Module 1

Introduction and Overview

Thousands of important, and sometimes life-altering, decisions are made every day. Who should we hire? Which students should be placed in accelerated or remedial programs? Which defendants should be incarcerated and which paroled? Which treatment regimen will work best for a given client? Should custody of this child be granted to the mother or the father or the grandparents? In each of these situations, a "test" may be used to help provide guidance. There are many vocal opponents to the use of standardized tests to make such decisions. However, the bottom line is that these critical decisions will ultimately be made with or without the use of test information. The question we have to ask ourselves is, "Can a better decision be made with the use of relevant test information?" In many, although not all, instances, the answer will be yes, *if* a well-developed and appropriate test is used *in combination* with other relevant information available to the decision maker. The opposition that many individuals have to standardized tests is that they are the sole basis for making an important, sometimes life-altering, decision. Thus, it would behoove any decision maker to take full advantage of other relevant information, where available, to make the best and most well-informed decision possible.

What Makes Tests Useful

Tests can take many forms from traditional paper-and-pencil exams to portfolio assessments, job interviews, case histories, behavioral observations, and peer ratings—to name just a few. The common theme in all of these **assessment** procedures is that they represent a sample of behaviors from the test taker. Thus, psychological testing is similar to any science in that a sample

is taken to make inferences about a population. In this case, the sample consists of behaviors (e.g., test responses on a paper-and-pencil test or performance of physical tasks on a physical **ability test**) from a larger domain of all possible behaviors representing a construct. For example, the first test we take when we come into the world is called the APGAR test. That's right, just 1 minute into the world we get our first test. You probably do not remember your score on your APGAR test, but our guess is your mother does, given the importance this first test has in revealing your initial physical functioning. The purpose of the APGAR test is to assess a newborn's general functioning right after birth. Table 1.1 displays the five categories that newborn infants are tested on at 1 and 5 minutes after birth: Appearance, Pulse, Grimace, Activity, and Respiration (hence, the acronym APGAR). A score is obtained by summing the infant's assessed value on each of the dimensions. Scores can range from 0 to 10. A score of 7–10 is considered normal. A score of 4–6 indicates that the newborn infant may require some resuscitation, while a score of 3 or less means the newborn would require immediate and intensive resuscitation. The infant is then assessed again at 5 minutes, and if his or her score still is below a 7, he or she may be assessed again at 10 minutes. If the infant's APGAR score is 7 or above 5 minutes after birth, which is typical, then no further intervention is called for. Hence, by taking a relatively small sampling of behavior, we are (or at least a competent obstetrics nurse or doctor is) able to quickly, and quite accurately, assess the functioning of a newborn infant to determine if resuscitation interventions are required to get the newborn functioning properly.

Table 1.1 The APGAR Test Scoring Table

	Points		
Sign	*0*	*1*	*2*
Appearance (color)	Pale or blue	Body pink, extremities blue	Pink (normal for non-Caucasians)
Pulse (heartbeat)	Not detectible	Lower than 100 bpm	Higher than 100 bpm
Grimace (reflex)	No response	Grimace	Lusty cry
Activity (muscle tone)	Flaccid	Some movement	A lot of activity
Respiration (breathing)	None	Slow, irregular	Good (crying)

The **utility** of any assessment device, however, will depend on the qualities of the test and the intended use of the test (see Web References 1.1 and 1.2 for a discussion of these desired qualities). Test information can be used for a variety of purposes from making predictions about the likelihood that a patient will commit suicide to making personnel selection decisions by determining which entry-level workers to hire. Tests can also be used for classification purposes, as when students are designated as remedial, gifted, or somewhere in between. Tests can also be used for evaluation purposes, as in the use of a classroom test to evaluate performance of students in a given subject matter. Counseling psychologists routinely use tests to assess clients for emotional adjustment problems or possibly for help in providing vocational counseling. Finally, tests can also be used for research-only purposes such as when an experimenter uses a test to prescreen study participants to assign each one to an experimental condition. If the test is not used for its intended purpose, however, it will not be very useful and, in fact, may actually be harmful. As Anastasi and Urbina (1997) note, "Psychological tests are tools . . . Any tool can be an instrument of good or harm, depending on how it is used" (p. 2).

For example, most American children in grades 2–12 are required to take standardized tests on a yearly basis. These tests were initially intended for the sole purpose of assessing students' learning outcomes. Over time, however, a variety of other misuses for these tests have emerged. For instance, they are frequently used to determine school funding and, in some cases, teachers' or school administrators' "merit" pay. However, given that determining the pay levels of educational employees was not the intended use of such standardized educational tests when they were developed, they almost always serve poorly in this capacity. Thus, a test that was developed with good (i.e., appropriate) intentions can be used for inappropriate purposes, limiting the usefulness of the test. In this instance, however, not only is the test of little use in setting pay for teachers and administrators, it may actually be causing harm by coercing teachers to "teach to the test," thereby trading long-term gains in learning for short-term increases in standardized test performance.

In addition, no matter how the test is used, it will only be useful if it meets certain psychometric and practical requirements. From a psychometric or measurement standpoint, we want to know if the test is accurate, standardized, and reliable; if it demonstrates evidence of validity; and if it is free of both measurement and predictive bias. Procedures for determining these psychometric qualities form the core of the rest of this book (see also Web References 1.1–1.7). From a practical standpoint, the test must be cost effective as well as relatively easy to administer and score. Reflecting on our earlier example, we would surmise that the APGAR meets most of these qualities of being practical. Trained doctors and nurses in a hospital

delivery room can administer the APGAR quickly and efficiently. Our key psychometric concern in this situation may be how often different doctors and nurses are able to provide similar APGAR scores in a given situation (i.e., the **inter-rater reliability** of the APGAR).

Individual Differences

Ultimately, when it comes right down to it, those interested in applied psychological measurement are usually interested in some form of **individual differences** (i.e., how individuals differ on test scores and the underlying **traits** being measured by those tests). If there are no differences in how target individuals score on the test, then the test will have little value to us. For example, if we give a group of elite athletes the standard physical ability test given to candidates for a police officer job, there will likely be very little variability in scores with all the athletes scoring extremely high on the test. Thus, the test data would provide little value in predicting which athletes would make good police officers. On the other hand, if we had a more typical group of job candidates who passed previous hurdles in the personnel selection process for police officer (e.g., cognitive tests, background checks, psychological evaluations) and gave them the same physical ability test, we would see much wider variability in scores. Thus, the test would at least have the potential to be a useful predictor of job success, as we would have at least some variability in the observed test scores.

Individual differences on psychological tests can take several different forms. Typically, we look at **inter-individual differences** where we examine differences on the same construct across individuals. In such cases, the desire is usually prediction. That is, how well does the test predict some criterion of interest? For example, in the preceding scenario, we would use the physical ability test data to predict who would be successful in police work. Typically, job candidates are rank ordered based on their test scores and selected in a top-down fashion, assuming the test is indeed linearly associated with job performance. As you will see as we move further into the book, however, it is rare that any single test will be sufficient to provide a complete picture of the test taker. Thus, more often than not, several tests (or at least several decision points) are incorporated into the decision-making process.

We may also be interested in examining **intra-individual differences.** These differences can take two forms. In the first situation, we may be interested in examining a single construct within the same individual across time. In this case, we are interested in how the individual changes or matures over time. For example, there have been longitudinal studies conducted by life-span

developmental psychologists that have looked at how an individual's cognitive ability and personality change over the course of his or her lifetime. In particular, these researchers are interested in studying intra-individual differences in maturation. That is, why do some individuals' scores on cognitive ability tests go up dramatically over time, while the scores of other individuals only go up a little or not at all or maybe even go down? Thus, the focus is not on group mean differences (as in inter-individual differences); rather, we are looking for different rates of change within individuals over time.

In the second form of intra-individual differences, we are interested in looking at a given individual's strengths and weaknesses across a variety of constructs, typically at one point in time. Thus, the same individual is given a variety (or **battery**) of different tests. Here we are usually interested in classifying individuals based on their strengths and weaknesses. For example, hundreds of thousands of high school students take the **Armed Services Vocational Aptitude Battery** (**ASVAB**) every year. The ASVAB consists of a series of 10 subtests that assess individuals' strengths and weaknesses in a wide variety of aptitudes. Those not interested in pursuing a military career can use it for career counseling purposes, while individuals interested in military service can use it to be placed or classified within a particular branch of the armed services or career path within the military based on their relative strengths and weaknesses. The key is that the ASVAB consists of a **test battery** that allows test users to see how individuals differ in terms of the relative strength of different traits and characteristics. Hence, the ASVAB is useful for several different constituents in the testing process.

Constituents in the Testing Process

Because the decisions that result from the uses of test data are so often of great consequence, the testing process is very much a political process. Each of the **constituents** or stakeholders in the testing process will have a vested interest in the outcome, albeit for different reasons. Obviously, the **test takers** themselves have a strong vested interest in the outcome of the testing process. Because they are the ones who will be affected most by the use of the test, they tend to be most concerned with the procedural and distributive (i.e., outcome) fairness of the test and the testing process. The **test users** (those who administer, score, and use the test) may be less concerned with an individual's outcome per se, focusing more on making sure the test and testing process are as fair as possible to all test takers. They are using the test, no doubt, to help make a critical decision for both the individuals and the organization using the test. Thus, they will also be concerned with many of

the psychometric issues that will be discussed throughout this book, such as reliability, validity, and test bias. The **test developer** tends to focus on providing the best possible test to the test user and test taker. This includes making sure the test is well designed and developed, in addition to being practical and effective. Test developers also need to collect and provide evidence that the test demonstrates consistency of scores (i.e., **reliability**) and that the concepts that are purported to be measured are, in fact, measured.

Thus, this book focuses on what you will need to know to be a qualified *test developer* and informed *test user*. You will learn how to develop test questions, determine the psychometric properties of a test, and evaluate test items and the entire test for potential biases. In addition, many practical issues such as test translation, dealing with response biases, and interpreting test scores will also be discussed. Each module includes case studies and hands-on exercises that will provide practice in thinking about and working through the many complicated psychometric processes you will learn about in the rest of the book. In addition, many modules also include step-by-step examples to walk you through the process that an applied practitioner would go through to evaluate the concepts discussed in that particular module. Thus, in short, conscientious use of this book will help you to better understand and apply the knowledge and skills you are developing as you study a wide variety of topics within advanced measurement theory.

Concluding Comments

Psychological testing, when done properly, can be a tremendous benefit to society. Competently developed and implemented assessment devices can provide valuable input to the critical decisions we are faced with everyday. However, poorly developed and implemented tests may, at best, be of little assistance and, in fact, may actually do more harm than good. Therefore, the rest of this book was written to help you become a more informed consumer of psychological tests as well as to prepare future test developers in terms of the critical competencies that are needed to develop tests that will be beneficial to society and acceptable (or at least tolerable) to all testing constituents.

Practical Questions

1. What specific goals do you want to achieve by taking a course in measurement?
2. What will likely be your major stake in the testing process once you finish your measurement course?

3. What alternative "test" to the APGAR could an obstetrics nurse or doctor use to assess newborn functioning? What would be the advantages and disadvantages compared to the APGAR test shown in Table 1.1?

4. Who are the major constituents or stakeholders in the psychological testing process?

5. What is the major purpose of examining inter-individual differences via test scores?

6. What are the different types of intra-individual differences?

7. What are the major purposes of the different forms of intra-individual differences in interpreting test data?

8. Can you provide examples of the uses of both inter-individual differences and intra-individual differences?

Case Studies

CASE STUDY 1.1: TESTING CONSTITUENTS AND THE STANFORD ACHIEVEMENT TEST

Professor Gilbert, an educational testing professor at a local state university, was contacted by a small school district that had decided to implement a Talented and Gifted (TAG) program for advanced students. The school district initially was going to use grade point average (GPA) as the sole basis for placement into the TAG program. However, several parents objected that the different tracks within the schools tended to grade using different standards. As a result, those students in Track A had much higher GPAs (on average) than those in the other two tracks. Thus, those in Track A were much more likely to be placed in the TAG program if only GPA was used than those in Tracks B and C.

Therefore, the school board decided to set up an ad hoc committee to provide recommendations to the board as to how entrance to the new TAG program would be determined. The committee was headed by Professor Gilbert (who also happened to have two sons in the school system) and included school psychologists, principals, parents,

teachers, and students. The committee's initial report recommended that teacher written evaluations, test scores from the Stanford Achievement Test, version 9 (SAT-9), and letters of recommendations be used, in addition to GPA, to determine entrance into the TAG program. As you might have guessed, the next meeting of the school board, where these recommendations were presented and discussed, was a heated affair. Professor Gilbert was suddenly beginning to ponder whether she needed to raise her consulting fees.

Questions to Ponder

1. Who are the major constituents or stakeholders in the testing process in this scenario?
2. What is Professor Gilbert's "stake" in the testing process? Does she have more than one?
3. What form of individual differences is the committee most likely to be focusing on? Why?
4. Should all of the different assessment devices be equally weighted?

CASE STUDY 1.2: DEVELOPMENT OF A VOLUNTEER PLACEMENT TEST

A local volunteer referral agency was interested in using "tests" to place volunteer applicants in the volunteer organizations it served. In order to do so, however, the agency needed to assess each applicant to determine where his or her skills could best be used. As a first step, the director of the agency contacted a local university and found out that Professor Kottke's graduate practicum class in applied testing was in need of a community-based project. Soon thereafter, Professor Kottke and her students met with the director of the agency to determine what her needs were and how the class could help.

In the past, the agency first conducted a short 15-minute telephone interview as an initial screen for each volunteer applicant. Those applicants who appeared to be promising were asked to come in for a half-hour face-to-face interview with a member of the agency staff. If the applicant was successful at this stage, a brief background check

was conducted, and the candidates who passed were placed in the first available opening. However, the agency was receiving feedback from the volunteer organizations that a large portion of the volunteers were participating for only a month or two and would then never return. In follow-up interviews with these volunteers, the most consistent reason given for not returning was that the volunteer placement was simply "not a good fit." Thus, Professor Kottke and her class were asked to improve the fit of candidates to the positions in which they were being placed. Unfortunately, Professor Kottke's 10-week course was already one third completed, so she and her students would have to work quickly.

Questions to Ponder

1. If you were in Professor Kottke's practicum class, where would you start in the process of trying to help this agency?
2. Does this seem to be more of an inter-individual differences or intra-individual differences issue? Explain.
3. Who are the constituents in this testing process?
4. What do you think Professor Kottke and her students can realistically accomplish in the six to seven weeks remaining in the term?

Exercises

EXERCISE 1.1: DIFFERENT USES FOR A GIVEN TEST

OBJECTIVE: To think critically about the wide variety of uses for a given test.

The Armed Services Vocational Aptitude Battery (ASVAB) was discussed in the module overview. Nearly one million people take this test each year, many of them high school students. The test consists of 10 different subtests measuring general science, arithmetic reasoning, word knowledge, paragraph comprehension, numerical operations, coding speed, auto and shop information, mathematics

knowledge, mechanical comprehension, and electronics information. The ASVAB is used primarily to select recruits for the different branches of the armed services and then to place those individuals selected into various training programs based on their aptitude strengths and weaknesses. In fact, a subset of 100 items (called the Armed Forces Qualification Test—AFQT) from the ASVAB is used by all the branches of the military to select recruits. Each branch of the military employs a slightly different cutoff score to select recruits.

Given the 10 subtests listed previously, what other purposes could the ASVAB be used for besides selection and placement (i.e., career guidance)?

EXERCISE 1.2: WHO ARE THE MAJOR CONSTITUENTS IN THE TESTING PROCESS?

OBJECTIVE: To become familiar with the major constituents in the testing process.

As noted in the overview, there are typically numerous constituents in any given testing process. These constituents may include the test takers, test developers, test users, and, more broadly, society in general. Each of these constituents will have a varying degree of interest in a given assessment device.

Who are the major test constituents with regard to the Armed Services Vocational Aptitude Battery (ASVAB) discussed in the module overview and in Exercise 1.1? What would be the major concerns of each of these different constituents with regard to development, refinement, administration, and use of the ASVAB?

EXERCISE 1.3: TESTING AND INDIVIDUAL DIFFERENCES

OBJECTIVE: To identify the major forms of individual differences commonly assessed with psychological tests.

Most tests are administered to identify some form of individual differences. These can include inter-individual differences, intra-individual differences, or both. Again, looking at the Armed Services Vocational

Aptitude Battery (ASVAB), what forms of inter- and intra-individual differences might be assessed with this particular test?

Internet Web Site References

1.1. http://trochim.human.cornell.edu/kb/contents.htm

This Web page provides the table of contents for the online textbook, *The Research Methods Knowledge Base,* by Dr. William M. Trochim of Cornell University. A wealth of testing-related information is available from subsidiary links. Links from within this table of contents that are relevant to specific modules are cited throughout this book. An enhanced and revised version of Trochim's book is available in print. A hard copy of the book can be ordered online at http://trochim.omni.cornell.edu/kb/order.htm.

1.2. http://www.ed.gov/databases/ERIC_Digests/ed385607.html

This Web page, from the Educational Resources Information Center (ERIC), provides key standards to consider when evaluating a test. The standards are based on the AERA/APA/NCME (1999) *Standards for Educational and Psychological Testing.*

1.3. http://www.apa.org/science/fairtestcode.html

This Web page presents the Code of Fair Testing Practices in Education (1988) as developed by the Joint Committee on Testing Practices.

1.4. http://www.intestcom.org/test_use_full.htm

This Web page provides the International Test Commission's Guidelines on test use.

1.5. http://www.vanguard.edu/faculty/ddegelman/amoebaweb/index.cfm?doc_id=853

This Web page, maintained by Dr. Douglas Degelman of Vanguard University, presents links to a wide range of topics related to testing and research methods.

1.6. http://www.cpsimoes.net/artigos/art_psycho_eng.html

This Web page provides a brief introduction to the concept of psychological testing.

1.7. http://www.siop.org/_Principles/principlesdefault.htm

This Web page provides access to the fourth edition of the *Principles for the Validation and Use of Personnel Selection Procedures,* a guidebook of the accepted professional practices in the field of personnel selection psychology.

1.8. http://www.unl.edu/buros/

This Web page provides **Mental Measurement Yearbook** (**MMY**) reviews of nearly 4,000 commercially available tests. A fee is charged for each test review accessed. Your library, however, might subscribe to this service. Check with your instructor.

1.9. http://davidmlane.com/hyperstat/Statistical_analyses.html

This Web page provides links to several free statistical programs that can be used for analyzing test data.

Further Readings

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed., pp. 2–31). Upper Saddle River, NJ: Prentice Hall.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory* (pp. 3–15). Belmont, CA: Wadsworth.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences* (pp. 9–30). New York: W. H. Freeman.

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing: Principles and applications* (5th ed., pp. 49–66). Upper Saddle River, NJ: Prentice Hall.