Preface

I wrote this book to fill a gap—to try to explain the basics of psychometrics to the people who need that understanding most: *test users*. Professionals such as school psychologists, special educators, counselors, speech–language pathologists, and social workers sometimes administer tests and scales, or review evaluation reports with test data, and need to make sense of the test scores they see. My own students are preparing to become these kinds of professionals; most of them are in a master's degree program, getting trained to become practicing school psychologists. I had five goals for the book, and I knew of no other books that met those goals at the same time. Specifically, I wanted the book's content to be (1) accessible, (2) intuitive, (3) concise, (4) clinically relevant, and (5) technically accurate:

1. *Accessible*—Most test users have only minimal background, if any, in statistics and may be avoidant (even somewhat afraid) of mathematical material. Unfortunately, many psychometrics books assume statistically sophisticated readers and are stingy in providing clear and detailed examples. I aim to cover mathematical formulas only when necessary for understanding and to focus on conceptual descriptions.

2. *Intuitive*—I wanted test users not only to know (for instance) the relationship between the reliability coefficient and the standard error of measurement, but also to understand *why* the relationship makes sense. It's easy enough to present formulas and definitions; it's harder to show readers why these formulas and definitions are logical rather than arbitrary.

3. *Concise*—My students don't have time for a lengthy treatise exploring everything that we can do with measurement; the basics are enough, provided that those concepts are clearly learned and understood. Similarly, most test users (and those studying to be the same) can't devote a full semester to psychometric theory. Therefore, I aimed to write a book that could be used as a textbook for just several class periods and that could serve as a quick reference and resource for practicing test users who need to "study up."

4. *Clinically relevant*—Many psychometrics textbooks take the point of view of a test developer or researcher who is interested in large datasets. My students are training to be clinicians; they are interested in applying psychometrics to the individual client that they have in front of them. I have tried to use clinical examples throughout, referencing a variety of professions.

5. *Technically accurate*—When offering accessible and intuitive information about psychometrics in a concise format, it is tempting to "fudge" some of the technical niceties—to say things that aren't exactly accurate, oversimplifying the material. While writing, I sometimes felt that a psychometrician was looking over my shoulder, and I tried not to incur their disapproval too often.

Aids for Learning

I've incorporated a number of features into the book to make it more useful for instructors and for students as a learning tool:

- Most of the chapters end with a set of applied exercises. Some of the exercises give students an opportunity to calculate various test score data, but generally the exercises focus on interpretation of test scores. Many exercises ask students how they would respond to questions or concerns about tests, and they give students an opportunity to practice explaining psychometric ideas to various audiences (examinees, parents, etc.). In addition, I have provided suggested answers to the exercises in Appendix B of the book.

- In Appendix A of the book, I have written an annotated guide to more advanced books and even some journal articles on psychometrics.

Although I worked hard to make this book accessible and concise, that also meant recognizing that it's only a first step for students who want to do psychometric research or otherwise obtain more advanced skills.

- The appendices are followed by a glossary of key terms (including any term used in **boldface** in the book). This provides a ready reference for readers who might not always recall a term from an earlier chapter or who use the book primarily as a resource in which to look things up.

- In the final chapter (Chapter 8), I include excerpts from sample evaluation reports to show how to further apply the principles discussed throughout the book. This chapter is designed as an additional resource for students and professionals who are writing up test score data for their own reports.

Acknowledgments

First, it's a pleasure to thank my editor at The Guilford Press, C. Deborah Laughton, for her encouragement, guidance, and feedback throughout the process. I'm also so grateful to experts who reviewed the book for Guilford and provided feedback prior to publication: J. Gayle Beck, Matt Burns, Randy Floyd, and Laura Spenceley. It was Laura whom I first approached with the idea of this book in the Fall of 2020, and her reaction motivated me to actually write it! At Guilford, it was another Laura—Laura Specht Patchkofsky—who deftly handled production issues.

More generally, while writing this book, I often thought about the courses in psychometrics that I took as a student, and I'm grateful to Hoi Suen at Penn State and Brian Martens at Syracuse for the excellent foundation they provided. While they have very different instructional styles, both professors put an interesting behavioral spin on psychometrics, and I continue to use the models and metaphors they taught me. While in graduate school, my advisor, Larry Lewandowski, and my clinical mentor, Michael Gordon, showed me both the power and limitations of psychometric data, as applied to high-stakes decisions; they both served as additional inspiration for this book. Finally, for assistance with reviewing chapters, checking exercises, and formatting materials, I'm grateful to my Teachers College colleague Matt Zajic and my students Thea Bucherbeam, Tess Schaberg, and Lisa Szczesniak.

This book is the culmination of many years spent thinking about measurement. In some sense, I've been interested in testing and measurement since childhood. On the rainy days when I "played school" with friends as a child, I would actually make tests for others to take. At the age of 16, I discovered psychology and psychological testing, and when I learned that whole careers in assessment were possible, I knew that I would have one. For over 15 years now, I've been teaching psychometrics and psychological testing in classes for undergraduate and graduate students, and I've been giving workshops on assessment issues for test users such as psychologists, teachers, learning specialists, and counselors. It's been a genuine privilege to be able to do all this work and now to finally include some of my material in a book.

If you are an instructor, I hope that this volume serves your needs in the classroom, and if you are a student or a practitioner, I hope that you find this volume helpful in your work. If you have any comments or suggestions, please contact me at *BL2799@tc.columbia.edu*.

CHAPTER 3

The Meaning of Test Scores

Thirty-six is the highest possible score on the ACT college admissions test, but 36 is also an extremely low IQ score. If you receive a 36 on a class midterm exam, you might be pleased if the score is out of 40 possible points, but you would probably be distressed if the 36 was out of 100. The same score carries very different meaning depending on the test's **scale**—the range and distribution of scores. This chapter covers common scales for different diagnostic tests and the ways that the resulting scores are interpreted.

Psychometricians make a broad distinction between two ways of interpreting scores. First, **norm-referenced score interpretations** involve comparison of different people to each other. A norm-referenced test score will tell the test user how the examinee performed or responded relative to other people. An IQ score is a typical norm-referenced score; an IQ of 100 doesn't mean that the examinee got 100 items right or 100% of the items right, but that the examinee performed exactly average for someone of their age, better than about 50% of people in their age group. Generally, norm-referenced tests are designed to show differences between individuals; if everyone received the same IQ score, the test would not be very useful. Norm-referenced tests are therefore useful for selection, classification, and similar decisions. Most diagnostic tests are norm-referenced, and most of this chapter will focus on them.

Another way of thinking about norm-referenced scores is that they tell us how common or rare someone's level of performance or functioning is. An IQ in the average range is, by the statistics of the normal distribution, very common. However, an IQ of 70 or below is exceedingly rare, obtained by only about 2% of people in the general population. In the same way, but in the opposite direction, an ACT score of 36 is quite rare, obtained by only the top <1% of the students taking the test in 2020–2021 (ACT, n.d.). Since norm-referenced scores tell us not just how one person did but how they did relative to others, the scores give us a good sense of whether the person's score was typical (near the average) or *unusually* high or low.

A second type of interpretation—a **criterion-referenced score interpretation**—involves the comparison of an examinee to an absolute standard (a criterion). A simple example would be the score from a road test used to license automobile drivers. Typically, the road test consists of a number of tasks, and errors cause the driver to lose points. If a particular road test has 100 possible points and scores of 90 and above are considered passing, the score of 85 has a direct interpretation compared to the mastery standard of 90; the hopeful driver has failed the test. Exams that school districts administer for school accountability are also criterion-referenced. The test developer or user sets standards for "proficient," "advanced," and other levels of skill, and a student is judged relative to those standards. If on a math test for eighth graders, scores of 525 and above represent proficiency in mathematics and scores of 566 and above represent advanced skills, a student who earns a score of 540 is thought to be proficient but lacks advanced skills in mathematics. A criterion-referenced score does not tell us how other examinees performed. If a school district implemented widespread instructional reforms that led to all eighth graders getting scores of 566 or above on the math test, the meaning of any individual's score would not change. Instead, the data would suggest that all eighth graders in the district now had advanced skills in math.

Test users often refer to norm-referenced and criterion-referenced *tests* or test *scores*. However, technically, it is the score *interpretation* that is norm- or criterion-referenced. For instance, the road test for the driver's license exam is typically interpreted in a criterion-referenced way, but if a city wishes to honor the most prepared young drivers, the Department of Motor Vehicles could identify those examinees who earned road test scores in the top 10% of those tested, interpreting the scores in a norm-referenced fashion. When I refer to norm-referenced tests or scores, I am referring to tests and scores that are typically (or designed to be) interpreted in a norm-referenced fashion. When I refer to criterion-referenced tests or scores, I am speaking about tests and scores that are typically interpreted in a criterion-referenced fashion.

Norm-Referenced Scores

Scores from diagnostic tests are typically interpreted in a norm-referenced fashion. One of the criteria for a clinical diagnosis (of depression, an expressive language disorder, etc.) is statistical rarity; we diagnose individuals whose trait levels are unusual in some way. If the trait is symptoms of anxiety, unusually high symptom levels might be part of the evidence underlying a clinical diagnosis of an anxiety disorder. If the trait is intelligence, unusually low levels of intelligence might be part of the evidence leading to a diagnosis of intellectual disability. Therefore, the core evidence involves comparing the examinee to other people to determine how unusual their responses are.

Who exactly is the examinee compared to on a norm-referenced test? The test is developed on a **norm group** (also known as a **standardization sample**). In the course of the test's creation, it is given to many people (often hundreds or thousands of people), and when the test is finally in applied use, each new examinee is compared to the norm group or a part of that group. The **norms** for a norm-referenced test show the distribution of scores in the norm group, so that a test user can find out if an examinee's score is average, unusually low, or unusually high. Often, the norm group is divided up into subgroups by demographic features, especially age. Therefore, if we are assessing mathematics skills in an 8-year-old child, we can compare their skills to those of just other 8-year-old children. Such a comparison group should always be specified, since the nature of the group can have a large impact on norm-referenced scores. I return to this topic later in the chapter, when discussing the importance of appropriate normative comparisons.

Norm-referenced scores are calculated by starting with an examinee's **raw score**. This might be the number of items they got correct on a reading comprehension test or the total number of points they earned on an essay test where the essay was scored out of 20 possible points. A raw score on a personality/psychopathology measure could be the number of symptoms that the examinee reported having, or the number of statements that they answered "yes" to. The examinee's raw score is then compared to the distribution of raw scores in the appropriate norm group block (which might be the people of the same age as the examinee) to check where their score is in relation to the average score of people in the block. Based on where the examinee's score falls within that distribution, it is transformed to a norm-referenced score. For instance, if the examinee's raw score on an IQ test (i.e., the points they earned for correct answers across all of the items) is at the exact average for their age group, the examinee will be assigned a norm-referenced IQ score of 100, since 100 is defined as the average IQ score at every age level. A norm-referenced score of 100 does *not* mean that an examinee earned 100 points; this score merely means that the examinee's IQ test performance was at the exact average of the score distribution for people of their age who were in the norm group.

Percentile Ranks

Perhaps the most useful norm-referenced scores are **percentile ranks** or percentile scores (sometimes abbreviated as %ile). They tell us what proportion of the population the examinee scored above. For instance, a student who scores at the 66th percentile of a test scored higher than 66% of the norm group, and the norm group is expected to represent the population. A score at the exact average[1] would be at the 50th percentile, and the average *range* is often taken to extend from the 25th to about the 75th percentile; thus, it is the middle 50% of the population. Many rating scales of attention-deficit/hyperactivity disorder (ADHD) symptoms suggest that a score is clinically significant if it is at the 93rd percentile or above—that is, if someone's symptom levels are in the top 7% of the population. The IQ cutoff for intellectual disability is at approximately the 2nd percentile; if a student has an IQ score in the bottom 2% of the population, that is part of the evidence needed for a diagnosis of intellectual disability.

Virtually all norm-referenced tests yield percentile ranks, along with other types of norm-referenced scores. Percentile ranks are easy to understand and to explain to clients, students, and families of children being assessed. Thinking in terms of proportions is intuitive, and laypeople understand why it is unusual that someone's test responses place them in, for instance, the top or bottom 5% of the population. However, on tests of academic skills, percentile ranks are sometimes confused with *percent correct* scores. A parent may hear that their daughter is at the 60th percentile in mathematics and think that the girl is almost failing (as a 60% class grade would suggest), when in fact she is doing better than most of her peers. Make

[1] Technically, the 50th percentile is the median, but in a normal distribution, the mean and median are the same.

sure to explain the difference between percentile rank and percent correct scores when presenting the percentile rank.

You need to consider two more technical caveats when interpreting percentile rank scores. First, they are not on an "equal interval scale"—that is, the difference in trait levels between (for instance) the 10th and 20th percentiles is not the same size as the difference between the 30th and 40th percentiles. Recall that in the normal distribution, most people are relatively close to the average, and the farther that you move away from the average, the fewer people you will find. Percentiles tell you how an examinee compares to other people, so percentiles will be clustered tightly near the average but will be spread out far at the edges of the distribution. The anxiety level of a client with an extremely high anxiety score (at the 99th percentile) might decrease substantially and still be at the 95th percentile. Meanwhile, if a client was at the 60th percentile to begin with, even a small (and clinically meaningless) decrease in anxiety might knock them down to the 50th percentile.

Given this caveat, note that you cannot perform meaningful arithmetic operations on percentile ranks. Calculating the mean of three different percentile rank scores is not accurate, for instance. Relatedly, you cannot meaningfully interpret the size of a gap between two percentile ranks without knowing the exact percentiles that the gap is between. To say, for instance, that a student increased 10 percentile rank units in reading comprehension between September and January of the school year could represent either a small or a large degree of growth depending on where along the normal distribution the growth occurred. Similarly, to say that one student is 10 percentile rank units above another student is not inherently meaningful. More information (such as the exact percentiles) can be helpful, but other types of norm-referenced scores that *do* have equal intervals are preferable for these purposes. In fact, many of the other norm-referenced scores that we cover are treated as having equal intervals.

There is yet a second caveat: in the norm-referenced scores that follow, I give percentile rank equivalents, but those equivalents assume an approximately normal distribution. Many score distributions, particularly for performance tests (cognitive and achievement tests), have approximately normal distributions, especially in children and adolescents. However, some neuropsychological tests do not (since almost everyone without brain damage does well), and many rating scales of clinical symptoms and problem behaviors also do not (since most people receive low raw scores, showing few symptoms, while those with clinical problems are spread far into the high end of possible raw scores). On tests with non-normal distributions, percentile ranks are particularly important, and they will not match up exactly with other types of scores. However, Figure 3.1 shows the expected relationships between the scores for normally distributed data.

z-Scores (*M* = 0, *SD* = 1)

We first encountered ***z*-scores** in Chapter 2, since these scores serve as landmarks along the normal distribution. A *z*-score tells us, quite literally, how many standard deviations away from the mean a score falls. If a counselor uses a norm-referenced scale to measure a client's level of anxiety and the client has a *z*-score of 0, the counselor knows that the client's reported anxiety level is exactly average, since the score is 0 standard deviations away from the mean. Negative *z*-scores are below the mean, whereas positive *z*-scores are above the mean. And almost everyone will have a *z*-score between –3 and +3. These features give *z*-scores some intuitive appeal to test users who understand psychometrics, but very few norm-referenced diagnostic tests actually use *z*-scores for their primary reporting method. I suspect that this

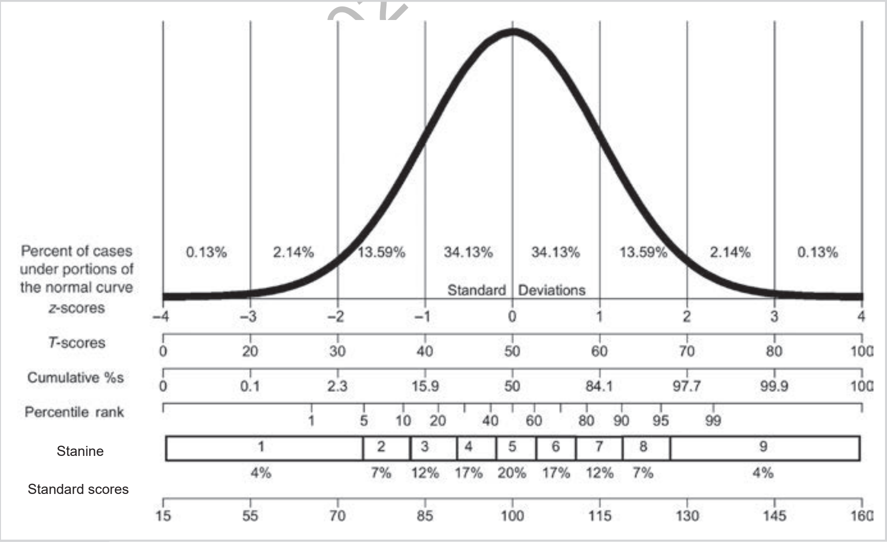

FIGURE 3.1. Normal distribution with norm-referenced score scales shown. From Bandalos (2018, p. 31). Copyright © 2018 The Guilford Press. Adapted by permission.

is because these scores are difficult to explain to laypeople. A score of 0 sounds like the examinee didn't get any items correct (or didn't report any symptoms), and a negative score is even harder to explain. Even so, *z*-scores are important for practitioners to know about, both when reading research articles and when thinking about how scores on different tests compare to each other. Indeed, *z*-scores provide a common metric for comparing the other types of norm-referenced scores presented in this section. If you are looking at scores from a battery of diagnostic tests, each of which uses a different type of norm-referenced scores, you can still think about the various scores that an examinee received as being, for instance, about half a standard deviation below the mean, two standard deviations above the mean, and so forth.

Standard Scores (*M* = 100, *SD* = 15)

Standard scores are common scores on tests of cognitive abilities, academic skills, adaptive behavior, language functioning, and related areas. They are used for the famous IQ test scale. Standard scores have an average (mean) of 100 and a standard deviation of 15, so a score of 70 is 2 standard deviations below the mean ($z = -2.00$) and an IQ of 70 is typically the cutoff for intellectual disability (the condition formerly known as mental retardation). On any given test, 68% of the population will have standard scores between 85 and 115, and 95% of the population will have scores between 70 and 130. The more you work with standard scores, the more you will develop an intuitive feel for what counts as a "high" score, a "bad" score, and so on. In particular, it is helpful to know the percentile ranks of common standard score landmarks: a standard score of 70 is at the 2nd percentile, a standard score of 80 is at the 9th percentile, a standard score of 90 is at the 25th percentile, and so on.

Subtest Scaled Scores (*M* = 10, *SD* = 3)

Many intelligence tests and measures of adaptive behavior (used to assess individuals with developmental disabilities) have subtests that use "scaled scores" with a mean of 10 and a standard deviation of 3. On this scale distribution, scores between 8 and 12 typically constitute the average range, being between the 25th and 75th percentiles. However, keep in mind that making important clinical decisions or other bold interpretations based on a single subtest can be hazardous, so make sure to examine the reliability and validity evidence for the specific subtests being interpreted.

T-Scores (*M* = 50, *SD* = 10)

T-scores are fairly common,[2] used in some cognitive ability and neuropsychological tests as well as many measures of psychological disorders. Behavior rating scales for children often use *T*-scores, as do clinical personality tests (e.g., the Minnesota Multiphasic Personality Inventory). Because they have a mean of 50, *T*-scores can be mistaken for percentiles, but they are actually very different. They are on an equal-interval scale, and because their standard deviation is 10, if the scores are normally distributed, almost all people will have scores between 20 and 80. Earlier I mentioned the ADHD symptom rating scales that use a 93rd percentile cutoff for clinically significant symptom levels. Most of these scales generate *T*-scores, and the cutoff is $T = 65$ (1.5 standard deviations above the mean, which is the 93rd percentile).

Stanines

Many achievement tests also have another type of norm-referenced score, the **stanine**. The stanine scale divides the normal distribution into 9 score ranges, with 5 being the middle range (the middle 20% of the distribution, incidentally), 1 the lowest, and 9 the highest. Typically, stanine scores of 4, 5, and 6 are considered the average range (together, they represent a bit more than the middle 50% of the distribution), with below-average and above-average scores being below and above that range, respectively. The stanine does not have any unique advantages over other norm-referenced scores, but because it only has 9 score ranges, some test users may find it simpler to interpret.

Age-Equivalent and Grade-Equivalent Scores

On tests in the areas of intelligence, academic skills, and speech/language, test users often have the option of recording **age-equivalent** and

[2] *T*-scores do not have anything to do with the *t*-test, an inferential statistic that is used for comparing two groups.

grade-equivalent scores. Also known as *developmental scores*, they are highly controversial. They are popular with some test users, but many scholars argue that they are so misleading that they do more harm than good.

If a student receives a grade-equivalent score of 5.3 on a test of reading skills, this is supposed to indicate that the student performed at the same level on the test as the median child in a sample of children in the third month of their fifth-grade year. Similar notation is used for age-equivalent scores; a child with an age-equivalent score of 8–11 is assumed to have performed like the median child in the norm group who is 8 years and 11 months old. Already there's one problem with this kind of definition: there may not be any children in the norm sample who were tested in exactly the third month of fifth grade, and so test developers might need to examine the scores of students at other developmental points nearby and infer where a child in grade 5.3 should perform. This process, known as *interpolation*, assumes that skills increase the same amount each month, which isn't necessarily the case.

Bigger problems with developmental scores come with misuse; not only is their official definition problematic, but many people interpreting the scores go far beyond that definition. For instance, if the student receiving a grade-equivalent score of 5.3 was just starting sixth grade (6.0), a teacher or parent might express concern that the student was "almost a year behind his peers," but in fact this would be inaccurate. It is common for a class of students to show fairly wide variability in academic skills, and so a grade-equivalent score of 5.3 likely puts the child near the average for their peers in that class. Such a misinterpretation also assumes that the scores are perfectly reliable, which is never the case (an issue discussed in detail in Chapter 4). Another common-but-incorrect interpretation is that the student reads the same way as students in the third month of fifth grade do. That assumes that the items on the test and the way that they're scored include all relevant aspects of the reading process, when they probably don't. Students of different age and grade levels will approach test items in different ways, even if their final scorable response is the same. Finally, some test users might infer that a student should be retained or given remedial instruction, or instead placed in a more advanced instructional setting, based solely on developmental scores: "He reads like a fifth grader, so why are we making him take sixth-grade reading lessons?" In fact, these are complicated decisions requiring far more information than someone's developmental score.

Each of these limitations applies equally to *age*-equivalent scores.[3] For these reasons, I cannot endorse the use of developmental scores, except in rare situations where the limitations of the scores are made clear. Other norm-referenced scores are easier to interpret and do not have all of these limitations to the same degree. Still, it is very important for test users to know about these scores; you will see them in practice, and you should know their limitations.

Transforming Norm-Referenced Scores

Figure 3.1 shows how several of the different norm-referenced scores relate to each other along the normal distribution. Extending a vertical line at any point along the distribution on the figure will show all of the norm-referenced scores at that point. In addition, you can find a "psychometric conversion table" on the internet to help with these transformations. Finally, if you are seeking a percentile rank and cannot find one, you can use an online calculator to transform whatever kind of score you have into a *z*-score, and then find out what percent of the population is below that *z*-score (i.e., the percentile rank) here: *www.calculator.net/z-score-calculator.html.*

What's Normal? What's Not?

In clinical assessment, regardless of the field, practitioners are typically focused on the question of whether a patient, client, or student is experiencing problems to an unusual degree. One criterion for a clinical diagnosis or a determination that clinical or educational services are needed is *statistical deviance*, or *deviance from the norm*.[4] Problems are present to a clinical degree in part because they are unusual. Norm-referenced tests can be extraordinarily helpful in these cases, but the term *unusual* is obviously ambiguous. There is no single point where someone's level of functioning

[3] The two types of scores (age-equivalent and grade-equivalent) also can lead to different conclusions, particularly among older students. As I discuss in more detail below, this is particularly problematic when assessing college students and other adults.

[4] Statistical deviance is not sufficient for a clinical designation, but it is one criterion. Typically, such a designation also requires functional limitations—that is, difficulties in everyday, real-world functioning of some kind, or at least significant distress.

suddenly goes from being normal to being abnormal. To take a medical example, if a systolic blood pressure of 140 is the cutoff for hypertension, this does not make a blood pressure reading of 139 perfectly fine!

Acknowledging that there is an element of arbitrariness in any cutoff for the "abnormal" range, we see that there are two sources of such cutoffs. The first is found in legal and policy regulations. For instance, perhaps a governmental agency decides that to be eligible for early intervention services, preschoolers must score below the 20th percentile in at least one area of development. When regulations define cutoffs, clinicians can easily cite and follow them. The second source of cutoffs is trickier to follow consistently: the narrative descriptions of test score ranges found in diagnostic test manuals. Test developers genuinely try to be helpful to practitioners by offering narrative descriptions such as "below average," "extremely low," "at risk," and "borderline clinical," but the terms and cutoffs differ from one test to another. A universal score interpretation system (Guilmette et al., 2020) has recently been proposed, but the current situation is unlikely to change quickly, since test publishers, test authors, researchers, and clinicians would all need to "get on the same page."

Earlier, as I was discussing the various norm-referenced score types, I mentioned some of the typical cutoffs for different tests. First, on measures of performance (e.g., cognitive, academic, neuropsychological tests), the "average range" is generally considered to be scores from the 25th up to the 75th percentile, which (in normally distributed scores) works out to standard scores between 90 and 110. (Sometimes, you will see the 74th percentile or a standard score of 109 used as the upper bound of the average range.) Many tests describe the 10 standard score points on either side of this range as "low average" and "high average." The idea is that standard scores between 80 and 90, and between 110 and 120, are not grossly deviant from average. Less than 20% of the population has standard scores either below 80 (which is at the 9th percentile) or above 120 (which is at the 91st percentile). On IQ tests, scores in these ranges are often referred to as "low" (on one side) or "superior" (on the other). Cutoffs even farther from the average are often described as "extremely low," "very superior," and so on.

On rating scales for psychopathology, high scores typically indicate higher levels of symptoms (i.e., more severe problems). Below-average and average range scores are generally viewed the same way, as simply indicating a lack of clinically significant problems. The cutoff for clinical significance is typically at either 1.5 standard deviations above the mean (z = 1.5, T = 65, or the 93rd percentile), or at 2 standard deviations above the mean (z = 2.0, T = 70, or the 98th percentile). At times, a lower standard (at z = 1.0 or 1.5) is used to define an "at-risk" threshold, suggesting a higher likelihood of problems developing.

Issues in Norm-Referenced Score Interpretation

The Importance of the Norm Group

On any norm-referenced test, the biggest determinant of the score is the group of people to which someone is being compared—that is, the norm group. Therefore, a key indicator of test quality is a good norm group, and a key to valid test interpretation is appropriate norms.

Size is one feature to look for in norm groups; all other things being equal, larger norm groups are better than smaller norm groups. However, what is most important with regard to size is *not* the norm group size as a whole (i.e., the total number of people included in the test development sample); it is the size of individual norm group *blocks*—the groups of people against which an individual's scores are compared. For instance, consider an IQ test that has been normed on 2,000 people—an impressive accomplishment! Dakin, a boy who is 8 years and 6 months of age, will not be compared to all 2,000 people; instead, he may only be compared to 100 children in the norm sample who are between 8 years, 5 months, and 8 years, 8 months, of age. It certainly makes sense to compare him to close age peers, but 100 is a far less impressive comparison group than 2,000. Tests with larger norm group blocks are preferred, regardless of the total sample size.

Representativeness is another important norm group feature. Generally, in the United States, norm groups are sought to be representative to the general population of the country. Often, test manuals will compare the demographic characteristics of the norm group to statistics from the U.S. Census, with particular regard to gender, ethnicity, and geographical location. For instance, if 85% of the people in a norm sample were men, this would vastly overrepresent men relative to their proportion in the general population. At times, age is another demographic factor matched to the Census, although on many tests, there are separate norms by age, making this type of matching unnecessary. In any case, test users should review the characteristics of the normative sample described in the test manual to ensure reasonably representative norms.

A final important norm group feature is *recency*. The average level of traits sometimes changes over time in a population. Therefore, all other things being equal, a test with more recent norms is preferable to one with older norms. For instance, to infer that a student's reading skills are at the 80th percentile relative to age peers, it is always most helpful to use a test where those peers (from the norming sample) were tested recently. Even IQ tests have shown average raw score performance changes over time, and so measurement of intelligence is most accurate when an examinee is compared to a norm sample from recent years. This is one reason why most diagnostic tests are revised and renormed every decade or so.

Norms Based on Demographic Groups

At times, test users have the option to compare someone's score to a group other than the general population or age peers.

Gender-Specific Norms

On many questionnaires and rating scales that measure emotional and behavioral problems, as well as personality inventories, norms are available (and are sometimes *only* available) by sex/gender. It can seem attractive to use gender-specific norms. For instance, when rating a young boy's level of hyperactivity symptoms, it might seem fair to compare him to other boys rather than all children his age, since boys are thought to be "naturally" more hyperactive. This comparison would avoid unfairly penalizing him for being a boy and risking pathologizing his typically male behavior.

From a diagnostic point of view, however, gender-specific norms have significant limitations. By definition, they erase actual gender differences in the traits they measure. For instance, if the 93rd percentile (e.g., a *T*-score of 65 or above) is the cutoff for clinically significant anxiety symptoms, gender-specific norms will make it so that 7% of males *and* 7% of females would meet clinical significance. But in fact females have far higher rates of clinically significant anxiety, often experiencing anxiety disorders at twice the rate of males (Hartung & Lefler, 2019). Similar gender differences (in both directions) are present for many other disorders and personality traits.

These differences can only be seen in norms that combine data across gender identities.

Gender-specific norms are also difficult to apply to the increasing number of clients who identify as transgender or nonbinary. In the case of transgender clients, using norms for the gender corresponding to their gender identity does not always work, particularly for children. When working with clients who identify as nonbinary, if a test only has gender-specific norms, it is best to score the test using both sets of norms and to view the client's true scores as lying somewhere between the two options (since that is what combined norms would yield). More generally, combined norms are to be preferred where they are available, except in specific cases where behavior relative to gender expectations is relevant.

Education-Group Norms

On many cognitive and achievement tests, norms are available not just for different age groups but also for different grade levels. A 12-year-old in sixth grade can be compared to other 12-year-olds or other sixth graders. Through the childhood and adolescent years, age- and grade-based norms yield similar scores for most examinees, the exception being students who are significantly older or younger than most people in their grade year. For individuals in late adolescence and adulthood, age and education norms often yield vastly different scores, since a significant proportion of the population does not attend higher education for long, if at all (Harrison et al., 2019). Comparing a 22-year-old college senior to other college seniors is quite different from comparing that student to all fellow 22-year-olds. Comparing a 25-year-old medical student to other students in their third year of graduate/professional school is *extremely* different from comparing the student to all 25-year-olds in the general population. Higher-ability individuals are more likely to seek more education and to be successful in their applications to education settings; moreover, education directly increases cognitive and academic skills.

Generally, for diagnostic and other clinical purposes, age norms are preferred, even in childhood, and they are certainly the most appropriate norms in older clients. However, educational norms can be helpful for making recommendations and inferences about a student's likelihood of success in various educational settings, and so they may be helpful to calculate when offering advice or counseling regarding educational placement decisions.

Norms Based on Race, Ethnicity, and Cultural Background

Many neuropsychological tests yield scores that are normed based on race or ethnicity as well as age, education, and gender. The original rationale for *race norming* was that neuropsychological tests were designed to assess the organic, biological effects of brain injury or degeneration, and the standard against which the examinee should be judged is that of their cultural peers, to eliminate the influence of cultural factors. Arguably, race norming could also affect the identification of students for special education services; currently (without race norming), a higher proportion of African American students are identified than other groups, although this appears to be explained by differences in academic performance and other factors besides race per se (Morgan et al., 2017).

Race norming is a controversial practice. Most recently, it has made news for its use in identifying neurological impairment among football players seeking compensation for play-related injuries (Associated Press, 2021). For clinical diagnostic decisions, combined norms are generally preferred, although in cases where the diagnosis depends on a decline in neuropsychological functioning, race norming continues to be used in some settings. As a consumer of evaluation reports, be careful to note whether any neuropsychological testing you review mentions "demographically corrected" norms, as they will likely include race, and this should affect your interpretation of the scores. Specifically, African American examinees may have significant increases in their scores on neuropsychological or cognitive tests in the presence of race norming, relative to the scores obtained with combined norms.

Extremity in Composite Scores

One of the most confusing situations for a practitioner involves a composite test score that is farther from the average than any of the scores making up the composite. For instance, consider a test of children's oral language skills, with a total language score made up of two subscores: one in expressive language (speech) and one in receptive language (listening). A child receives the following scores (on the standard score scale):

Expressive Language = 82
Receptive Language = 78
Total Language = 73

If the child's overall language skills are, in some sense, an average of their expressive and receptive language, why isn't the total score in the middle of the other two scores? At times, these phenomena lead to composite scores that meet clinical cutoffs for severe deficiencies, when none of the subscores making up the composite meet the cutoff. In these cases, it can be difficult to explain the situation to clients, families, and administrators.

Remember that norm-referenced scores tell you how rare a score is. In the above example, the Expressive Language score (82) is at the 12th percentile, meaning that only 12% of the population had lower scores than that. The Receptive Language score (78) is at the 7th percentile, so only 7% of the population had lower scores. A composite score based on those two subscores must consider how rare it is to have significant deficits in expressive *and* receptive language. This is rarer than just a single low score in one area of language.[5] Therefore, the composite score will be lower than either of the subscores that make it up. The Total Language score of 73 is at the 4th percentile, suggesting that only 4% of children have such poor overall language skills.

Whenever both (or all) of the subscores are on the same side of the mean, the composite will not be in the middle of the subscores—it will be farther from the mean than the average of the subscores. The degree to which the composite will be more extreme will depend on how correlated the subscores are, but some amount of composite extremity is the rule, not the exception. This occurs in either direction from the mean and should be expected, and it should be explained to clients and others in terms of the rarity of having multiple areas of functioning (the subscores) that are below (or above) average.

Base Rates of Extreme Scores in Batteries

We just saw how multiple extreme scores are rarer than a single extreme score. Relatedly, the more tests (or subtests) that are given, the more likely it is that an extreme score will be found somewhere in the battery. This can occur just by chance, or a few extreme scores in a lengthy battery can

[5] Consider the chance that there will be a hailstorm tomorrow in the city where you live. (The probability is likely relatively low.) Now consider the chance that there will be a hailstorm tomorrow *and* another hailstorm the day after tomorrow. That probability is even lower. The same is true of the probability of one low test score versus *two* low scores.

indicate genuine but very narrow strengths and weaknesses. Regardless, it should not be seen as unusual or statistically deviant to see individual extreme scores in batteries.

Recent studies have tried to quantify the *base rate* (the general population prevalence) of individual extreme scores—in particular, extreme low scores—using the data from normative samples of major diagnostic tests. The base rates have been found to be quite high. For instance, in one of these studies, Brooks (2010) found that *most* children in the normative sample of the Wechsler Intelligence Scale for Children (the fourth edition, the WISC-IV) had at least one subtest score that was at the 16th percentile or below. Almost half of the normative sample (43.4%) had at least *two* subtest scores meeting that criterion. Similar findings have been published regarding other tests.

The results of these studies have important clinical implications. First, avoid going on "fishing expeditions," evaluating a client in areas where there is no referral concern. If you keep assessing different areas, an apparent problem will "turn up," where or not it is meaningful. Second, and relatedly, insist on evidence beyond norm-referenced test scores before diagnosing a problem as a disorder or disability. The additional evidence might come from real-world (nondiagnostic) tests, clear self-reports and informant-reports, structured clinical observation, and so on. Finally, seek converging evidence from multiple diagnostic tests of similar areas of functioning to ensure that apparent problems are not just a statistical fluke.

Conclusions

Diagnostic tests generally provide norm-referenced scores that describe someone's functioning relative to a group of people on which the test was developed. Different types of norm-referenced scores look quite variable, but they can all be equated with percentiles, which is easy to do when the score distribution of a test is approximately normal. Moreover, a particular *z*-score always corresponds to the same *T*-score, standard score, and so on. When deciding how to interpret norm-referenced scores, keep several principles in mind. First, be sure that the norm group is reasonably large and representative of the population it is supposed to embody. Second, avoid norms based on particular demographic groups (other than age) except in unusual circumstances. Third, understand the relationshipt between subtest and composite scores, and why composite scores can sometimes be more extreme than any of the subtest scores making up the composite. Finally, be aware of the base rate of extreme subtest scores, and demand additional evidence to validate interpretations of these scores.

APPLIED EXERCISES

1. Consider the test scores for an adolescent undergoing a psychoeducational evaluation at school, shown in Table 3.1. Where do these scores fall relative to the average range? (You can assume a roughly normal distribution for this exercise.) What kinds of problems appear to be present? What areas of functioning are unimpaired? What areas of functioning are perhaps *better than typical*? To help justify your answers, describe the (approximate) percentiles of these scores.

2. At a special education committee meeting one day in June, the school principal points out that Briana's age-equivalent language development score is a year below her actual chronological age. The principal therefore suggests that the committee consider retaining Briana (not passing her to the next grade) on that basis. Briana just turned 7, but her age-equivalent language score is 6.0. How would you explain the meaning of an age-equivalent score to the principal and

TABLE 3.1. Test Scores for an Adolescent

Type of test/subtest or area of functioning	Type of score	Score
Intelligence—verbal	Standard score	115
Intelligence—nonverbal	Standard score	103
Reading—reading individual words aloud	Subtest scaled score	8
Reading—answering comprehension questions	Subtest scaled score	11
Math—performing calculations	Subtest scaled score	5
Math—application/word problems	Subtest scaled score	7
Parent-report of attention problems	*T*-score	62
Parent-report of hyperactivity	*T*-score	48
Parent-report of anxiety	*T*-score	73
Parent-report of depression	*T*-score	65
Parent-report of conduct problems	*T*-score	34

advise the committee regarding its limitations? What score(s) would you suggest focusing on instead?

3. Robert is a 20-year-old man who recently transferred to Farmland State University after finishing 2 years at Polk County Community College, where his grades were mostly Cs with a few Bs. He has now been experiencing trouble at Farmland State on his exams, and he is actually in danger of failing some of his classes. A psychoeducational evaluation finds that his reading comprehension standard score is 87, based on norms from college juniors nationally. What might be going on here?

4. Jane is a 25-year-old woman who has sought counseling services because she is still quite upset about a romantic break-up that occurred a few months ago. She and her ex-girlfriend had been together for several months, and she is hoping that counseling will help her to move on from that relationship. The counselor, Marla, gives a lengthy trauma symptoms rating scale to all of her clients. The scale has eight subscales, each of which is for a different cluster of trauma symptoms and each of which yields a *T*-score where higher scores indicate more symptoms and where the cutoff for "clinically significant symptoms" is $T = 65$. Jane's *T*-scores are under 60 on seven of the eight subscales, but her score on the remaining subscale is 68. Marla concludes tentatively that Jane is suffering from clinically significant trauma-related symptoms. Why is this conclusion premature, and if you were supervising Marla, what advice would you give her, both in completing this evaluation and for future evaluations?

Guilford Publications
370 Seventh Avenue
New York, NY 10001
212-431-9800
800-365-7006
www.guilford.com