect measure (such as by asking participants to respond to questions that aim to tap into their underlying thoughts and emotions). The Behavioral Regulations in Exercise Questionnaire (BREQ; Mullan, Markland, & Ingledew, 1997) poses a series of statements such as "It's important to me to exercise regularly" and "I value the benefits of exercise" in order to measure motivation. Participants respond on a numeric scale ranging from 0 ("Not true for me") to 4 ("Very true for me"). Their responses are indicative of their underlying level of motivation for exercise. Even quite straightforward constructs such as body fatness often require indirect methods of measurement, because much of our body fat is hidden deep within our body, around our organs. The body plethysmograph (or Bod Pod) measures body fatness by enclosing the body in an airtight box and measuring air displacement. It thus allows us to estimate the body's volume, from which we can calculate body density, and, using assumptions about the density of fat and fat-free tissue, we then estimate the percent of body weight that is body fat. This is quite indirect!

The term *testing* therefore is used synonymously with *measurement*. *Measurement* describes the application of a process to an individual or group of individuals that results in them obtaining a score (usually, but not always, a numeric score). A *test* is the tool or process that is applied. From Table 2.1 and the previous examples, you will see that the measurement instruments in kinesiology can consist of several components, including the necessary equipment; a measurement protocol, or set of procedures; a response format or score format; a tester or team of testers; and the participant. The context and purpose of testing also varies within the disciplines of kinesiology. For example, tests can serve an educational purpose. Not only do they provide teachers with valuable information about their students, but the test can itself be a learning experience. By participating in the Loughborough Soccer Passing Test (LSPT; Ali et al., 2007), students can learn soccer skills in a controlled, safe environment and use their scores to monitor their own learning. This information can be used for formative evaluation purposes without counting toward grades. Tests can be used for diagnosis in health settings. Fasting blood glucose levels can help to screen for poor glucose regulation, one of the components of the metabolic syndrome, which is a prediabetic health condition that can be improved via physical activity (Ha, Kang, Han, & Hong, 2014). Within sports medicine, tests are often used to diagnose injury. Quite recently, much attention has been paid to the validation of tests such as the Standardized Assessment of Concussion (SAC; McElhiney, Kang, Starkey, & Ragan, 2014) for diagnosing concussion in order to decide about return to the playing field. This recognizes the short- and long-term consequences of playing sport following mild traumatic brain injury. Sporting competitions are also a form of testing context. The rules and regulations that govern an event such as the high jump have much in common with the instructions for administering many tests—there is a correct way of completing the task, and incorrect attempts are not counted, similar to many performance tests. Tests are also used widely in a research context in order to understand human behavior or to evaluate the success of an educational or health intervention. This is the basis for evidence-based practice; professionals in kinesiology should use practices and programs that have been demonstrated to be effective, and testing is a core part of the process leading to that evidence.

The practical use of tests therefore requires an understanding of the nature of tests from both a theoretical and pragmatic perspective. A one-size-fits-all approach to testing does not work in kinesiology because of the variety of testing methods and purposes, but many general principles apply to all testing situations, and an understanding of the practical aspects of testing allows us to determine which principles are relevant in a given testing situation.

Stages of the Testing Process

Regardless of the reasons for testing and the testing context, certain stages of the testing process should be completed. At each stage, you should apply the information provided in this text to make decisions about how you will use testing. In this section, we describe the stages of test selection, test preparation, test administration, data processing, and decision making and feedback.

Test Selection

The first step in testing is to decide what test you should use. The Internet now makes it easy to search for existing tests. This is very helpful because it enables us to not only locate existing tests, but also to search for evidence supporting their use. Important questions to ask yourself at this point are: "What construct do I want to test?" "Who do I plan to test?" "What is my purpose in administering this test?" and "What decisions do I hope to make using these test scores?" When searching the Internet, search engines such as Google, Yahoo, and Bing work perfectly well. It is important to bear in mind that (1) not all sources of online information are trustworthy and (2) not everyone may call the construct by the same name that you are familiar with.

In terms of trustworthiness, look for information that was generated by impartial and informed organizations, such as universities and professional bodies. The latter include the American College of Sports Medicine (ACSM), the Society of Health and Physical Educators (SHAPE-America, formerly known as AAHPERD), and the National Athletic Trainers' Association (NATA). When entering search terms, familiarize yourself with alternative names for the constructs you are interested in testing. For example, aerobic fitness is described by many alternative terms—the terms *cardiovascular* and *cardiorespiratory* are often used instead of *aerobic*, and *endurance* is often used instead of *fitness*. Body fatness may also be called *body composition*, physical activity may also be called *exercise*, concussion may also be called *traumatic brain injury*, and so on. In psychological measurement, the "jingle-jangle fallacies" refer to the fact that two tests with similar names may not measure the same construct (the jingle fallacy), and, conversely, two tests with very different names may not test different constructs (the jangle fallacy). Essentially, do not assume that the name of a test is proof that it measures a certain construct. For example, the Attitude Toward Physical Activity Inventory (ATPA; Kenyon, 1968) does not measure what we typically mean by "attitude" (a positive or negative disposition toward something). Instead, it is designed to assess people's motives for participating in physical activity (e.g., to experience the beauty or artistry of movement, to meet people and maintain friendships, to experience the element of thrill, or to improve health and fitness). The characteristics you should look for in selecting a test are presented in a later section of this chapter.

Test Preparation

Having selected an appropriate test, the next step is to prepare for testing. The goal is to be confident that our preparation is adequate and that the test administration will proceed smoothly. *Pretest planning* is all of the preparation that occurs before test administration. It involves a number of tasks: knowing the test, developing test procedures, developing directions, preparing the individuals to be tested, planning warm-up and test trials, securing equipment and preparing the test facility, making scoring sheets, estimating the time needed, and practicing the test.

Whenever you plan to administer a test for the first time, read the directions or test manual several times. This is the best way to avoid overlooking small details about procedures, dimensions,

necessary equipment, and the like. Once you are familiar with the test, develop procedures for administering it. These may be described in the existing test manual, but often you will need to consider the specific constraints of the situation, such as how many testers you have, how much time is available, what equipment can be used, and other factors specific to each testing situation. These include selecting the most efficient testing procedure, deciding whether to test all the individuals together or in groups, and determining whether one person will do all the testing or pairs of individuals will test each other. If you plan to administer several performance tests on the same day, consider the optimal order to present the tests. Resting measures should be administered first, as well as collection of any demographic information. This helps to put the participant at ease and allows you to introduce yourself and assess the participant's readiness for being tested. Do not give consecutive tests that tire the same muscle groups. Also, plan to administer very fatiguing events, such as distance runs or other maximum exertion tests, last. If a questionnaire is extremely long, plan to give the test-taker a small break in the middle to avoid "questionnaire fatigue." Be familiar with the exact scoring requirements and units of measurement. If you have never administered a specific test before, try one or two practice administrations before the actual test. This is a good way not only to see what changes and additions to the procedures must be made, but also to train testing personnel.

After you have determined the procedures, you should develop instructions for the participants in order to standardize the procedures across all participants. If you retest participants at a later date (in order to assess progress, for example), having written directions will help to ensure that the test is run the same way both times. These can be read to the participant before administering the test. The directions should be easy to understand and should specify the following:

1. Administration procedures
2. Instructions on performance
3. Scoring procedure and the policy on incorrect performance
4. Pointers on key techniques or aspects of performance to ensure trustworthy scores

Test Administration

If you have planned properly, the testing should go smoothly. Although your primary concern on the day of the test is the administration of the test itself, you should also be concerned with participant preparation, motivation, and safety. If, after you have administered the test, you can say, "The participants were prepared and the test was administered in a way that I would have liked if I were being tested," the testing session will have been a successful experience for the test-taker. As the test administrator, your primary goal is to obtain test scores that are as close as possible to each participant's true score. To understand how to achieve this, consider the following formula:

$$X = t + e$$

This mathematical formula comes from *classical test theory*. In classical test theory, the X in the formula represents the score that a person receives, or obtains, as a result of being tested. The t represents the person's "true" score (the score the person would, or should, get if the test works perfectly). The letter e refers to *measurement error*. This is the sum of all of the sources of error that affect the test-taker's score. Some sources of error will have a numerically negative effect on the participant's score. An example would be that a skinfold tester pinched the skinfold too hard and placed the caliper too close to the pinch. This would contribute to a score that was numerically less than the participant's true skinfold thickness. Some sources of error will have a numerically positive effect on the participant's score. For example, if a timer was too late stopping a stopwatch, this would contribute

to a recorded time for a 60-meter dash that would be numerically greater than the participant's true time. Note that in this context, the terms *negative* and *positive* do not equate to "bad" or "good" error, but the numeric effect of the source of error. All measurement error is bad because it leads to inaccurate scores. Overall, the various sources of error (some will be negative, some positive) on any occasion will have an overall numerically negative or positive effect on the participant's score (or a zero effect if they all numerically cancel each other out). In the previous example, the fact that the skinfold tester pinched the skinfold too hard might be partly counteracted by a caliper that has a worn, weakened spring (thus contributing a positive source of error to the measured skinfold thickness).

This may all seem rather abstract for a chapter about the practical aspects of testing, but the concept is introduced here in order to help you consider how to minimize measurement error (*e*) through appropriate testing practices. Mahar and Rowe (2008) provided a set of practical guidelines for minimizing measurement error when administering youth fitness tests, but their advice is also relevant to other forms of testing within kinesiology and is a recommended supporting resource for this chapter. On any testing occasion, we will not know what the actual effect of measurement error is, but by recognizing the various sources of error we can work to minimize them as much as possible. The four main sources of error when testing are (1) the tester (the person who is administering the test, collecting data, conducting the measurement, and/or judging the performance); (2) the test-taker (the person being tested); (3) the test (including all of the aspects of the test procedures described earlier in this chapter); and (4) the environment in which the testing takes place (including the built/physical environment, the atmospheric environment, and the social context on the day of testing). These sources of error are listed in **Table 2.2**, along with specific examples of each. Some sources of measurement error are quite literally due to an error (mistake) on the part of the participant or the test administrator. Others are not due to human error but other factors. The importance of the four sources of error varies between testing situations, and you should consider which are most influential in your situation. For example, a polished gym floor (environment) will influence a basketball dribble test much more than a push-up test. The experience of a tester (judge) will be much more influential for scoring a handstand test than for scoring height or weight (because a greater level of skill is required to judge handstand performance). Participant (test-taker) fatigue will have a greater effect in a cognitive test than for a test of flexibility. The quality of the instructions will have a greater effect in tests that require a complex response, such as a soccer dribbling test, than one in which the task is quite straightforward, such as wearing a pedometer to measure physical activity.

By considering these factors and the likelihood they will play a role in a given testing situation, you can ensure that the test-taker's score does justification to the person's true status on whatever you are trying to measure. Some useful strategies for reducing the many types of measurement error are given below.

PREPARE THE PARTICIPANTS FOR TESTING

The participants' readiness to be tested is particularly important for performance tests. Providing a suitable warm-up that is similar to the test task will ensure that participants are physiologically or cognitively prepared to perform to their true ability. Sometimes, arranging tests in an appropriate order can assist in preparing participants for later tests. Testing sit-ups and push-ups before a sit-and-reach or the Scratch Test of flexibility will ensure that the trunk and shoulder girdle are adequately warmed up to enable stretching. The warm-up should not be so extensive as to tire the participant, because

TABLE 2.2 The Main Sources of Measurement Error and Specific Examples

Source of Error	Examples
Test (instrument)	Ambiguous instructions or questions Faulty or poor-quality equipment Insufficient number of trials or items Inappropriate for the given population
Tester (judge, test administrator)	Visual and auditory awareness Lack of understanding or experience Skill Motivation Fatigue Inattentiveness Poor description or presentation of test requirements Poor organization
Test-taker	Innate inconsistency Motivation Fatigue Anxiety Illness Misunderstanding the task Preparation
Testing environment	Noise Temperature Humidity Precipitation Surface Room size Cultural or social context (e.g., peer pressure, clinical environment) Time pressures

participant fatigue is a source of measurement error. If the task involves a cognitive performance, such as a reaction time task or even responding to a questionnaire, giving example tasks first will help to prepare the participant mentally.

The tester is also responsible for ensuring that the participant is appropriately motivated. By explaining the importance or relevance of the test, the tester can ensure that the participant is sufficiently motivated to perform well or, if the test is not a performance test, sufficiently motivated to take the task seriously. Many testing situations are potentially anxiety producing, so the tester should attempt to put the test-taker at ease. There is a delicate balance between motivation and anxiety, and experienced testers will recognize what is needed to motivate or calm a participant. In some situations, it may be helpful to emphasize the importance of the test (in order to increase motivation), whereas at others it may be necessary to play

down the importance and assure the participant that trying is important, rather than focusing on the score (in order to reduce anxiety).

In some situations, ensuring that the participant is adequately prepared may entail communicating important information about what clothes to wear, how much to eat and when, whether to drink fluids prior to participation, and a reminder to get sufficient sleep the night before testing.

EXPLAIN THE TEST TASKS CLEARLY

Some tasks are inherently understandable and straightforward, whereas others are more complex and conceptually abstract. Ideally, the test-taker will have been given the opportunity to practice the test prior to the testing day. This is particularly important when testing participants whose performances are innately less consistent (e.g., children). In some situations, this is not possible. Educational settings offer this opportunity, and test tasks can form part of the normal practice within lessons. By practicing the mile run within a previous lesson, students will realize that the goal is not to start as fast as possible, but that pacing is required. Tests such as the mile run should never be administered to participants who are unfamiliar with the test, because this can lead to embarrassment, a poor performance, or even illness. On the day of the test, first explain the test instructions and procedures, using a standard script if possible to ensure that you do not forget to include an important part of the instructions. Place appropriate emphasis on key aspects of the test scoring criteria. For example, in a throwing test for distance, explain clearly whether the throw is measured from where the ball first hits the ground or from where it ends up. Provide a demonstration (for performance tests) or an example of the response format (on questionnaires or written tests). In educational settings it can be helpful to ask a student to demonstrate (however, ensure beforehand that the student knows how to perform the task correctly and is willing to demonstrate!). During and after the demonstration and explanation, ask questions of the participants to determine whether they understand correctly. Also, ask the participants whether they have any questions. When the skill or procedures are particularly complicated, let the participants run through a practice trial of the test, if this is feasible. During the practice trial, give feedback on whether participants are performing the task correctly, or provide corrective advice if they appear to misunderstand (although in group settings, avoid identifying specific individuals who are making mistakes).

ENSURE THAT THE TESTERS ARE TRAINED

Some testing protocols require training and practice. Testing is itself a skill. Even basic measures such as height and weight can be inaccurate if performed incorrectly. Usually, participants should stand with feet together when height is measured, and should be asked to take in and hold a deep breath while the measure is being taken. For more complicated testing skills, extensive training may be required. Instruction manuals and training videos can help with training. For novice testers, it may be helpful to check their accuracy against an experienced tester. Even for experienced testers it is advisable to conduct occasional quality control to ensure that their skills have not drifted. Observational skills can be validated using videos of performance or behavior. For example, Senne et al. (2009) used videos of physical education lessons to train and check the observational accuracy for using the System for Observing Fitness Instruction Time (SOFIT). If more than one person will conduct testing and the test requires skill using the equipment or expert subjective judgment of a performance, it may be useful to check interrater objectivity before testing.

USE A RELIABLE INSTRUMENT OR TEST PROTOCOL

As referred to previously, test selection is the first stage in testing. It is vital to ensure not only that there is supporting evidence for the reliability and

validity for the test selected, but also that this evidence was gathered on a similar population and in similar settings. A test that has been shown to produce reliable scores in high school children may not be reliable in middle school children. Data supporting the reliability of skinfold measures taken with a Lange skinfold caliper may not apply to the use of a cheaper, plastic skinfold caliper. Many existing tests have a standard protocol that can be followed in all test situations. Where aspects of the test can be modified (such as the number of trials), we recommend collecting pilot data on the population you intend to test in order to determine the optimal test protocol for your population. Depending on the type of test and the time available, often the most practical method of improving test reliability is to increase the number of trials or the length of the test (e.g., the number of items on a questionnaire). The collection of pilot data may also serve as an opportunity for the tester and test-taker to practice the test in a less high-stakes situation. If equipment is to be used, check the technical manuals for which makes and models are trustworthy, and calibrate equipment on a regular basis. Check the accuracy of pedometers regularly by conducting a simple 100-step test, for example, to ensure that the internal mechanism is still working accurately. Use calibration blocks of known thickness to check that skinfold calipers are still accurate.

ENSURE AN OPTIMAL TESTING ENVIRONMENT

In many test settings, the tester may have limited control over the testing environment. If construction noise outside of the school gym is causing a distraction to testing or it is raining on the day of a mile run test, the tester cannot do much about this. The tester has considerable influence on the social environment of a test situation and can reduce the effect of peer pressure or embarrassment in various ways. If space is limited, large-group testing can be adapted so that half of the group is tested on one day while the other half is engaged in a different activity.

Data Processing

After a test has been given, the scores should be analyzed using the appropriate techniques. This usually requires entering the data into a computer so that analysis, record keeping, and data retrieval are possible. Analysis serves to reveal characteristics that could influence the teaching procedures or program conduct and to provide information for the group tested and prepare the data for grading or other evaluation purposes. People are usually interested in their scores, their relative standing in the group or class, and how much they have changed since the last test. Reporting test results to participants is an effective motivational device and an ethical responsibility of the tester.

RECORDING TEST RESULTS

The recording of test results is often no more challenging than placing the scoring sheets and your analysis of them in an appropriate data file. The information makes possible comparisons between classes or groups within and between years, program evaluation over years, and the development of norms. Ethical responsibilities associated with recording and storing data are presented later in this chapter.

Decision Making and Feedback

After administering a test and obtaining a score (measurement), evaluation often follows, taking the form of a judgment about the quality or meaning of a performance. Suppose, for example, that each participant in a class or exercise program ran a mile, and their scores were recorded by the teacher or exercise specialist. When the tester classifies these scores as "healthy" or "unhealthy" or "A," "B," "C," she is making an evaluation.

Evaluation can sometimes be subjective: the judge uses no posttesting standards for each classification and/or evaluates during the performance without recording any measurements. The objectivity of evaluation increases when it is based on defined standards. The standards, or criteria, for evaluation may be inherent in the scoring

system described in the test manual. The tester should ensure that test scores are used for some purpose, and that feedback is provided to the test-taker in situations where this is an explicit reason behind the testing process. Feedback should also be provided promptly. This may even aid motivation for subsequent testing. Particularly in educational settings, having to wait for test results can be tedious and demotivating. In subsequent testing, the student will remember having to wait for results last time, and so encouragement by the tester to try hard may not be taken seriously. Formative evaluation is interpretation or decision making based on scores obtained early or midway through a program, whereas summative evaluation is interpretation or decision making based on scores obtained at the end of a program. Sometimes there is an overlap between formative and summative evaluation. For example, a mid-semester evaluation may provide feedback enabling a student to make changes and improve (formative evaluation), and the test performance will also contribute to an end-of-semester grade (summative performance). It is important to make a distinction between feedback that results from a formal measurement process and feedback in less formal situations. In a teaching or coaching context, a skills test may be administered and feedback provided on what areas require improvement. This feedback is formative evaluation. More informal feedback, for example encouragement or suggestions given out in the middle of a physical education lesson or coaching session, is not formative evaluation, because no formal measurement preceded or informed the feedback.

Important Test Characteristics

When searching for a test, there are certain characteristics you should look for. We introduce these characteristics here to start you thinking about the practical implications of searching for a test that is appropriate for your needs. As you search for and begin using tests, you will realize that these desirable characteristics often overlap or are related. For example, a test that is not reliable cannot yield scores that are valid. A test that is more practical may be less reliable, and so on. The goal is to find a test that is sufficiently high in the necessary characteristics and has the balance of those characteristics that suits your needs. Sometimes practicality may be the most important characteristic, whereas at other times obtaining the most accurate, valid scores may be the most important priority.

Reliability

As described earlier in this chapter, we want scores that are as free from measurement error as possible. *Reliability* is the extent to which scores are error-free. This abstract concept is more readily understood as the consistency of test scores across a range of situations. If, in the formula presented earlier ($X = t + e$), measurement error (e) is small, when participants are measured repeatedly their scores will be similar every time. For reliable tests, this will be unaffected by which instrument you use, which trial is used, the occasion on which the person is tested, or whoever is administering the test. The latter (inter- and intrarater reliability) is also called *objectivity*. Information documenting the reliability evidence for a test should be available either in test manuals or in published articles in journals such as *Measurement in Physical Education and Exercise Science* and *Research Quarterly for Exercise and Sport*.

Validity

Although reliability is an important test characteristic, it only speaks to the accuracy or consistency of test scores as a measure of "something." It does not tell us what that "something" (the underlying construct) is. Knowing about test consistency therefore does not tell us anything about the meaning of the scores. *Validity* is related to the meaning of the test scores. For example, we can

determine that two pedometers, worn simultaneously, provide very similar step counts at the end of each day. This would mean that pedometers have a high degree of interinstrument reliability or agreement. It does not tell us how to interpret the scores or whether pedometers capture physical activity performed during household activities, such as vacuuming, doing laundry, and making beds. Earlier in the chapter, we suggested that when selecting a test you should ask yourself what you want to measure, in whom, and for what purpose. Validity information helps you to determine what a test measures, for whom it is appropriate, and for what purposes. For example, controlled studies have shown that many mechanical pedometers are insensitive to (do not measure) many steps taken during household activities. They are therefore more appropriate for measuring ambulatory activity (walking and running at moderate speeds and faster) rather than incidental physical activities (colloquially described as "pottering about" activities). As described elsewhere, validity is a characteristic of a particular use of test scores rather than a property of the test itself. However, the trustworthiness of the test itself plays a role in validity.

Discrimination

Variability in test scores is a necessary test characteristic so that we can discriminate among individuals throughout the total range of ability (or whatever construct we are measuring). *Discrimination* is the ability to differentiate between people who are truly different. Ideally, there should be many possible scores, and the distribution of the scores in a group of test-takers should be relatively normally distributed. Ideally, no one (or very few people) should receive a perfect score (or the maximum score possible), and no one (or very few people) should receive a zero (or the minimum score possible). This assures us that there is sufficient variability to distinguish between people who are at the high end or low end on the construct we are measuring. Consider the problem of two individuals who both receive the minimum possible or maximum possible score. Although two individuals who receive a zero on a pull-up test both have a low level of upper body strength per pound of body weight, they are probably not equal in strength (even though they both obtained the same score). If many people get the minimum possible score (called a *floor effect*) or many people get the maximum possible score (called a *ceiling effect*), this is particularly problematic, because it means you are unable to use the test to differentiate between many people who truly are different.

Often a construct has several components that are independent (different in nature). Typical examples are health-related fitness, sport skills, and many psychological constructs. In these situations, it is important that a test discriminates between the different components, so you will measure it using a battery composed of several tests. The tests in a battery should be unrelated—that is, the correlation between the tests should be low—both to save testing time and to be fair to the individuals being tested. When two tests are highly correlated, they are probably measuring the same ability. This means that it is redundant to administer both tests because they tell us about the same construct. For example, skinfold measures can be taken at many different sites on the body (triceps, subscapula, thigh, calf, and so on). When selecting a test protocol for body composition, we would decide how many different sites we should measure, because scores from the different sites are often correlated. In most settings, it is sufficient to measure skinfolds from only three sites, because this gives us sufficient information—measuring at seven sites does not give us enough additional accuracy to justify the extra time spent measuring the additional sites.

Practicality and Mass Testability

When we want to test a large number of people in a short period of time, *mass testability* is a critical test characteristic. The longer it takes

to administer each test, the fewer the number of tests that are likely to be administered. With large groups it is essential that people are measured quickly, either successively or simultaneously. A test can be mass testable when a participant performs every few seconds. A sit-up test can be mass testable when half the group is tested while the other half helps with the administration. Remember also that short tests or tests that keep most of the participants active at once help to improve the test-taker's experience of being tested. In educational settings, it can prevent discipline problems that may result from student inactivity and prevent dissatisfaction of participants in research or fitness programs. *Practicality* refers to the feasibility of conducting a test across a variety of applied settings without the need for extensive resources (such as time, personnel, and expensive equipment). In professional settings, we can become so concerned about mass testability and practicality that reliability of the data and, thus, validity of the interpretations based on the data suffer. With careful thought and planning this can be avoided.

Documentation

As referred to previously, test documentation helps the tester to determine whether a test is reliable and will yield valid scores for the intended purpose. Detailed documentation of recommended test procedures also helps to ensure reproducibility of test scores both within a test context (a single test session, for example) and across test contexts (where different people administer the test in different areas of the country; for example, across all schools within a state school system). Documentation of standardized procedures also adds confidence to comparisons across research studies and allows for the comparison of scores against national norms (because we can be confident that the scores used to develop norms were derived from the same testing process as we used to collect our data). A good example of documentation is the Senior Fitness Test. This functional fitness test battery was designed for older adults by Roberta Rikli and Jessie Jones over many years. Via a series of published studies, the reliability and validity evidence for the subtests was documented, and the test manual includes norms developed on data from several hundred participants. The test manual (Rikli & Jones, 2013) also contains a clear explanation of the test protocol, the equipment needed, and other important information to standardize administration of this test.

Ethical Responsibilities of the Test User

Test use carries ethical responsibilities, primarily toward the person being tested. Test scores are often used for high-stakes decision making. In an educational setting, the physical education teacher has to decide what tests to administer and how to use scores for the purposes of grading, for example. In a clinical setting, an exercise scientist may use tests as the basis for determining health risk or the need for an exercise intervention. On the sports field, the athletic trainer has to evaluate an injury quickly and accurately in order to make decisions about acute treatment or to determine whether the athlete can continue playing. We generally use tests because we want there to be some positive outcome; however, there can be adverse side effects of testing. Unintended adverse effects of testing include embarrassment, demotivation, and injury. The American Psychological Association recognized the ethical consequences of testing and the responsibilities of test-users in the 1999 version of its guide to standards for educational and psychological testing (American Psychological Association, 1999). These ethical guidelines apply to professional settings as well as to research data gathering. Some would say that consideration of ethical consequences is the most important responsibility of the test user. Evidence supports this viewpoint. An example of the potential for

unintended negative outcomes is fitness testing in schools. Particularly for students who do not score highly, fitness testing can be a negative experience if not used responsibly. Wiersma and Sherman (2008) wrote an excellent set of recommendations for making fitness testing a positive experience for school children, and is a recommended read for anyone interested in fitness testing. Many of the recommendations set out by Wiersma and Sherman apply equally to noneducational settings and tests other than fitness tests. In the final section of this chapter, we discuss briefly some of the more important ethical aspects of responsible test use.

Fairness

Fairness in testing means that every participant should be given equitable treatment and an equal opportunity to succeed. In most countries, legal and ethical standards govern professional activities so that they do not discriminate against individuals based on, for example, sex, ethnic origin, or disability. We should ensure that the tests we select take these factors into account. For example, standards for youth fitness testing should take into account the participant's age and sex. Separate from the legal implications, in any testing situation we should provide the participants with an equal opportunity to obtain their best score. This is particularly important for performance-related tests. If one physical education class is larger than another, reducing the warm-up time for the larger class in order to allow more time to get through testing may disadvantage the students in the larger class. It is also important to be able to demonstrate lack of bias in subjective scoring. This can be achieved by maintaining a standardized scoring rubric denoting objective criteria for different levels of achievement and establishing interrater objectivity.

Privacy and Confidentiality

Some of the test scores that we collect in kinesiology settings may be sensitive for the person being tested. Many students and participants in fitness or rehabilitation programs would prefer that others did not know how well or poorly they scored on a test, for example. Often students are embarrassed when they receive a test score that is considerably better or worse than those of their peers. Participants in fitness and rehabilitation programs have similar feelings or just do not think that their score should be known by others. All people conducting measurement programs should be sensitive to this issue. Testing people one at a time rather than in a group may be the only way to satisfy this concern. In educational settings, this is usually not feasible because of resource limitations. In such cases, the tester can help to alleviate these concerns by promoting a healthy social environment surrounding the test process and setting ground rules regarding personal information and conduct.

In some circumstances, we should not collect participant-identifying information at all. Consider student class evaluations—it would be inappropriate to require students to include their names on the evaluative comments they make about their class teachers. In many research settings, allowing complete anonymity (no participant ID records at all) will provide more valid data, because participants may respond more honestly if they know their identity is not recorded.

Data Ownership and Data Protection

Beyond the testing situation itself, data security is also important. Most institutional settings will have data protection regulations, and these are also governed by national and state legislation. The test user should take reasonable precautions to protect the data of the people being tested. Score sheets should be kept in secure (locked) cabinets rather than lying around the office. If hard copies of score sheets are to be discarded after data entry, they should be shredded first. Sensitive electronic data should be stored on password-protected

computers, and sensitive data should not be carried around on an unsecure memory stick that could be lost or stolen. For particularly sensitive data, it is advisable not to include participant-identifying information (e.g., name, date of birth) on the data record. Instead, use a system of participant ID numbers, and keep a separate record of which ID number is linked to which participant. Such practices add to the burden of the test administrator or kinesiology professional, but are a necessary ethical responsibility.

Recognize that the participants own their data; it is not the sole property of the tester. Depending on the setting, the participants' scores may be subject to legislation regarding freedom of information. Regardless, the participants should be provided with the results of their testing if they wish to have them. Where scores are linked with norm- or criterion-referenced standards, provide each participant with an evaluation of his or her score. Where this is an explicit expectation of the testing situation (e.g., in an educational or clinical setting), provide feedback within a reasonable time frame.

Participant Safety

Clearly, you should not use tests that endanger the people being tested. Even seemingly innocuous test procedures can carry a risk of injury. In kinesiology, we use many physical performance tests that require a moderate to high level of physical exertion. A participant may slip and fall when getting onto and off of a treadmill even at slow speeds, may sprain an ankle while running an agility test, or may collide with another participant in a basketball dribble test. Examine each test's procedures to see whether individuals might overextend themselves or make mistakes that could cause injury. The use of spotters in weight-lifting tests; soft, nonbreakable marking devices for obstacle runs or the marking of testing areas; and nonslip surfaces and large areas for running and throwing events are always necessary. Adequate hydration should be provided for participants in maximal effort tests, especially during hot, humid conditions. Some participants have underlying clinical conditions that put them at increased risk during tests that require physical exertion. The American College of Sports Medicine (ACSM, 2006) offers guidelines for risk-screening participants prior to administering submaximal and maximal exercise tests or for participating in an exercise program. This risk-screening process is itself an example of criterion-referenced standards. The ACSM Position Statement on Exertional Heat Illness during Training and Competition (Armstrong et al., 2007) provides guidelines that could be useful for avoiding heat illnesses during testing in hot or humid conditions. Among other things, ACSM recommends that if the wet bulb globe temperature exceeds 28°C (82°F), consideration should be given to canceling or rescheduling training or competitive events, and similar caution might apply to excessive exertion in testing under such conditions. No matter how rarely it happens, injury is almost unavoidable in physical performance tests, so we should conduct risk assessments before testing in order to recognize and minimize the risk of injury or illness. Appropriate first aid and emergency procedures should be in place in the event of an injury during physical performance testing.

The Participant Experience

Always try to ensure that testing is a positive, enjoyable experience for the test-taker. If several tests are available, choose the option that minimizes discomfort or inconvenience. When individuals enjoy taking a test and understand why they are being tested, they are motivated to do well, and their scores ordinarily represent their true score. To be enjoyable, a test should be interesting and challenging, within reason. People are more likely to enjoy a test when they have a reasonable chance to achieve an acceptable score. Testing comfort is also an aspect of enjoyment.

Although certain aerobic capacity tests and other maximum effort tests can be uncomfortable, avoid any test so painful that few people can do it well.

Summary

When using tests, a specific sequence of procedures should be followed. Whether you develop your own test or select a preconstructed test (which is advisable and easier), you should ensure that the instrument has certain important attributes. These characteristics make the measurement procedure both efficient and meaningful. Understanding the theoretical underpinnings of reliability and validity will enable you to take practical steps to ensure that the scores you obtain will be trustworthy. Although the successful administration of a test depends on many factors, the key to success is good planning in the pretest stage and attention to the details of that planning during and after the testing procedure.

Professionals in all areas of kinesiology will be responsible for testing individuals of varying abilities and backgrounds. They have an ethical responsibility toward test-takers that extends beyond simply collecting data. Every test situation is different and test users should carefully consider the ethical implications for each testing situation before administering the test.

FORMATIVE EVALUATION OF OBJECTIVES

Objective 1 Describe the key elements of the five stages of testing.

1. Five sequential stages of testing are presented in the chapter. What are the key responsibilities of the tester at each stage?
2. Several key elements will ensure that test administration runs smoothly. What are they?

Objective 2 Describe practical ways by which the tester can minimize measurement error during test administration.

1. Minimizing measurement error can be achieved through anticipating sources of error specific to the tests being used. Describe example situations where (numerically) positive and negative measurement error would occur in the following tests.
   a. A timed bent-knee sit-up test
   b. A basketball dribble test
   c. An agility run test
2. The four major sources of measurement error are tester, test-taker, test, and environment. Evaluate the extent to which each of these might contribute to measurement error in each of the following tests.
   a. A skinfold test
   b. A pedometer measure of daily physical activity
   c. A 1-mile run test or some other cardiovascular test

Objective 3 Identify the important characteristics of a test.

1. What is the difference between reliability and validity?
2. In what situations would mass testability and practicality be particularly important?
3. Describe how a test would be able to discriminate between people of differing abilities.

Objective 4 Describe the ethical responsibilities of test users.

1. Why are ethical responsibilities important?
2. What might be the most important ethical considerations when administering the following tests?
   a. A maximal treadmill test
   b. A questionnaire given to college athletes, asking about eating disorders
   c. A measure of body fatness
3. What measures should a test-user take to ensure confidentiality, privacy, and data protection?

ADDITIONAL LEARNING ACTIVITIES

1. From the material in this chapter and other physical education measurement tests, develop a summary of test characteristics and a checklist of pretest planning procedures.
2. Select a test with which you are unfamiliar and administer it to individuals following the pretest, administrative, and posttest procedures outlined in the chapter.
3. Select a test with which you are unfamiliar and evaluate it, using information in test manuals and published articles, from the perspective of the most important test characteristics described in this chapter.

BIBLIOGRAPHY

Ali, A., Williams, C., Hulse, M., Strudwick, A., Reddin, J., Howarth, L., . . . McGregor, S. (2007). Reliability and validity of two tests of soccer skill. *Journal of Sports Sciences*, *25*, 1461–1470.

American College of Sports Medicine. (2006). *ACSM's guidelines for exercise testing and prescription* (7th ed.). Philadelphia: Lippincott Williams & Wilkins.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Armstrong, L. E., Casa, D. J., Millard-Stafford, M., Moran, D. S., Pyne, S. W., & Roberts, W. O. (2007). Exertional heat illness during training and competition. *Medicine & Science in Sports & Exercise*, *39*, 556–572.

Ha, C-D., Kang, H-S., Han, T-K., & Hong, H-R. (2014). Effects of Taebo exercise on cardiorespiratory fitness and metabolic risk factors in elementary school children. *Korean Journal of Obesity*, *23*, 58–63.

Kenyon, G. S. (1968). Six scales for assessing attitude toward physical activity. *Research Quarterly*, *39*, 566–574.

Mahar, M. T., & Rowe, D. A. (2008). Practical guidelines for valid and reliable youth fitness testing. *Measurement in Physical Education and Exercise Science*, *12*, 126–145.

McElhiney, D., Kang, M., Starkey, C., & Ragan, B. (2014). Improving the memory sections of the Standardized Assessment of Concussion using item analysis. *Measurement in Physical Education and Exercise Science*, *18*, 123–134.

Mullan, E., Markland, D., & Ingledew, D. K. (1997). A graded conceptualisation of self-determination in the regulation of exercise behaviour: Development of a measure using confirmatory factor analytic procedures. *Personality and Individual Differences*, *23*, 745–752.

Rikli, R., & Jones, C. J. (2013). *Senior fitness test manual* (2nd ed.). Champaign, IL: Human Kinetics.

Senne, T., Rowe, D. A., Boswell, B., Decker, J., & Douglas, S. (2009). Factors associated with adolescent physical activity during middle school physical education: A one-year case study. *European Physical Education Review*, *15*, 295–314.

Wiersma, L. D., & Sherman, C. P. (2008). The responsible use of youth fitness testing to enhance student motivation, enjoyment, and performance. *Measurement in Physical Education and Exercise Science*, *12*, 167–183.

Example Tests with a Rigorous Set of Supporting Documentation

Plowman, S. A., & Meredith, M. D. (eds.). (2013). *Fitnessgram/Activitygram reference guide* (4th ed.). Dallas, TX: The Cooper Institute.
(Also see the special issue of the *Journal of Physical Activity and Health* on the reliability and validity of FITNESSGRAM® tests [2006, Supplement 2], and supplement to the *American Journal of Preventive Medicine* on evidence for criterion-referenced standards for FITNESSGRAM® [2011, Supplement 2]).

Rikli, R., & Jones, C. J. (2013). *Senior fitness test manual* (2nd ed.) Champaign, IL: Human Kinetics.
(Also see multiple published studies by Rikli and Jones reporting reliability and validity evidence for the tests within the Senior Fitness Test battery, published in journals such as *Research Quarterly for Exercise and Sport*, *Medicine & Science in Sports & Exercise*, the *Journal of Aging and Physical Activity*, *The Gerontologist*, and *Science & Sports*.)

McKenzie, T. L. (2005). *SOFIT procedures manual*. San Diego, CA: San Diego State University.
(SOFIT is the Systematic Observation of Fitness Instruction Time. Also, see similar test procedures manuals by McKenzie for SOPLAY and SOPARC, two other observational tools for measuring physical activity. Multiple validation studies exist for SOFIT, SOPLAY, and SOPARC, published by McKenzie and colleagues.)