CHAPTER 2

Assessment Administration, Scoring, and Interpretation

Learning Objectives

- Describe the role and importance of administration protocols.
- Understand standardization and its implications for interpretation.
- Explain how percentile ranks and standard scores are used to interpret assessment results.
- Compare the purposes and interpretive frameworks of norm-referenced and criterion-referenced assessments.
- Identify the components of effective scoring guidelines and rubrics in behavioral assessments.
- Understand the importance of assessment conditions in determining the generalizability of assessment outcomes.
- Recognize how subject-matter expertise and standard setting contribute to fair and valid score interpretation.

In the previous chapter we gave readers a brief introduction to, or more likely a refresher on, the foundational dimensions of ABA. Then, we began our introduction to assessment and measurement. Although not unique to ABA, assessment is a critical function for any type of systematic data collection to support decision making. Behavioral assessment has a long tradition in ABA,

but we contend that there are exciting opportunities to apply assessment and measurement principles from other disciplines to bolster research and evaluation efforts in ABA. We concluded the previous chapter by stating that this book is designed to be a guide through assessment-related considerations and decision points, so that's where this chapter begins: conditions and considerations for **assessment administration**, scoring, and interpretation.

The use of behavioral assessments has increased sharply over the past two decades. The increased usage is driven in part by the development of new instruments to aid in determining skill repertoire, requirements by insurance companies to administer standardized assessments, and the recognized need for standardization in the assessment process. Research regarding the measurement properties and quality of standardized behavioral assessments is ongoing and will be discussed later in this book, but there are important aspects of administration, scoring, and interpretation with immediate relevance. Standardized assessments are being administered to individuals daily. In this chapter, we will discuss considerations for administration, scoring, and interpretation. With respect to administration, we provide a thematic representation of the *Standards for Educational and Psychological Testing* (AERA et al., 2014), which is a must-read text for those conducting research in the social sciences. With respect to scoring, we present various types of scores or reporting structures as well as the scoring considerations for norm- and criterion-referenced assessments. Different types of scores have different interpretive considerations that we discuss as well. The content of this chapter is important in its own right, and the topics discussed here also have important implications for reliability and validity evidence gathering, which are the foci, respectively, of Chapters 3 and 4.

ADMINISTRATION

The manner and conditions in which an assessment is administered could have bearing on the way individuals respond to items/prompts, which is why it is so important that the procedures and conditions be standardized to the extent possible. When individuals' behaviors are reflective of specific administration procedures or **assessment conditions** rather than the item or prompt from the assessment, the decisions made based on the assessment results could be flawed. Fortunately, it is within the control of researchers in many cases to closely follow the administration procedures to minimize the potential impact on behavior and/or results. In this section, we will discuss standardization and assessment conditions, which can support valid decisions made based on the assessment results.

Standardization

When thinking of "standardized tests" many readers will recall taking tests as a part of public education or for admission to schools or universities. As a

result of this association, it may be common to think of standardized assessments as being limited to cognitive or psychological assessments. It is true that many cognitive assessments are standardized, but there are many psychological and behavioral assessments that are standardized as well. **Standardization** refers to the conditions under which an assessment is administered and scored—that is, a standardized assessment should involve administering the same items or materials across all individuals and occasions in a manner that is consistent with the procedures stipulated by the assessment developer (American Educational Research Association et al., 2014). All aspects of a standardized protocol, such as materials, time limits, instructions, and scoring rules, must be consistent to allow us to conclude that any differences in observed behaviors are due to differences between individuals, time, or occasions rather than differences in assessment procedures (Barrios & Hartmann, 1986). Furthermore, if a standardized protocol is followed there should not be any differences in outcome between evaluators. The assessment should also be scored using a predefined set of scoring rules such that scoring can be performed consistently (American Educational Research Association et al., 2014).

Standardized assessments typically include, at minimum, a manual or set of administration guidelines for the assessment. Some assessments, like PEAK Comprehensive Assessment (PCA), even offer commercial trainings for those interested in administering and scoring the assessment. The assessment manual will specify instructions, possible time limits, form of item presentation and response, and which materials can or should be used during the administration (American Educational Research Association et al., 2014). The manual is an essential tool to ensure that an assessment is administered as intended, which contributes to consistent administrations. For example, a standardized behavioral assessment would include a script to read when presenting an item to an individual and would ideally have a robust set of possible responses corresponding to specific scores (i.e., scoring rules) to promote consistency across administrations. The more prescriptive an assessment protocol, the more reflective the assessment results are of differences between individuals, time, or occasions—rather than being reflective of the assessment administration procedures. In other words, the administration and environment are more controlled.

The conceptual purpose of using standardization in behavioral assessment is common in behavior research. In the ABAB design (in which the first A refers to the baseline phase; the first B refers to the intervention phase; the second A refers to the return to baseline; and the second B refers to reintroduction of the intervention), for example, the behavior researchers demonstrate experimental control over the independent variable while accounting for confounding variables. In doing so, any differences that are observed between phases can be attributed to the effect of the independent variable because it is the only element of the design that varies. Treatment fidelity is evaluated in many behavior research studies for the same reason—to ensure that the observed behaviors are not due to inappropriate or inconsistent administration

of the treatment. Failure to administer a treatment with fidelity is a failure to control confounding variables, which threatens the internal validity of the study. Similarly, if the same assessment was given to two individuals following a standardized protocol, then a researcher could reasonably conclude that any differences in observed behavior were due to differences between the individuals rather than the way the assessment was conducted.

With respect to assessments commonly used in ABA, the VB-MAPP offers suggestions for administration conditions whereas PEAK includes a diagram on administration. For the PEAK, ABLLS-R, and AFLS, the first round of assessment can be done using interviewers or informants rating the items, so the administration may lack standardization. Needless to say, assessments that do not require consistent administration procedures are unstandardized, but adaptations may be necessary at times. From a measurement perspective, the kinds and nature of adaptations have important implications for the kinds of decisions that can be supported based on the data.

Accommodations and Modifications

In traditional assessment of cognitive or psychological phenomena, certain changes to the assessment or protocol may be necessary to increase the access of some individuals, such as those with diverse linguistic or cultural backgrounds or those with disabilities (American Educational Research Association et al., 2014). The changes are commonly classified as accommodations or modifications, depending on the nature and type of the adaptation and the purpose of the assessment. **Accommodations** are those changes that increase access to the test for individuals and preserve the comparability of scores for individuals who took the assessment with and without accommodations. Examples of accommodations include allowing an individual additional time to complete an assessment, allowing an individual with a physical disability to use an assistive device to complete an assessment, or allowing an English Language Learner to use a translation dictionary to complete an assessment of learning in a non-language-related area. **Modifications** are those changes that increase access but do not preserve the comparability of scores for individuals who took the assessment with and without modifications. That is, the interpretation of the scores is not the same for modified and unmodified assessments. Examples of modifications include having a reading assessment read aloud (i.e., the assessment no longer reflects reading ability), allowing use of a calculator on an arithmetic assessment (i.e., the assessment no longer reflects arithmetic ability), or allowing an English Language Learner to use a translation dictionary on an assessment of English vocabulary.

It should be noted that accommodations or modifications may not, or more likely will not be necessary due to the nature and purpose of behavioral assessment. Many behavioral assessments are specifically designed for individuals with intellectual or developmental disabilities where comparisons of scores across individuals are rarely of primary interest. Instead, many

behavioral assessments are designed to provide information about an individual's current level of a particular skill by presenting stimuli (e.g., items or prompts) that evoke the target behavior that can be observed. We include this brief presentation of accommodations and modifications for the completeness of our measurement overview for readers.

Assessment Conditions

Assessment conditions refer to the place, time, and/or setting when an assessment is administered. Ideally, the conditions would also be standardized (i.e., the same for all individuals), but this of course is rarely feasible or, in many cases, possible. Assessment developers do not expect all administrations to be identical and may offer guidance for necessary adjustments. The assessment conditions have bearing on the kinds of generalizations that are supported. For example, behaviors observed in a clinical setting are likely to be different than behaviors exhibited at home. Decisions made about behaviors based on clinical observations may not apply to an individual at home. If an assessment is designed to support decisions about individuals in schools, then administering the assessment in a school setting would best align the assessment with desired inferences or decisions.

NORM-REFERENCED ASSESSMENTS

For some assessments, the goal is to determine how much or little of a skill, attribute, or trait a person has relative to others. These assessments are norm-referenced. An individual's score(s) on an assessment is compared to the scores of others in order to determine if the individual is above average, below average, or somewhere in between. In such scenarios, it is critical to know the characteristics of the comparison group before reaching such conclusions. The group against which individual scores are compared is a so-called normative sample. The assessment must have been administered to a normative sample in order to know which score(s) are considered "average" and how much variability can be expected in the scores. For most norm-referenced assessments, the normative sample is made up of a representative sample of children, adolescents, or adults, depending on the target population. This can complicate interpretation of scores for individuals with disabilities because, in many cases, norm-referenced assessments are not designed to support decisions about or normed with a representative sample of the general population of individuals with disabilities.

Scoring Rules

As one might anticipate, **scoring rules** are guidelines by which observations or events are coded and/or combined. Assessments are commonly administered

to determine *how much* of a particular kind of skill a person possesses, although assessment may also be administered to determine *what kind* of skill a person possesses. In either case, the behaviors or responses elicited by an assessment are coded in some way in accordance with scoring rules. After all prompts have been administered and behaviors or responses have been scored, the scoring rules provide guidance on how to combine the set of scores into a type of summary, or composite, score. It should be noted that the term "composite score" is ubiquitous in cognitive and psychological assessment when referring to this type of summary score. A **composite score** has the benefit of being one score representing all the individual prompt or item scores on an assessment, and can be very useful for expressing how much or little of a skill a person is able to demonstrate. That is, higher scores typically reflect a higher level of some skill, function, development, or attribute whereas lower scores reflect lower levels of these characteristics. A composite score is commonly the sum or average of a set of item scores with or without weighting. The scoring rules outline the steps necessary to compute the composite score. In the next section, we will discuss how the interpretation of these scores is dependent on the types of conclusions that are desired about each individual.

Reporting and Interpretation

When scores are interpreted relative to the scores of other people, the scores must be transformed in order to incorporate information from others. The transformed scores can then be interpreted because they contain the necessary comparative information. That is, they reflect the position of the observed score relative to the scores of others. The two most common measures of position are percentile ranks and standard scores.

Percentile Ranks

A common way of reporting the relative position of a score compared to the scores of others is by reporting the percentile rank of the score. It should be noted that percentile ranks and percentiles are related but different concepts. Depending on which text or online resource one reads, the definitions for these two terms are slightly different, but the essential difference between the terms is that a percentile is a score and a percentile rank is a percentage. A **percentile rank** of a score is the percentage of scores that are less than or equal to the score. A percentile is the score in a distribution at or below which some percentage of scores fall. Below we demonstrate how to manually calculate the percentile rank of a score, but we acknowledge that most users will not be calculating percentile ranks manually.

In order to determine the percentile rank of a score, all scores must first be ordered from largest to smallest, and the frequency of each score recorded. Consider the set of scores in Table 2.1. The highest score recorded was 40 and the lowest score recorded was 26. The "Frequency" column indicates how

many times each score was observed. For example, only one individual had a score of 40, 10 individuals had a score of 37, and 17 individuals had a score of 29. There are 167 total individuals reflected in this distribution. The "Cumulative frequency" column shows how many scores are at or below a given value. For example, there are 45 scores at or below 29, there are 101 scores at or below 32, and all 167 scores are at or below 40.

The formula for percentile rank for some score, x, is

$$PR_x = \frac{cf_b + .5(f_x)}{N} \times 100 \quad [2.1]$$

where cf_b is the cumulative frequency of scores below score x, f_x is the frequency of score x, and N is the total number of scores. Using the scores in Table 2.1 to compute the percentile rank for a score of, say, 30, the calculation would be

$$PR_{30} = \frac{cf_{29} + .5(f_{30})}{N} \times 100 = \frac{45 + .5(18)}{167} \times 100 = 32.3$$

TABLE 2.1. Frequency Distribution of Scores

Score	Frequency	Cumulative frequency
40	1	167
39	1	166
38	5	165
37	10	160
36	13	150
35	16	137
34	10	121
33	10	111
32	25	101
31	13	76
30	18	63
29	17	45
28	15	28
27	9	13
26	4	4

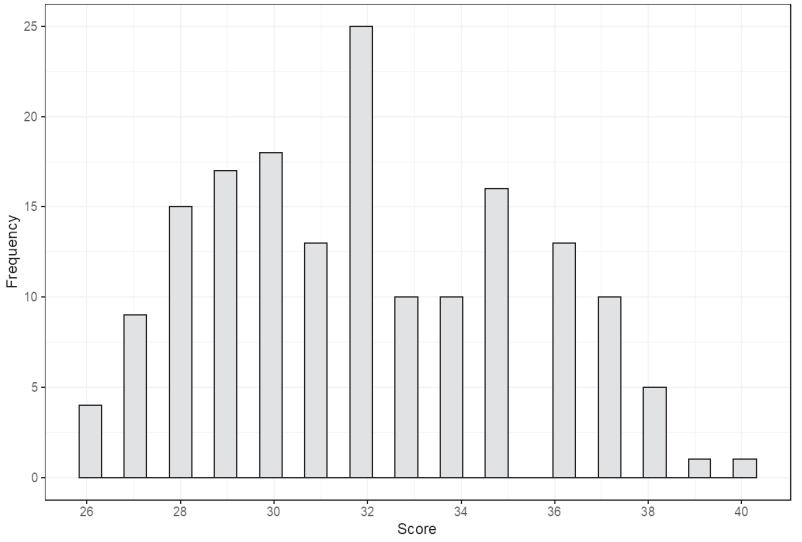

FIGURE 2.1. Frequency distribution of scores from Table 2.1.

The interpretation would then be that 32.3% of the scores are at or below a score of 30. Because a relatively small portion of the distribution is at or below the score of 30, the percentile rank of 32.3 shows that the score of 30 is lower than most scores in the distribution. The distribution of the scores is plotted in Figure 2.1.

Standard Scores

A second common way of expressing the position of a score relative to the scores of others is to convert the score to a **standard score**, also sometimes referred to as a *z* score. Not only does a standard score express how far away a score is from the average, but standard scores are on a scale that allows comparisons across assessments. In other words, standard scores allow the relative position of a score on one instrument to be compared with the relative position of a score on a different instrument.

To begin, one must determine how far away a score is from the average, or mean. This is done by subtracting the mean from the score to produce what is called a deviation score. That is,

Observed score – Mean score = Deviation score

Positive deviation scores indicate that the observed score is larger than the mean. Negative deviation scores indicate that the observed score is smaller than the mean score. Deviation scores of zero indicate that the observed score

and the mean score are equal. In order to convert deviation scores to a standardized scale, the deviation scores must be divided by the standard deviation of the set of scores. The formula for the standard deviation is beyond the scope of this text but can be found in any introductory statistics text or by doing an internet search for "standard deviation." For most users, a computer program will be used to obtain the standard deviation. Regardless of how one chooses to obtain the standard deviation of a set of scores, the standard deviation expresses how much scores tend to differ from the mean, on average. The formula for a standard score is shown in Equation 2.2.

$$z = \frac{\text{Observed score} - \text{Mean score}}{\text{Standard deviation}} = \frac{x - \bar{x}}{s} \quad [2.2]$$

For the scores shown in Table 2.1, the mean score is 32 and the standard deviation is 3.3. Using the same score as before, 30, the standard score is

$$z = \frac{\text{Observed score} - \text{Mean score}}{\text{Standard deviation}} = \frac{30 - 32}{3.3} = -0.61$$

The standard score of –0.61 indicates that the score of 30 is 0.61 standard deviation units below the mean. Suppose the same individual received a standard score of –1.2 on a different assessment; then, we would immediately know that the individual's performance on the first assessment was relatively better than her performance on the second assessment because the standard score of –0.61 is higher than –1.2.

CRITERION-REFERENCED ASSESSMENTS

Criterion-referenced assessments are used to determine an individual's performance by comparing it to a predetermined criterion or standard for the purpose of making a decision or classification (e.g., skill level, mastery, proficiency, certification). According to Crocker and Algina (1986), the commonly used term *criterion-referenced measure* is actually an abbreviation for the term "criterion-behavior-referenced measurement," which implies that measurements are to be interpreted in terms of the criterion behaviors an individual can exhibit. These types of assessments make no direct reference or comparison to the performance of other examinees. Criterion-referenced instruments either indicate the likely proportion of correct responses that would be obtained on some larger domain of similar items or indicate that an examinee's level of tested skill is adequate to perform successfully in some other setting (AERA et al., 2014).

There have been a variety of criterion-referenced assessments developed within the ABA framework to identify an individual's strengths and weaknesses in order to develop skill acquisition programs. This type of measurement

is well-suited to meet the needs of behavior analysts. Criterion-referenced assessments have made a huge impact on how behavior analysts identify target behaviors, develop treatment plans, and monitor progress to enhance an individual's skill repertoire (Padilla, 2020).

Over the last 30 years, several instruments have been developed within the ABA framework that specifically target skill acquisition for individuals diagnosed with ASD. Assessments commonly used include the Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP), Assessment of Basic Language and Learning Skills—Revised (ABLLS-R), and the Promoting the Emergence of Advanced Knowledge (PEAK) Relational Training System, with the VB-MAPP being the most common (Austin & Thomas, 2017; Padilla, 2020). The VB-MAPP (Sundberg, 2014) and PEAK (Dixon, 2014) were both initially developed in 2008 whereas the ABLLS was initially published in 1998 (Partington, 2006). According to its manual, the VB-MAPP is described as a criterion-referenced assessment, curriculum guide, and progress-monitoring tool designed for parents and professionals to gain information regarding their child's language and social skills for individuals aged 0–48 months.

Scoring Guidelines

In the simplest case, scoring guidelines may provide guidance for how to code a behavior that should or should not be classified as an occurrence. The administration protocol that accompanies a standardized assessment will most likely provide instructions for or examples of acceptable responses or behavior. In such cases, the response or behavior may be coded as "1" to indicate that the behavior was a positive example of the target behavior or skill. Some assessments may allow for what may be thought of as *partial credit* where the behavior or response elicited does not entirely meet the criterion necessary for classifying the behavior as a positive example (i.e., score = 1), but the response was partially related to the target behavior (i.e., score = 0.5). There may also be more complex rules for assigning scores to observed responses. Given a prompt or stimulus on a standardized assessment, different responses may be weighted more or less heavily than others. Scoring rules would provide guidance on which behaviors should be coded with higher scores versus those that should be coded with lower scores.

There are a number of important considerations during the development of such scoring guidelines. Guiding questions may include, who determines which behaviors or responses are acceptable or worth different point values? On what information are these decisions to be based? After composite scores are calculated, how are cutoff scores established for making decisions about when an individual's level of skill or behavior is high enough for a certain classification? The answers to these questions will depend on such factors as the target populations and behaviors, the purpose of the assessments, the format of the assessments, and so on. The most common source of information for

answering the questions posed above are consultants who are subject-matter experts, or SMEs. These experts should have specific knowledge and experience in the relevant theories and previous research that can (and should) justify the answers. It is also critical to ensure that the consulting experts include representatives of populations with characteristics important for making assessment decisions; such characteristics may include gender, race/ethnicity, geography, viewpoints, and training history. Ultimately, demonstrating diversity of these types of characteristics among the consultants can be used to support the fairness of decisions made based on the assessment results, should the decisions ever be questioned.

The considerations above apply to both item- and task-level scoring, as well as scores from an assessment as a whole. The general term for a set of scoring guidelines for items, tasks, performances, or behaviors is rubric. A **rubric** is a tool that aids in scoring by describing the levels or types of responses needed for understanding behavior, achievement, or some other characteristic of an individual. In education, for example, rubrics are used for scoring student performances to make decisions about proficiency or mastery of some content. They are commonly used for scoring performances and products, such as essays, speeches, or projects, that require human observers to assign scores. Rubrics promote consistency of scoring across individuals for the same human observer, as well as scoring across observers by providing a set of criteria for scoring, or classifying, behavior. Within the context of behavioral assessment, rubrics could be used to aid observers when they are scoring or classifying behaviors as part of an intervention or when they are seeking to identify the function of an individual's behavior. In fact, agreement or consistency between observers is typically based on the scores or classifications of behaviors recorded using a rubric. Furthermore, rubrics could be used to evaluate whether an assessment used to identify the function of an individual's behavior was conducted appropriately. These types of FBAs are discussed in more detail in Chapter 5 of this text, but our point here is to demonstrate the widespread utility of rubrics as scoring guidelines in social and behavioral research.

To this point, we have described the use of rubrics as scoring aids for criterion-referenced assessments, but criteria can also be established for interpreting total scores. As mentioned above, assessment developers commonly consult with subject-matter experts (SMEs) to establish the criteria against which individuals' total scores are compared. One general process used for determining these interpretive criteria is called **standard setting**, which often involves multiple rounds of review/scoring by experts and can span several days. Interested readers should see such texts as Cizek (2012) or Cizek & Bunch (2007) for excellent and thorough presentations of standard setting, which is common in educational and psychological measurement. Regardless of the specific process used to establish the interpretive criteria, a group of experts with knowledge and experience directly related to the target behavior or domain are generally involved.

Reporting and Interpretation

For criterion-referenced assessments, the total scores, domain scores, or some type of summary score per individual is commonly reported, but the scores are not compared with the scores from any other individuals. Instead, they are compared with the criteria that were preestablished. The criteria are often cutoff scores, meaning that scores below the cutoff are classified one way and scores above the cutoff are classified another way. One common example of this type of system is an assessment used for granting licensure or certification; that is, examinees must meet or exceed a certain cutoff score in order to be licensed or certified in some professional area. A second common example with which many readers will be familiar are end-of-year tests used in public education; in this case, students need to meet or exceed a certain score on the test in order to be promoted to the next grade level. As can be seen, both examples involve comparing an individual's scores against some criterion in order to determine what the individual knows or can do. The same is true for behavioral assessment. An individual who has undergone a skill-based assessment may receive scores for domains such as tact, mand, echoics, play, and so on. The score for each domain is as an indication of the level of this skill in the individual's repertoire. Based on this information, intervention plans could be designed to develop the skills on which the individual may have a deficit. Again, the domain scores for this individual are not compared to those of any other individual; rather the domain scores are the output from an observer comparing and scoring behaviors in accordance with interpretive criteria. Depending on the assessment and scoring criteria, there may be more prescriptive guidelines about present levels of behavior or function based on specific domain scores or domain score ranges. We provide further discussion about assessment-specific scoring, reporting, and interpretation in Chapter 6 of this text.

CONCLUSIONS

In this chapter, we have discussed considerations for the administration, scoring, and interpretation of norm- and criterion-referenced assessment. There are several commonly used types of scores used for expressing the relative position of a score within a larger distribution. Although using relative scoring in ABA may not seem fair or common on its face due to the characteristics of the individuals that are often assessed in ABA research and practice, norm-referenced assessments are increasingly being used in ABA (Padilla, 2020). We present and discuss assessment use in ABA research and practice in Chapter 5. Before that presentation, deeper conceptual knowledge of the concepts of reliability and validity, as well as associated types of evidence, is needed. Reliability and validity are the focus, respectively, for Chapter 3 and Chapter 4.