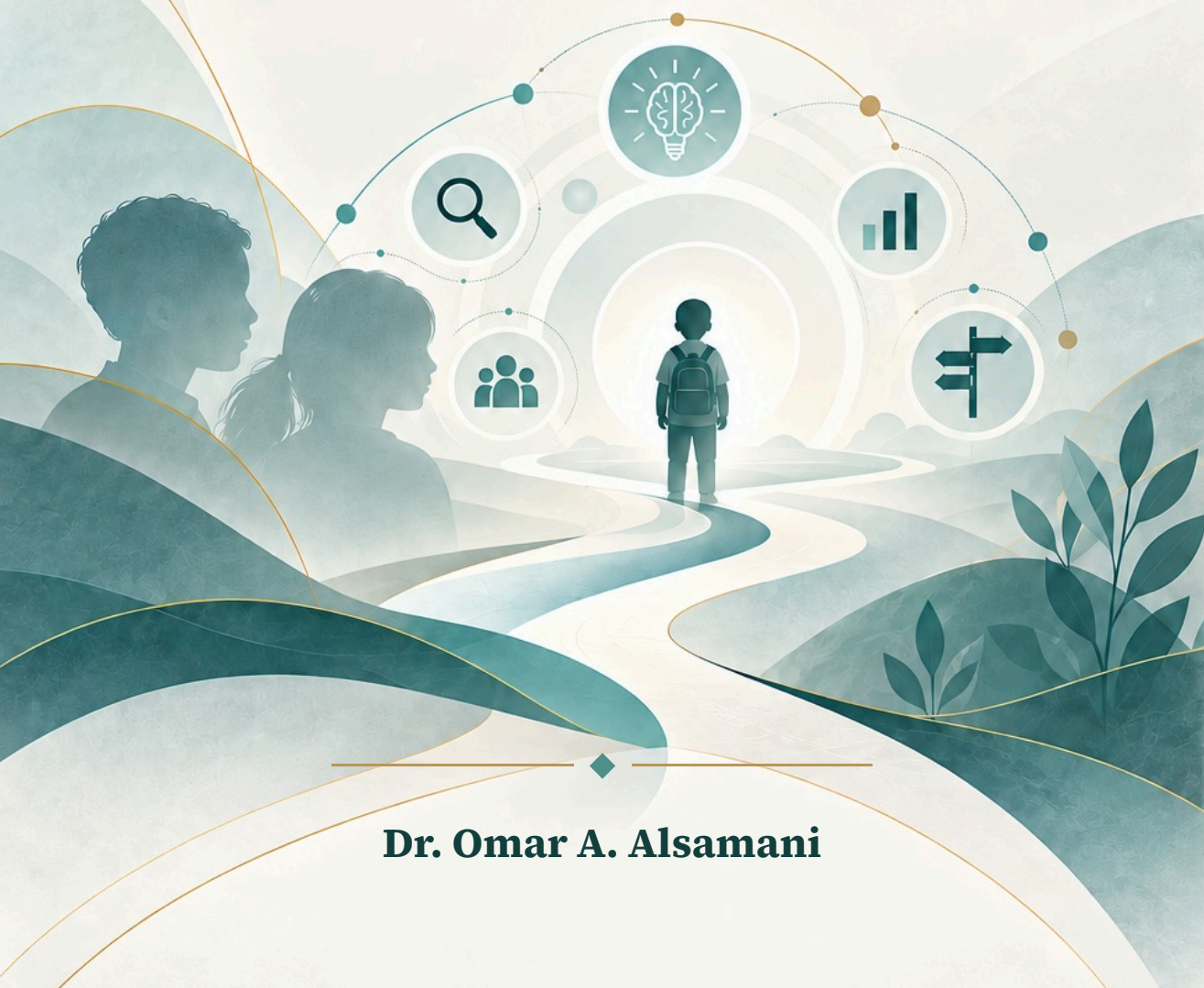


Diagnosis and Identification of Exceptional Learners

Learners with disabilities, the gifted
and the twice-exceptional



Dr. Omar A. Alsamani

Diagnosis and Identification of Exceptional Learners

Learners with disabilities, the gifted
and the twice-exceptional

Dr. Omar A. Alsamani

Diagnosis and Identification of Exceptional Learners

Learners with disabilities, the gifted and the twice-exceptional

This book is published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International licence (CC BY-NC-ND 4.0). It may be copied, distributed and shared, in full or in part, in any medium or format, for non-commercial educational and research purposes, provided it is attributed to its author and its text is not altered. Licence text: creativecommons.org/licenses/by-nc-nd/4.0/

Website: alsamani.com

© 2026 Omar A. Alsamani

Trademarks and instrument names mentioned in this book belong to their owners and are cited for descriptive and educational purposes only. The book reproduces no items from any standardized instrument; diagrams depicting forms are schematic.

First edition, 1448 AH / 2026

Legal deposit no. 1448/7008

ISBN 978-603-06-7714-6

2

Chapter 2

Measurement fundamentals

What does a score of 78 mean? How does a percentile rank differ from a percentage? And why does the psychologist write two more numbers on either side of a score? In this chapter we provide the keys for reading assessment reports of every kind, and we show what a number can tell us and where its meaning stops.

Scene from a team meeting

Umm Majed sits in the school meeting with a six-page report in front of her. She reads: “Full-scale score 72, percentile rank 3, confidence interval 65 to 79.” She looks up and asks the teacher: “Does this mean my son passed with three percent? Does he have a disability or not?”

The teacher hesitates. He knows the number is not a pass rate, but he does not know how to explain that, nor why the report lists two additional numbers around the score.

We return to Umm Majed at the end of the chapter to give her two questions a precise answer. The concepts presented here recur throughout the category chapters, for they are the shared language in which every diagnostic report is written.

Chapter objectives

After studying this chapter, the reader should be able to:

- 1 Classify a reported assessment result as norm-referenced or criterion-referenced, and justify the classification by the question the result answers.
- 2 Convert a score among the common standard score metrics, and locate it on the normal curve and within its descriptive classification.
- 3 Interpret a percentile rank correctly, and correct a statement that confuses it with a percentage or an age equivalent.
- 4 Calculate the standard error of measurement, the confidence interval and the standard error of the difference, and use them to judge a score close to a cut-off or a difference between two scores.
- 5 Calculate positive and negative predictive values from sensitivity, specificity and prevalence, and explain why they differ between universal screening and a referred group.
- 6 Evaluate whether an instrument is appropriate for a particular student in light of the five sources of validity evidence, reliability, norms, language and fairness evidence.
- 7 Identify the ethical or professional principle at stake in a described assessment situation, and propose a course of action consistent with it.

Introduction

Assessment results reach teachers and families as numbers: a standard score, a percentile rank, a confidence interval and a descriptive classification. The percentile rank is the percentage of the student's peers in the norming sample (the representative group from which the instrument's norms were derived) who scored lower than the student did; a percentile rank of 84 means the student outperformed 84% of them. Each of these numbers carries a specific meaning and has limits it cannot exceed, and misreading it leads to a

mistaken decision: a percentile rank taken as the proportion of correct answers, as Umm Majed understood it in the opening scene; a score treated as an exact figure free of error; or a firm decision made when the confidence interval straddles the cut-off. In most categories, the diagnostic judgment rests on where the student stands relative to peers, on how precisely that position is estimated, and on whether the instrument suits its purpose and the student to whom it was administered.

We open the chapter by distinguishing norm-referenced from criterion-referenced measures (Section 2.1), then present the mean and standard deviation (Section 2.2), the normal curve, standard scores and their descriptive classifications (Section 2.3), and percentile ranks and their pitfalls (Section 2.4). We then take up reliability/precision and the standard error of measurement, together with the related matters of score differences, regression toward the mean and floor and ceiling effects (Section 2.5), followed by validity, sensitivity, specificity and fairness (Section 2.6), and then norms, standardization and test adaptation (Section 2.7). We apply all this to reading a realistic score table (Section 2.8), and close with ethics, professional responsibility and levels of qualification (Section 2.9).

2.1 Norm-referenced and criterion-referenced measures

Before we ask “How much did the student score?” we must ask “Compared with what?” A score on its own has no meaning. Eighty points may be excellent on a hard test and weak on an easy one. Measures are therefore divided into two main types according to what the score is compared with:

Norm-referenced measures

Norm-referenced

Where does the student stand relative to peers?

Criterion-referenced

measures Criterion-referenced

Has the student mastered the target skill?

The student's score is compared with the scores of a large group of same-age or same-grade peers, called the **norming sample**. The result tells us the student's position among them: average, above average or below average.

Examples: intelligence tests, adaptive behavior scales, and rating scales such as the Gilliam Autism Rating Scale, Third Edition (**Gilliam Autism Rating Scale, GARS-3**).

The student's score is compared with a predetermined level of performance. The result tells us what the student can and cannot do, regardless of peers.

Examples: "Reads 40 words per minute with 95% accuracy," or "Adds two two-digit numbers with regrouping on 8 of 10 problems."

The two types complement each other. Diagnosis usually calls for norm-referenced measures, because its question is "Is the student's performance unusually far from that of peers?" Instructional planning, by contrast, calls for criterion-referenced measures, because its question is "What skill should we teach next?"

A single tool may combine both logics. One example is **curriculum-based measurement**, a set of brief, repeated probes used to track a student's reading or mathematics week by week (defined and discussed in detail in Chapter 4). It serves as a general indicator of growth: its scores are interpreted normatively when compared with seasonal grade-level norms, such as fall, winter and spring percentile ranks, and at the same time against a specific growth goal set for the student. It thus draws on both frames of reference at once.

□ An example of the difference

Saad, a fourth grader, obtained a percentile rank of 50 on a standardized reading test, placing him exactly in the middle of his peers. This is norm-referenced information, and it does not point to a problem. On a criterion-referenced test, however, his teacher noticed that he could not distinguish between letters that share the same shape and differ only in their dots, such as ب (bā'), ت (tā') and ث (thā'). The first test does not reveal this, yet it is precisely what the teacher needs to know in order to decide what to teach tomorrow.

2.2 The mean and the standard deviation

Every standard score is built on two numbers: the **mean (Mean)**, the center of the norming sample's scores, and the **standard deviation (Standard deviation, SD)**, the degree to which scores are spread around that center. A small standard deviation means people's scores lie close together; a large one means they are widely dispersed.

To make the idea of standard deviation concrete, imagine two classes of twenty students each, both with an average height of 140 centimeters. In the first class, heights range from 135 to 145; in the second, from 120 to 160. The mean is the same, but the spread differs. A student who is 150 centimeters tall would be considered very tall in the first class and fairly ordinary in the second. It is the standard deviation that tells us so.

This gives us the rule on which diagnosis rests: we measure how far a student lies from the mean in standard deviations rather than in points. A distance of two standard deviations from the mean signifies the same thing on any test, however its scores are calculated.

2.3 The normal curve and standard scores

When we measure a trait such as intelligence or height in a large number of people, the scores usually fall into a bell shape: most people cluster near the mean, and their numbers dwindle the further we move from it in either direction. This shape is the **normal curve** (**Normal curve**), and among its properties are fixed proportions that do not change:

- About 68% of people fall between one standard deviation below the mean and one standard deviation above it.
- About 95% fall between two standard deviations below the mean and two above it.
- Only about 2% fall more than two standard deviations below the mean, and roughly the same proportion more than two standard deviations above it.

This last point is the origin of thresholds that recur throughout this book. The criterion of “two standard deviations below the mean” is used for intellectual disability, and “two standard deviations above the mean” appears in many definitions of giftedness. By this statistical criterion alone, then, each of the two groups represents about 2% of people at either end of the distribution.

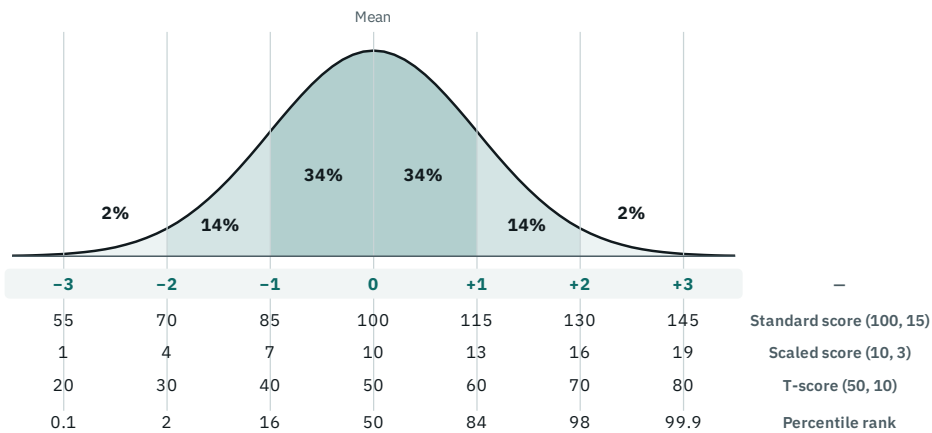


Figure 2.1: The normal curve, with the common score types aligned beneath it by position. Each column represents the same position expressed in different metrics: a full-scale score of 70 on an intelligence test is equivalent to a subtest score of 4 and to a T-score of 30, and all three mean two standard deviations below the mean.

Types of standard scores

Tests convert the raw score, the number of correct answers, into a standard score with a fixed mean and standard deviation. The benefit is that standard scores can be compared across ages and tests. The most widely used types are:

Table 2.1: Common types of standard scores.

Type	Mean	—	Where you will encounter it
z-score (z)	0	1	In calculations and research; it is the basis from which all other types are derived
Standard score	100	15	Composite and index scores on intelligence and achievement tests and adaptive behavior scales
Scaled score	10	3	Subtests of intelligence scales, and the subscales of the GARS-3
T-score (T)	50	10	Behavioral and emotional rating scales, including many measures of attention-deficit/hyperactivity disorder (Attention-Deficit/Hyperactivity Disorder, ADHD)

To convert any score to another type, we first convert it to a z-score and then from the z-score to the desired type:

$$z = (X - M) \div SD$$

z-score = (student's score – mean) ÷ standard deviation

Worked example

- 1 Hind obtained a standard score of 82 on an intelligence test with a mean of 100 and a standard deviation of 15.
- 2 The z-score: $(82 - 100) \div 15 = -1.2$, meaning she is one and one-fifth standard deviations below the mean.
- 3 Expressed as a T-score, it would be: $50 + (-1.2 \times 10) = 38$.
- 4 Her percentile rank is about 12, meaning her performance exceeds that of about 12% of her peers in the norming sample.

 Watch the direction of the score

On ability tests, a higher score means better performance. On many behavioral rating scales, however, a higher score means more problems. A T-score of 70 is excellent on a reading test and worrying on a hyperactivity scale. Always check the test manual or the report for what a high score means.

Descriptive classifications

Tests translate the standard score into a verbal description that is easy to understand. The following example shows the classifications used by the Wechsler Intelligence Scale for Children, Fifth Edition (**Wechsler Intelligence Scale for Children, Fifth Edition, WISC-V**), together with the official English label for each category, since specialists frequently read reports written with these labels (Wechsler, 2014):

Table 2.2: Descriptive classifications of composite scores on the Wechsler Intelligence Scale for Children, Fifth Edition.

Standard score range	Descriptive classification	Approximate percentage of people
130 and above	Extremely high	2%
120–129	Very high	7%
110–119	High average	16%
90–109	Average	50%
80–89	Low average	16%
70–79	Very low	7%
69 and below	Extremely low	2%

Three cautions apply to this table. First, the labels differ across tests, and even across editions of the same test. The fourth edition of this very scale called the 70–79 range “borderline” (**Borderline**) and anything above 130 “very superior” (**Very Superior**). So if you read the phrase **Very High** in a foreign report describing a score of 125, bear in mind that a higher category sits above it. Second, the “average” category here (90–109) is a descriptive classification specific to the Wechsler scales; it is narrower than the 85–115 range, which spans one standard deviation on either side of the mean and is the range used in this book’s figures. Third, and most important, a descriptive classification is not a diagnosis. The label “extremely low” describes only where the score falls; a judgment of intellectual disability requires other criteria, such as adaptive behavior, that is, the everyday living skills the student actually uses at home and at school (Chapter 7 addresses this in detail).

2.4 Percentile ranks and their pitfalls

The percentile rank is, on the surface, the easiest score for families to understand, and it is also the one most prone to misunderstanding. It means **the percentage of peers in the norming sample who scored lower than the student**. A percentile rank of 30 means the student outperformed about 30% of his peers, and that about 70% of them outperformed him.

The first pitfall: confusing it with percentage correct

The percentile rank has nothing to do with the percentage of correct answers. A student who answered 80% of the questions may have a percentile rank of 20 if the test was easy and most of his peers answered more than that. This is the trap Umm Majed fell into when she thought her son had “passed with three percent.”

The second pitfall: the distances between ranks are unequal

Because most people cluster around the mean, moving a few percentile ranks near the mean takes only a small difference in performance, whereas moving the same number of ranks at the extremes takes a large one. Compare these two moves, each spanning a similar number of ranks:

From rank 50 to rank 60

4 standard score points

From rank 90 to rank 99

16 standard score points

Figure 2.2: What each move is worth on a standard score scale with a mean of 100 and a standard deviation of 15. Ten ranks near the mean equal about 4 points, whereas nine ranks at the extreme equal about 16 points.

For this reason, percentile ranks should not be subtracted from one another or averaged, nor used to measure a student's progress. If a student moves from the 5th to the 10th percentile, his progress may be far greater than the two numbers suggest.

The third pitfall: age and grade equivalents

Age equivalents (**Age equivalent**) and grade equivalents (**Grade equivalent**) express performance in terms of an age or a school grade. A teacher may read in a report that “the student's reading age equivalent is 7 years, 3 months,” while the student is actually ten, and conclude that he reads like a seven-year-old. That reading is inaccurate. An age equivalent means only that the student's raw score equals the average raw score of children of that age; it does not mean that his skills or errors resemble theirs. These scores are also unevenly distributed, and they do not show whether the gap is large relative to the student's peers. This is why the Standards for Educational and Psychological Testing urge great caution in their use (AERA et al., 2014), and why standard scores and percentile ranks are preferred.

2.5

Reliability/precision and the standard error of measurement

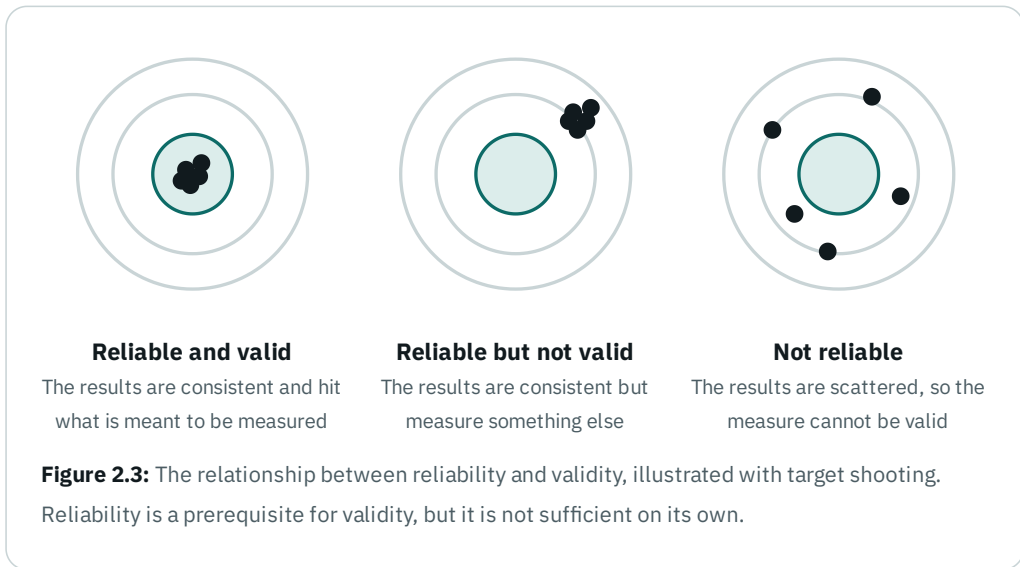
Reliability (Reliability) is the consistency of a measure's results. If the test were given to the same student twice under similar conditions, would he obtain two similar scores? A scale that shows a different weight each time is useless, however accurate it may be on average. The Standards for Educational and Psychological Testing use the compound term "reliability/precision" (**Reliability/precision**) to cover traditional reliability coefficients as well as every other index of score precision, such as the standard error of measurement and the information function, which we present at the end of this section (AERA et al., 2014).

Types of reliability

Table 2.3: Types of reliability and the question each one answers.

Type	Question it answers	How it is calculated
Test-retest reliability	Does the score remain stable over time?	The test is given twice with a short interval between administrations, and the correlation between the two is calculated
Internal consistency	Do the test items measure the same thing?	Calculated from the item intercorrelations in a single administration; the best known index is Cronbach's alpha (Cronbach's alpha)
Interrater reliability	Do two different raters assign the same score?	Two people rate the same student and their ratings are compared; this is especially important for observation and rating scales
Alternate-forms reliability	Do two forms of the test yield similar results?	Two equivalent forms are administered, and the correlation between them is calculated

Reliability is expressed as a coefficient ranging from zero to one. Salvia and colleagues (Salvia et al., 2017) proposed widely used benchmarks: at least 0.60 when scores are used to describe groups, at least 0.80 for screening decisions, and at least 0.90 for important decisions about an individual, such as diagnosis and classification.



No measure is perfectly reliable. A student's score on any given day is affected by fatigue, attention, guessing and the conditions in the room. The score we obtain is therefore an **estimate** of the true score, the score the student would have obtained had he been measured without any error. The standard error of measurement (**Standard Error of Measurement, SEM**) is a number that tells us how much the student's score would be expected to fluctuate if we measured him many times over:

$$\text{SEM} = \text{SD} \times \sqrt{1 - r}$$

standard error of measurement = standard deviation $\times$ square root of (1 – reliability coefficient)

From it we build a **confidence interval (Confidence interval)**: a range of scores within which we can say, with a given degree of confidence, that the true score lies. For a 95% confidence interval, we multiply the standard error by approximately 1.96, then add the result to the score and subtract it from the score.

Worked example: Majed's score

- 1 Majed obtained a full-scale score of 72 on an intelligence test with a standard deviation of 15 and a reliability coefficient of 0.95.
- 2 The standard error of measurement: $15 \times \sqrt{1 - 0.95} = 15 \times 0.224 \approx 3.4$
- 3 The 95% confidence interval: $72 \pm (1.96 \times 3.4) = 72 \pm 6.6$, that is, from about 65 to about 79.
- 4 The result: we can be 95% confident that Majed's true score on this test lies between 65 and 79. This range **straddles the cut-off score of 70**.

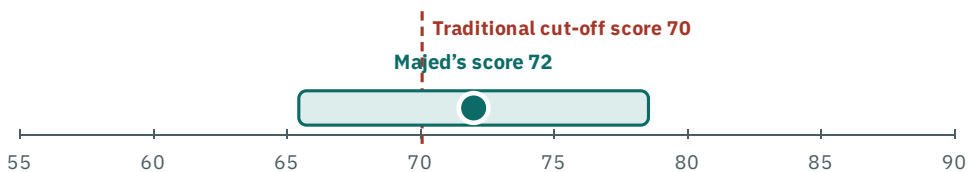


Figure 2.4: Majed's score and the confidence interval around it. The point lies above the cut-off score, but the band extends to either side of it, so we cannot say on the strength of the number alone that he "exceeded" it or "fell short" of it.

A **cut-off score (Cut-off score)** is the score that separates two categories in a decision, such as "refer" versus "do not refer" on a screening tool, or "meets the criterion" versus "does not meet it" for a diagnostic criterion. We use this term throughout the book.

It is precisely for this reason that the American Association on Intellectual and Developmental Disabilities (**American Association on Intellectual and Developmental Disabilities, AAIDD**) calls for the standard error of measurement to be taken into account when determining significant limitations in intellectual functioning, so that 70 is not treated as a hard dividing line (Schalock et al., 2021). A score of 72 or 74 does not automatically rule out intellectual disability, and a score of 68 does not automatically establish it. What settles the matter is the convergence of evidence, especially evidence of adaptive behavior.

 The rule to take with you

Every score in a diagnostic report is a point inside a band. The less reliable the measure, the wider the band. And when the band falls on both sides of the cut-off score, the decision becomes a professional judgment that rests on other evidence besides the number.

✓ **For the specialist: Why are confidence intervals in reports not symmetrical around the score?**

The way we calculated Majed's interval is a simplified method that centers the interval on the observed score itself. Many modern tests, however, including the Wechsler Intelligence Scale for Children, center the interval on the **estimated true score (Estimated true score)** and use the **standard error of estimate (Standard error of estimation, SEE)** to do so (Sattler, 2018):

$$T' = M + r (X - M)$$

$$SEE = SD \times \sqrt{r (1 - r)}$$

estimated true score = mean + reliability × (observed score – mean); standard error of estimate = standard deviation × square root of [reliability × (1 – reliability)]

Because reliability is less than one, the estimated true score always lies closer to the mean than the observed score does. Applied to Majed, his estimated true score would be $100 + 0.95 \times (72 - 100) \approx 73.4$ and the standard error of estimate about 3.3, so the interval would run from about 67 to about 80. The interval is thus skewed toward the mean rather than symmetrical around 72, yet it straddles the cut-off score of 70 under both methods. This is why readers see intervals in report tables such as “84: 78–92” and take them for a typo, when they are nothing of the kind (see Table 2.8). When you read a report, check the test manual for how its intervals were constructed before comparing them with the intervals of another test.

 In the Saudi context: Cut-off scores in the regulations

The Special Education Organizational Guide stipulates that among the conditions for a student's eligibility for intellectual disability programs is "that the student's IQ not exceed 75 on the Wechsler test, or 69 on the Stanford-Binet (Fifth Edition), or the equivalent of either," together with limitations in at least two areas of adaptive skills. It also requires "not relying on the results of a single scale" and that instruments be administered and interpreted "by qualified specialists" (Ministry of Education, 1436 AH [Arabic]). The earlier Regulations had defined the "educable" range as an IQ between 55 and 75 on the Wechsler or between 52 and 73 on the Stanford-Binet (Ministry of Education [Al-Ma'arif], 1422 AH [Arabic]). The different figures in the Regulations can be understood in light of the different standard deviation of the older editions of the Stanford-Binet (16 rather than 15); both fall at roughly the same distance from the mean. In the current Guide, however, the current editions of the two scales share the same standard deviation (15), and we found no rationale in the text for the two different figures (75 and 69). Nor did we find any provision directing that the standard error of measurement or the confidence interval be taken into account when these figures are applied.

A question for reflection: A student scores 77 on the Wechsler scale, with a confidence interval of 72–83, and shows clear limitations in three adaptive domains. How should the team handle the cut-off score of 75 in light of what you have learned in this section?

2.5.1 Differences between scores

Many diagnostic decisions rest on a **difference** between two scores: between ability and achievement in the discrepancy model, between two indexes on an intelligence test, or between a student's scores before and after an intervention.

Each of the two scores carries its own error, and the two errors combine in the difference. A difference is therefore less precise than either score, and it must be relatively large before we can be confident that it is real.

When we find a difference between two scores, we ask two distinct questions. The first concerns **statistical significance**: is the difference too large to be attributed to measurement error alone? The second concerns **rarity**, or the **base rate (Base rate)**: what proportion of the norming sample itself shows a difference of this size or larger? A difference may be significant, that is, real rather than a product of error, and yet common among typically developing children, in which case it cannot be treated as a sign of a disorder. Good test manuals present the two in separate tables.

✓ **For the specialist: The standard error of the difference**

When the two scores share the same standard deviation, the **standard error of the difference** (Standard error of the difference, SE_{diff}) is calculated as follows:

$$SE_{\text{diff}} = SD \times \sqrt{2 - r_1 - r_2}$$

standard error of the difference = standard deviation $\times$ square root of (2 – reliability of the first score – reliability of the second score)

Worked example: A difference between two indexes (from Table 2.8)

- 1 The Verbal Comprehension Index is 84, with a reliability of 0.94, and the Visual Spatial Index is 102, with a reliability of 0.92. The difference is 18 points.
- 2 The standard error of the difference: $15 \times \sqrt{2 - 0.94 - 0.92} = 15 \times \sqrt{0.14} \approx 15 \times 0.37 \approx 5.6$
- 3 The critical difference at the 0.05 level of significance: $1.96 \times 5.6 \approx 11$. The observed difference (18) exceeds it, so it is **statistically significant**.
- 4 Next we ask about rarity. Suppose, purely for illustration, that the base rate table in the manual shows that about 15% of children in the norming sample have a difference of this size or larger between these two indexes. The difference is real, but it appears in roughly one child in seven, so it is **not rare** and cannot by itself serve as evidence of a disorder.
- 5 Compare this with the difference between Working Memory (76, reliability 0.93) and Processing Speed (80, reliability 0.90): the standard error of the difference is $15 \times \sqrt{0.17} \approx 6.2$, the critical difference about 12, and the observed difference only 4. We cannot, then, say that processing speed is “stronger” than working memory.

The reliability of the difference score itself has a formula that shows the problem from another angle: $r_{\text{diff}} = [(r_1 + r_2) \div 2 - r_{12}] \div (1 - r_{12})$, where r_{12} is the correlation between the two scores. If the two indexes had reliabilities of 0.94 and 0.92 and correlated at 0.60, the reliability of the difference would be $(0.93 - 0.60) \div 0.40 \approx 0.83$, lower than that of either index. The higher the correlation between two scores, the lower the reliability of the difference between them. This is the root of the criticism leveled at decisions based on differences alone, as in the ability–achievement discrepancy model and in patterns of strengths and weaknesses models (to which Chapter 6 returns).

2.5.2 Regression toward the mean

When people who obtained extreme scores are measured again, their average score the second time tends to move closer to the population mean. This statistical phenomenon is called **regression toward the mean (Regression toward the mean)**. It occurs because an extreme score usually combines genuinely high (or low) ability with measurement error that happened to fall in the same direction, and that error does not necessarily recur the second time. The estimated true score we met in the previous “For the specialist” box is a direct application of this idea.

The phenomenon has practical consequences in two places. The first is giftedness: if we select students who scored above 130 on a test with a reliability of 0.90, the expected score of a student who obtained 135, on a retest with the same instrument, is about $100 + 0.90 \times 35 \approx 132$, and it drops further if the second test is a different one that correlates with the first at only about 0.80 (to about 128). A decline in the scores of some selected students at reevaluation is therefore to be expected; it means neither that the original decision was wrong nor that the student has “lost” his giftedness (Lohman & Korb, 2006). The second is disability and learning disabilities: some of those selected for their

very low scores will score higher when remeasured even without any intervention, so this gain cannot be credited entirely to the program in the absence of a comparison group.

2.5.3 Floor, ceiling and item gradient

A test's **floor (Floor)** is its capacity to discriminate among very low levels of performance, and its **ceiling (Ceiling)** is its capacity to discriminate among very high levels. If a test has few easy items, a child with very low performance and a child with low performance will both receive the same minimum score, and the test can no longer tell them apart. Likewise, if there are few difficult items, high performers bunch up at the highest possible score. On many individual intelligence tests, the full-scale score range bottoms out at about 40 and tops out at about 160. That suffices for most children, but not in two situations: distinguishing levels of severity among the more severe cases of intellectual disability, to which Chapter 7 returns, and distinguishing among exceptionally gifted students, to which Chapter 12 returns. For this reason, some publishers have issued **extended norms (Extended norms)** for both tails of the distribution.

Closely related is the **item gradient (Item gradient)**, the amount by which the standard score changes when the raw score increases by one point. If one additional correct answer moves a child from a scaled score of 4 to 7, the test is coarse in that range, and a small scoring error or a momentary lapse can change the whole classification. Bracken proposed minimum standards of technical adequacy that have since become widespread in the literature, among them that a raw score of 1 on a subtest should place the child more than two standard deviations below the mean if the floor is to be adequate, and that a one-point increase in raw score should not produce a change of more than a third of a standard deviation (Bracken, 1987). The problem is most pronounced with young children and with children at the youngest or oldest ages a test covers. If a child is at the edge of a test's age range, examine the norm tables before trusting the score.

2.5.4 Item response theory

Everything we have presented so far belongs to **classical test theory**, which assumes a single standard error for all scores on a test. Most modern tests, however, were built, wholly or in part, within another framework: **item response theory (Item Response Theory, IRT)**. Teachers need only two ideas from it: that a test's precision varies across ability levels and is highest where its items are concentrated; and that this is what makes computerized adaptive testing possible, in which the computer selects for each student the items suited to his level.

✓ **For the specialist: The item characteristic curve and the information function**

The theory describes each item by an **item characteristic curve** (**Item characteristic curve**), which links the student's ability, denoted θ and expressed on a scale resembling the z-score, to the probability of answering the item correctly. The curve's most important features are two parameters: **difficulty** (b), the ability level at which the probability of a correct answer reaches 50%, and **discrimination** (a), the slope of the curve at that point. A third parameter for guessing may be added for multiple-choice questions (Embretson & Reise, 2000).

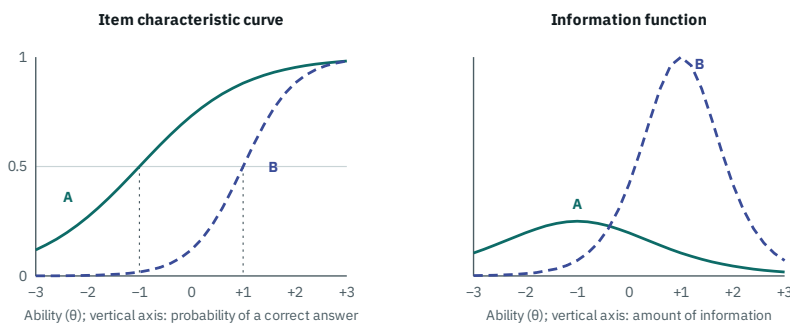


Figure 2.5: Two hypothetical items. Item “A” (solid line) is easy and moderately discriminating, and item “B” (dashed line) is difficult and highly discriminating. In the left panel, the probability of a correct answer reaches 50% at $\theta = -1$ for item “A” and at $\theta = +1$ for item “B.” In the right panel, each item yields the most information near its difficulty, and the more discriminating item provides more information over a narrower range.

From the curve we derive the **information function** (**Information function**), which shows how much the item contributes to the precision of the ability estimate at each level. What all the items contribute is summed in the **test information function**, and the standard error of the ability estimate at any level equals the inverse of the square root of the information at that level. A test whose items are concentrated around the mean therefore measures

average students with high precision and the two tails with less, which is a more precise formulation of the floor and ceiling problem discussed above. A **computerized adaptive test (Computerized adaptive testing, CAT)** begins with an item of moderate difficulty, estimates the student's ability after each response, and selects the next item from the item bank so as to yield the most information at that estimate. It thus reaches acceptable precision with fewer items, and each student faces items close to his level. The validity of this approach, however, depends on calibrating the item bank on a sample resembling the students we are measuring, a requirement that brings us back to Section 2.7.

2.6 Validity

If reliability asks, “Does the scale measure consistently?”, validity asks, “Does it actually measure what it is meant to measure? And are the decisions we base on its scores sound?” In the modern view, validity is a property of the interpretation of a scale's scores for a particular purpose, and it changes with the purpose and with the person tested. A scale that is valid for screening three-year-olds may not be valid for diagnosing an adolescent. The Standards for Educational and Psychological Testing treat validity as a single, unified concept, sometimes called **construct validity**: the degree to which the evidence supports interpreting the score as representing the intended construct. The construct is what all the evidence, taken together, attests to. The Standards draw this evidence from **five sources** (AERA et al., 2014):

Table 2.4: The five sources of validity evidence in the Standards for Educational and Psychological Testing, with examples from adapting scales for Arabic-speaking settings.

Source of evidence	Question	Example
Content	Do the items cover the facets of the construct adequately, and do they suit the setting in which they are used?	An autism spectrum disorder scale whose items cover the two core domains in the diagnostic manual: difficulties in social communication and interaction, and restricted, repetitive behaviors and interests. For the Arabic version, local experts review each item: does it describe a behavior one would expect to see in a Saudi home or school?
Response processes	Do test takers or raters actually engage the processes the construct assumes when they respond?	Saudi children are asked to think aloud while solving the items of a reasoning test, or parents are asked what they understand by a statement on a rating scale. If it turns out that a child fails because he does not know a Standard Arabic word in the question, the item is measuring vocabulary, not reasoning
Internal structure	Do the items and dimensions relate to one another as the theory predicts?	Factor analysis of the Arabic version: does it retain the structure of the original? A factor analysis of the Arabic version of the Childhood Autism Rating Scale arrived at a three-factor model of 14 items (Alotaibi & Alotaibi, 2021), and an Omani study of the Adaptive Behavior Assessment System (ABAS-3) supported a single general-factor model among children with intellectual disability (Emam et al., 2020)

Source of evidence	Question	Example
Relations to other variables	Do scores correlate with what they are expected to correlate with, predict the criterion, and distinguish between groups?	The Saudi version of the Wechsler Preschool and Primary Scale of Intelligence, Third Edition, correlated with the Saudi version of the Stanford-Binet at 0.81 (Ejbara, 2024). In a validation of the Arabic version of the Modified Checklist for Autism in Toddlers across nine Arab countries, sensitivity reached 0.86 and specificity 0.80 (Seif Eldin et al., 2008)
Consequences	What are the effects, intended or unintended, of using the scores in decisions?	Does using a scale for nomination to gifted programs lead to lower-than-expected representation of students from villages or of a particular group? (Chapter 14 returns to this question)

The older division into “content validity,” “criterion validity” and “construct validity” still appears in many test manuals and Arabic textbooks. In that scheme, “criterion validity” corresponds to part of the fourth source, while “construct validity” corresponds to validity as a whole in the modern view. The second and third sources, however, are the most important questions we ask when evaluating an Arabic adaptation: Does the Arab child understand the item as intended? And has the adaptation retained the structure of the original?

Sensitivity and specificity, and why the value of a positive result changes

For screening and diagnostic tools, validity evidence based on relations to other variables is usually reported through two concepts: **sensitivity** (**Sensitivity**), the proportion of those who have the condition whom the tool correctly detects, and **specificity** (**Specificity**), the proportion of those who do

not have the condition whom the tool correctly rules out. Yet two other figures matter more to teachers and families. The first is the **positive predictive value (Positive predictive value)**: if a student's result is positive, how likely is it that he actually has the condition? The second is the **negative predictive value (Negative predictive value)**: if his result is negative, how likely is it that the condition is truly absent? This second figure is the most important question in screening, because a negative result means the student will not be followed up.

These two probabilities depend jointly on the tool and on how prevalent the condition is in the group being tested. Consider a hypothetical example with illustrative figures: a screening tool with 85% sensitivity and 95% specificity is administered once to 1,000 children in a universal screening where the condition has a prevalence of 2%, and once to 200 children referred because of concerns, among whom the prevalence is 30%:

Table 2.5: The same tool (sensitivity 85%, specificity 95%) in two groups with different prevalence. Hypothetical figures.

	Universal screening (1,000 children, prevalence 2%)	Referred group (200 children, prevalence 30%)
Have the condition	20	60
True positive (detected by the tool)	17	51
False negative (missed by the tool)	3	9
Do not have the condition	980	140
False positive	49	7
True negative	931	133

	Universal screening (1,000 children, prevalence 2%)	Referred group (200 children, prevalence 30%)
Positive predictive value	$17 \div 66 \approx 26\%$	$51 \div 58 \approx 88\%$
Negative predictive value	$931 \div 934 \approx 99.7\%$	$133 \div 142 \approx 94\%$

In universal screening, roughly three-quarters of the positive results are false, yet a negative result is highly trustworthy. In the referred group the picture is reversed: most positive results are correct, but about one in every sixteen children with a negative result actually has the condition. This is why a positive screening result is not a diagnosis, and why a negative result in a high-risk group is not sufficient reason to close the file.

2.6.1 Fairness and bias in measurement

The Standards for Educational and Psychological Testing make **fairness** (**Fairness**) a third foundation of measurement alongside validity and reliability/precision, devoting their third chapter, “Fairness in Testing,” to it (AERA et al., 2014). Fairness in this sense is at heart a question of validity: Does a score mean the same thing for a student from a different linguistic, cultural or geographic group, or for a student with a disability? A failure in this respect is called **measurement bias**, and it takes several forms:

Table 2.6: Forms of measurement bias and how they are examined.

Form	What it means	How is it examined?
Construct bias	The construct itself is not defined, or does not manifest, in the same way in the two groups, as when “appropriate social behavior” is measured by the standards of a single culture	Expert review from both cultures, and comparison of factor structures
Differential item functioning (Differential item functioning, DIF)	Two students of equal ability from different groups do not have the same probability of answering a particular item correctly	Comparing the two groups' performance on each item after matching them on total score or estimated ability; experts then review the items on which a difference appears
Measurement non-invariance	The relationship between the items and the construct differs across the two groups, so their scores cannot be compared directly	Sequential tests in multigroup confirmatory factor analysis: equivalence of structure, then of loadings, then of intercepts
Predictive bias	The same score predicts different later performance in the two groups	Comparing the regression lines of criterion on score in the two groups: a difference in slope or intercept is evidence of bias

A difference in group means may be real; the two groups may genuinely differ in the measured skill for reasons related to opportunities to learn, and the scale may be describing that difference accurately. Bias is when the same score means something different depending on the group. Nor is examining the instrument enough: fairness also includes making the construct accessible to every test taker, as when a student with low vision is given the test in large

print, provided this does not change what it measures. Fairness in the referral and identification stages, and indices of proportional representation, are taken up again in Chapter 14.

2.7 Norms, standardization and test adaptation

A standard score means “the student's position among his peers in the norming sample.” The **norming sample** (**Norm sample**) is the group to which the scale was administered to derive its norms. If the norming sample does not resemble the student, the score loses its meaning. We therefore ask three questions of any instrument before using it:

1. **Who is in the norming sample?** Does it represent the children we are assessing in age, environment and language? A test normed on American children compares a Saudi child with American children rather than with peers in his own environment.
2. **When was it normed?** Average performance on intelligence tests has risen across generations in many countries, a phenomenon known as the Flynn effect (Flynn, 1984). A meta-analysis, that is, a study that statistically combines the results of many studies, estimated it at about two to three standard-score points per decade (Trahan et al., 2014). The phenomenon, however, is neither universal nor constant in direction; a slowdown or even a reversal has been documented in some countries in recent decades (Bratsberg & Rogeberg, 2018). Where the effect exists, old norms inflate the scores of today's children, masking an existing disability at one end and swelling the number identified as gifted at the other. The practical lesson is the same in both cases: know the year in which the norms were collected, and treat scores near cut-off points with caution when the norms are old.
3. **In what language and culture?** Translation alone is not enough. A question about “snow” or “Thanksgiving” is familiar to a child in one setting and strange to a child in another.

From translation to standardization

To simplify the path an instrument travels from one culture to another, this book distinguishes three levels. The division is the author's own teaching simplification, built on what international guidelines require (see the next box):

Table 2.7: Three levels in transferring an instrument from one culture to another (the author's simplification).

Level	What happens	What can be done with the scores
Translation	Items are rendered into Arabic literally or nearly so	The original norms cannot properly be used to judge the child
Test adaptation	Items and examples are modified to suit the culture, with back-translation (translating the Arabic version back into the original language and comparing it with the original), expert review, item analysis, and verification of measurement invariance across the two versions and of the absence of differential item functioning (see 2.6.1)	Better than translation, but the original norms remain unsuitable
Local standardization	The adapted version is administered to a representative sample of the local population, new norms are derived, and reliability and validity are studied in that population	Standard scores become meaningful for children in that population

✍ **For the specialist: What do the International Test Commission guidelines actually say?**

The second edition of the International Test Commission Guidelines for Translating and Adapting Tests consists of eighteen guidelines grouped into six categories. Unlike the transfer levels in our table, they follow the stages of the work (International Test Commission, 2017):

1. **Precondition:** obtaining permission from the copyright holder, and verifying before starting that the construct being measured is meaningful in the new culture.
2. **Test development:** translating and adapting the items with a team qualified in both languages and in the subject matter, and trying out the items on small samples.
3. **Confirmation:** gathering empirical evidence of the equivalence of the two versions: measurement invariance, examination of differential item functioning, reliability, and validity in the new population.
4. **Administration:** controlling administration instructions and conditions so as to reduce sources of bias in the new population.
5. **Score scales and interpretation:** scores are compared across the two populations only to the extent that the established level of equivalence allows.
6. **Documentation:** documenting every change made to the original, and providing guidance for those who use the new version.

Local standardization in our table thus corresponds roughly to the completion of the “confirmation” and “score scales” stages, together with the derivation of new norms. For a version that has not gone through the confirmation stage, a well-crafted translation is not enough.

📍 **In the Saudi context: Where do the norms of instruments in use stand?**

The instruments used in the Kingdom vary widely in where they stand among the three levels. Among the published examples we found:

- **Wechsler Intelligence Scale for Children:** The Saudi standardization cited in the literature is that of the revised edition (**WISC-R**) for ages 6–16, carried out within the gifted identification and care project at King Abdulaziz City for Science and Technology (Al-Nafi et al., 2000 [Arabic], as cited in Ejbara, 2024). We found no published Saudi standardization of the fourth or fifth edition at the time of writing. In light of the foregoing, this means that the norms of the locally standardized edition are old, while the norms of the current edition are not local.
- **Wechsler Preschool and Primary Scale of Intelligence, Third Edition:** A Saudi version administered to 788 children aged 4 years to 7 years 3 months in the Jeddah education region between 2018 and 2022, derived from the Jordanian version (Ejbara, 2024). The sample is large, but it comes from a single region.
- **Stanford-Binet Intelligence Scales, Fifth Edition:** Two Gulf versions are in use, compared in a study in the Northern Borders region (Mohammed, 2016 [Arabic]); we found no details of either norming sample in a published source.
- **Gilliam Autism Rating Scale, Third Edition:** An Arabic version studied with 178 participants aged 3 to 22 years from educational and clinical settings in the Kingdom, with a Cronbach's alpha of 0.97 (Alasmari et al., 2024). This is a study of the version's properties; national norms are a later step.

A question for reflection: Which of these examples reaches the level of “local standardization” in Table 2.7? And what does each one lack to be suitable for a student in your city?

❓ **Common misconceptions**

- **“An accurately translated version carries the validity, reliability and norms of the original.”** Validity, reliability and norms are properties of using scores in a particular population; they do not travel with the translation. New evidence from the new population is essential.
- **“As long as the reliability coefficient is high, the scale is valid.”** Reliability is a precondition for validity, not evidence of it (see Figure 2.3). A scale can consistently measure something other than what we intend.
- **“Any large difference between two scores indicates a problem.”** The difference may not be statistically significant, or it may be significant yet common in the norming sample (see 2.5.1).

⚠ Before you trust an Arabic version of an instrument

- Is the version licensed by the original publisher?
- Has a published adaptation study been conducted, or was it merely translated? Were measurement invariance and differential item functioning examined?
- Does it have local norms? When were they derived? From which regions was the sample drawn?
- Have its reliability coefficients and validity evidence in the local setting been published, and from which of the five sources does this evidence come?

If the answer to most of these questions is no, use the scale's results with great caution, confine them to description rather than diagnosis, and state this explicitly in the report.

2.8 Reading a realistic score table

Intelligence test reports usually present a table summarizing the composite scores. Table 2.8 presents hypothetical values for a ten-year-old student, modeled on the indexes of the Wechsler Intelligence Scale for Children, Fifth Edition:

Table 2.8: Hypothetical scores for a ten-year-old student, modeled on the indexes of the Wechsler Intelligence Scale for Children, Fifth Edition.

Index	Standard score	Percentile rank	95% confidence interval	Descriptive classification
Verbal Comprehension	84	14	78–92	Low average
Visual Spatial	102	55	94–110	Average

Index	Standard score	Percentile rank	95% confidence interval	Descriptive classification
Fluid Reasoning	98	45	91–105	Average
Working Memory	76	5	70–85	Very low
Processing Speed	80	9	73–91	Low average
—	86	18	81–92	Low average

The confidence intervals in the table are built around the estimated true score, as in the original scale, which is why they are not symmetrical around the score (see the “For the specialist” box in 2.5). We assumed the following reliability coefficients: 0.94 for Verbal Comprehension and Fluid Reasoning, 0.92 for Visual Spatial, 0.93 for Working Memory, 0.90 for Processing Speed, and 0.96 for the Full Scale IQ (**Full Scale IQ, FSIQ**). How, then, should a teacher read this table?

- 1. Look first at the full scale score and the limits of what it tells you.** The full scale score is 86, with an interval of 81–92, in the low average range. It is the most reliable score in the table and the strongest predictor of achievement.
- 2. Then examine the differences for significance and rarity before interpreting the pattern.** The gap between the highest and lowest index is 26 points. The questions that concern us are: Is it statistically significant? And is it rare in the norming sample? We saw in 2.5.1 that the difference between Verbal Comprehension and Visual Spatial is significant but not rare, and that the difference between Working Memory and Processing Speed is not significant at all. Only if the differences prove both significant and rare does it make sense to speak of a “pattern,” at which point the full scale score alone no longer describes the student adequately.
- 3. Note the strengths.** His reasoning and visual thinking are in the average range. This is important instructional information: new material can be

presented visually, and teachers can draw on his ability to make inferences.

4. Note the weaknesses and what they might mean in the classroom.

Working memory and processing speed are low. He may lose track of long instructions, need more time to finish tasks, and appear inattentive when in fact he cannot keep up with the pace of the class. These are hypotheses awaiting verification through classroom observation.

5. Read the confidence intervals. The working memory interval runs from 70 to 85, so describing it as “very low” in the scale's classification is correct for the observed score, but the interval extends into the low average range, so the degree of weakness is not settled.

6. Ask: What does this mean for diagnosis? This table alone diagnoses nothing. It does, however, call for considering the possibility of learning disabilities or attention-deficit/hyperactivity disorder, and for gathering further evidence from achievement, observation and measures of attention.

✓ **For the specialist: The scholarly debate over profile analysis**

School psychology has an ongoing debate about the value of **profile analysis**, that is, interpreting the pattern of strengths and weaknesses across indexes. On one side, critics argue that the indexes add little to the full scale score in predicting achievement, that the reliability of differences between them is much lower than that of the full scale score, and that patterns have not proven useful in selecting a particular instructional intervention (McGill et al., 2018). On the other side, its defenders argue that disciplined interpretation of the indexes, grounded in differences that are significant and rare and supported by evidence from outside the test, helps in understanding the student and in guiding subsequent assessment, particularly with twice-exceptional learners, whose uneven abilities the full scale score may conceal (Kaufman et al., 2016). The position this book takes lies between the two: begin with the full scale score and the limits of what it tells you, move to interpreting the pattern only after the differences have been examined statistically, and treat the pattern as a hypothesis requiring further evidence. Chapter 6 returns to this debate in discussing patterns of strengths and weaknesses models.

2.9 Ethics and professional responsibility

Those who assess children answer to two kinds of references that should not be confused. The first is a **technical reference**, chief among them the Standards for Educational and Psychological Testing, issued by three American associations (AERA et al., 2014). The Standards specify the validity, reliability/precision and fairness that tests and their use should meet, and devote their third chapter to fairness in testing. In international practice they are complemented by the International Test Commission's guidelines on test use, which set out the competencies needed by test users (International Test

Commission, 2001). The second is the **codes of ethics** that professional bodies issue for their members, among them the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct (American Psychological Association, 2017), the National Association of School Psychologists' professional standards, which include its principles for professional ethics (National Association of School Psychologists, 2020), and the Council for Exceptional Children's ethics and professional practice standards (Council for Exceptional Children, 2015). Among the principles on which these references agree, several bear directly on the teacher's work:

- **Informed consent:** The family knows why the assessment is being conducted, what will be administered and who will see the results, and gives consent before it begins. In the Kingdom, the Special Education Organizational Guide requires the guardian's written consent after formal notification (Ministry of Education, 1436 AH [Arabic]), and the Regulations stipulate obtaining his permission “for the purpose of diagnosis” and explaining his child's condition to him “in understandable language” (Ministry of Education [Al-Ma'arif], 1422 AH [Arabic]).
- **Confidentiality:** A student's results are not discussed in the staff room or sent in messaging groups, and only those who need them to serve the student have access to them. In the Kingdom, the Personal Data Protection Law, issued by Royal Decree No. (M/19) in 1443 AH, applies, and it classifies health data as sensitive data.
- **Test security:** Test pages are not photographed, and students are not coached on test items, since this undermines the validity of their results and those of others.
- **Appropriate use:** A scale is not used for a purpose it was not designed for, and no important decision rests on a single score. This is what the Organizational Guide stipulates in the phrase “not relying solely on the results of a single scale.”
- **Equity:** The assessor takes into account the student's language, culture and other disabilities; a student with low vision is not judged on intelligence by

a visual test (see 2.6.1).

- **Boundaries of competence:** The assessor does not administer a tool on which he has not been trained, nor interpret the results of a student from a linguistic or cultural background he does not know without consulting someone who does. Cultural competence is not acquired by reading a manual, and one of its boundaries is that the assessor knows when he needs a qualified interpreter or a colleague from the same community.
- **Conflicts of interest:** When the person who diagnoses is the same person who provides the paid service he recommends, or when a teacher's evaluation of a student is tied to incentives or to the number of students admitted to the teacher's program, judgment is affected, even if unintentionally. This should be disclosed, and an independent opinion sought where possible.

Among these principles lies a tension that recurs in practice: **the family's right of access** versus **test security**. The family has the right to know everything related to their child's assessment and program, and to request re-diagnosis if they believe he has not been placed in the appropriate program (Ministry of Education [Al-Ma'arif], 1422 AH [Arabic]). On the other hand, copying and handing over test items exposes the test to circulation and may violate the publisher's rights. The solution toward which professional codes lean is to explain the results to the family in detail and to show them their child's responses and how they were scored in the presence of the specialist, without handing over the test booklet or copying its items.

Qualification levels: Who administers what?

Many publishers classify their instruments into qualification levels that determine who may purchase and use them. The details vary from one publisher to another, but the general idea is the same: a first level for tools that teachers and others can administer after reading the manual; a second level requiring training in measurement or a professional license; and a third level

for complex tools such as individual intelligence tests and diagnostic instruments, which requires graduate study in a specialized field or an advanced professional license. The Wechsler Intelligence Scale for Children is an example of this last level. What the Saudi documents we reviewed stipulate is summarized in the following table:

Table 2.9: Who administers what in the Kingdom, according to the published documents we reviewed at the time of writing.

Role	What the documents stipulate	What we found no provision for
General education teacher	Member of the school's assessment team (Ministry of Education, 1436 AH [Arabic])	The tools he is entitled to administer. His practical role is referral, observation and informal tools
Special education teacher	Member of the assessment team (Ministry of Education [Al-Ma'arif], 1422 AH [Arabic]; Ministry of Education, 1436 AH [Arabic])	A list of the standardized scales he administers
Psychologist in schools and institutes	The Regulations, under the title “behavioral training teacher (psychologist),” assign him assessment and diagnosis using standardized intelligence scales, adaptive behavior scales and informal methods, as well as re-diagnosis when needed (Ministry of Education [Al-Ma'arif], 1422 AH [Arabic]). The Guide requires that tools be administered and interpreted by “qualified specialists” and that they be “approved by the Ministry of Education” (Ministry of Education, 1436 AH [Arabic])	A published list of approved scales, and the qualification criteria required for each scale

Role	What the documents stipulate	What we found no provision for
Psychologist in the health sector	Classified and licensed by the Saudi Commission for Health Specialties. In the edition of the classification manual we reviewed, a bachelor's degree holder is classified as an “assistant specialist” working under supervision, and is classified as a “psychology specialist” after no less than three years of clinical experience (Saudi Commission for Health Specialties, 2009 [Arabic]). The psychology licensing examination assesses, among other things, the application of knowledge in psychological assessment and diagnosis (Saudi Commission for Health Specialties, 2024 [Arabic])	The current edition of the classification manual
Physician	Participates in the school's assessment team “if available” (Ministry of Education, 1436 AH [Arabic])	A general provision in education regulations specifying the diagnoses reserved for physicians alone

We found no published provision in the Kingdom corresponding to the qualification levels publishers apply to the purchase and administration of standardized psychological tests. Until such a provision is issued, the practical reference remains the publisher's own requirements, together with the principle of boundaries of competence described above.

📍 **In the Saudi context: Two ethical dilemmas**

The first dilemma: consent by silence. The Organizational Guide states that if the guardian does not respond within a period not exceeding fourteen days from the date of written notification, “this shall be considered permission and consent” (Ministry of Education, 1436 AH [Arabic]). The provision protects the student's right not to have his assessment delayed because a notice was ignored. Yet the principle of informed consent assumes that the family has understood what it is consenting to. What should the team do if it learns that the father cannot read, or that the notice was sent home with the student and never arrived? The provision permits the team to proceed; ethics requires attempting to reach the family by other means, and documenting that attempt, before settling for silence.

The second dilemma: the mother who asks for the items. A mother asks for a copy of the booklet of the intelligence test administered to her son, because she wants to “train him on the questions he got wrong” before his reassessment next year. The mother has a right to access and understanding, and the test has a security that protects the validity of her own son's results next time. How would you word a reply that respects both rights?

 For the teacher: Back to Umm Majed

The teacher can now answer Umm Majed's two questions along these lines: "The number 3 is not a passing grade. It means that Majed's performance on this test is higher than that of about 3% of his peers of the same age. In other words, he needs clear support. As for your question about disability, the test alone cannot answer it. His score is close to the score used in making the decision, and the band around it extends on both sides, so the team will gather information about his everyday skills at home and at school before reaching any judgment."

Conclusion

What we have presented in this chapter rests on a single unifying idea: a score is an approximate estimate of the student's position among his peers. Standard scores of every kind are different languages for one position on the normal curve; a percentile rank is the proportion of peers the student outperformed; every score is surrounded by error whose extent the confidence interval shows; and the difference between two scores is less precise than either one. It follows that a decision in which the confidence interval crosses the cut-off score is a professional decision requiring other evidence; that a difference between two scores supports interpretation only if it is both significant and rare; and that an extreme score can be expected to regress toward the mean on retesting.

Precision alone is not enough. Whether a score suits its purpose is a question of validity, whose evidence is gathered from multiple sources, and whether it suits the student is a question of norms and fairness: norms are appropriate only if their sample resembles the student and is recent enough, and an accurate translation carries over neither the validity nor the norms of the original. Whenever an assessor uses a translated tool or norms from another

setting, one question remains that no calculation can settle: Does the norming sample resemble the student in front of me? The answer depends on the technical evidence available and on the limitations the report states openly.

2.10**Summary, key terms and activities**

Chapter summary

- Norm-referenced measures compare the student with peers and serve diagnosis; criterion-referenced measures compare the student with a mastery standard and serve instruction; and some tools, such as curriculum-based measurement, are interpreted both ways.
- All standard scores are different languages for one position on the normal curve, and they can be converted into one another via the z-score. Descriptive labels vary across tests and their editions.
- Two standard deviations below or above the mean mark off about 2% of people at each tail, and many diagnostic cut-off scores derive from them.
- A percentile rank is the proportion of peers the student outperformed, not the proportion of correct answers, and the distances between ranks are unequal.
- Every score is an estimate surrounded by error, and the confidence interval shows its extent. When the interval crosses the cut-off score, the decision becomes a professional one resting on other evidence.
- The difference between two scores is less precise than either one, and two questions must be asked of it: Is it statistically significant? And is it rare in the norming sample?
- Extreme scores regress toward the mean on retesting, and a test's precision weakens near its floor and ceiling.

- Validity is a single unified concept whose evidence is gathered from five sources, and fairness is a third foundation alongside it and reliability/precision. The predictive value of a result changes as prevalence changes.
- Norms are appropriate only if their sample resembles the student and is recent enough, and an accurate translation carries over neither the validity nor the norms of the original.
- Testing standards are a technical document, the codes of ethics of professional bodies are a separate reference, and local regulations add what is binding on the assessor in his own country.

Glossary of terms

Norm sample	The group to which a scale was administered to derive its norms, and against which each student's score is later compared.
Raw score	The number of correct answers or the total points before conversion; it has no meaning on its own.
Standard score	A converted score with a fixed mean and standard deviation, the best known having a mean of 100 and a standard deviation of 15.
Percentile rank	The percentage of the student's peers in the norming sample (of the same age or grade) who obtained a lower score; a percentile rank of 84 means the student's score is higher than those of 84% of them. It is not the percentage of correct answers.

Standard deviation	A measure of the spread of scores around their mean.
Cut-off score	The score that separates two categories in a given decision, such as referral versus no referral, or meeting versus not meeting a criterion.
Reliability/precision	The consistency of a scale's results across time, items or raters, and the precision of the score in general.
Standard Error of Measurement, SEM	An estimate of how much an individual's score fluctuates around his true score owing to imperfect reliability.
Confidence interval	A range of scores within which the true score is expected, with a specified probability, to fall.
Standard error of the difference	The amount of error expected in the difference between two scores, from which the smallest statistically significant difference is calculated.
Base rate	The proportion of the norming sample showing a particular difference or pattern, which indicates whether the difference is rare or common.
Regression toward the mean	The tendency of extreme scores to move closer to the mean on retesting.
Floor and ceiling	A test's capacity to discriminate among very low and very high levels of performance.

Item Response Theory, IRT	A measurement framework that describes each item by a curve linking ability to the probability of a correct response, in which measurement precision varies with ability level.
Validity	The degree to which evidence supports interpreting a scale's scores for a particular purpose; its evidence is gathered from five sources.
Differential item functioning, DIF	A difference between two groups of equal ability in the probability of answering an item correctly.
Measurement invariance	A scale's measuring the same construct in the same way in two groups, so that their scores can legitimately be compared.
Flynn effect	Change in average performance on intelligence tests across generations, which was upward for most of the twentieth century; where it occurs, old norms inflate scores.
Norm-referenced test	A test in which a student's score is interpreted by comparison with the scores of peers in the norming sample.
Criterion-referenced test	A test in which a student's score is interpreted by comparison with a predetermined level of performance or mastery.
Age equivalent	The age at which children's average raw score equals the student's raw score; it does not mean the student's skills resemble those of children of that age.

Sensitivity	The proportion of those who have the condition whom the tool correctly detects.
Specificity	The proportion of those who do not have the condition whom the tool correctly rules out.
Positive predictive value	The probability that the condition is actually present in someone with a positive result; it is affected by the prevalence of the condition in the group tested.
Negative predictive value	The probability that the condition is actually absent in someone with a negative result.
Test adaptation	Modifying a scale's items and examples to suit a new culture, together with scientific study of the result; it is broader than translation and narrower than local standardization. It differs from “accommodation” in the sense of supports provided to a student during testing.

Review questions

1. Classify each of the following results as norm-referenced or criterion-referenced, and justify your answer: (a) “Read 42 words correctly per minute; his end-of-term goal is 50.” (b) “His percentile rank is 30 on a standardized reading test.” (c) “Mastered 7 of 10 skills on an addition skills checklist.” (Objective 1)
2. Convert a T-score of 35 into a standard score with a mean of 100 and a standard deviation of 15, and give its descriptive classification in Table 2.2. (Objective 2)

3. A student obtained a percentile rank of 98 on a mathematics test. What does this mean? What does it not mean? And why is it incorrect to say that a student who moved from rank 45 to 55 progressed as much as a student who moved from 85 to 95? (Objective 3)
4. Two attention scales: the first has a reliability of 0.70 and the second 0.92. Which would you prefer for a diagnostic decision, and why? Calculate the standard error of measurement for each if the standard deviation is 10. (Objective 4)
5. In Table 2.8, calculate the standard error of the difference between Fluid Reasoning (reliability 0.94) and Working Memory (reliability 0.93), then decide: Is the difference between them significant at the 0.05 level? And what question remains before you interpret it? (Objective 4)
6. A screening tool with 90% sensitivity and 90% specificity is administered to 400 children among whom the condition has a prevalence of 5%. Calculate the positive and negative predictive values, and explain to a family what a positive result means for their child. (Objective 5)
7. An Arabic version of a behavior rating scale is being used at a diagnostic center with its American norms. For each of the five sources of validity evidence, write one question you would put to whoever prepared the version, and state the potential risks for a Saudi student. (Objective 6)
8. A colleague posts in the teachers' messaging group a photo of the summary page of a student's diagnostic report, asking what one of the scores means. Which principles does this action touch on? And how would you respond to him? (Objective 7)

Case for application

Noura is thirteen and in the first year of intermediate school. Her teacher referred her because of marked underachievement over several years. She was given an Arabic-adapted individual intelligence test whose norms were collected about eighteen years ago on a sample from a region other than her own. Her full scale score was 74, with a full scale reliability coefficient of 0.96.

Her Verbal Comprehension index was 86 (reliability 0.94) and her reasoning index 72 (reliability 0.90). An adaptive behavior form completed by the teacher alone indicates a deficit in one domain. One team member holds that “the score is below 75, so the student is eligible for the intellectual disability program.”

Calculate the confidence interval for the full-scale score and the standard error of the difference between the two indexes. Then reach a judgment: is the evidence sufficient for an eligibility decision? Support your judgment with at least three points you have learned in this chapter, and name the missing evidence you would request before deciding. (Objectives 4, 6, and 7)

Learning activities

- 1. A score ruler:** On a sheet of paper, draw a normal curve and place on it the scores of five hypothetical students expressed in different metrics: standard scores, T-scores, scaled scores, and percentile ranks. Then order them from lowest to highest. *Product:* one drawing with a conversion table. *Time:* about 30 minutes. (Objectives 2 and 3)
- 2. Translating for the family:** Rewrite Table 2.8 as a five-sentence paragraph that a family with no training in measurement could follow, mentioning the confidence interval and what the table does not mean. If you wish to practice on a real report, do so only with the consent of the family and the school administration, after removing everything that could identify the student, under the supervision of a specialist, and without using the product of the exercise in any decision. *Product:* one paragraph. *Time:* about 20 minutes. (Objectives 3, 4, and 7)
- 3. Examining a test manual:** Choose an instrument whose technical manual you can access, or a published study of an Arabic version of an instrument, and extract from it: a description of the norming sample, the year the norms were collected, the reliability coefficients, the validity evidence classified by the five sources, and any examination of differential item

functioning or measurement invariance. Then judge: is it suitable for a student in your city? *Product*: a one-page table. *Time*: about two hours. (Objective 6)

Chapter references

- Mohammed, H. A. A. (2016). [Differences between the two Gulf versions of the Stanford-Binet Intelligence Scales, Fifth Edition, in assessing and diagnosing intelligence and mental abilities among persons with disabilities in the Northern Borders Region] [In Arabic]. *Jazan University Journal of Human Sciences*, 5(Suppl.), 236–254.
<https://search.mandumah.com/Record/1103082>
- Personal Data Protection Law, Royal Decree No. (M/19) (9/2/1443 AH), as amended by Royal Decree No. (M/148) (5/9/1444 AH) [In Arabic].
- Saudi Commission for Health Specialties. (2009). [*Professional classification manual for health practitioners*] [In Arabic] (5th ed.).
https://cams.ksu.edu.sa/sites/cams.ksu.edu.sa/files/attach/Professional_Registration_Classification_Manual_vr.6.pdf
- Saudi Commission for Health Specialties. (2024). *Saudi Psychology Licensure Examination (SPSLE): Applicant guide*. <https://scfhs.org.sa/sites/default/files/2024-03/SPSLE%20Applicant%20Guide%202024.pdf>
- Ministry of Education. (1436 AH). [*Special education organizational guide*] [In Arabic] (1st ed., academic year 1436–1437 AH). <https://moe.gov.sa/en/aboutus/aboutministry/RPRLibrary/الدليل%20التنظيمي%20للتربية%20الخاصة.pdf>
- Ministry of Education [Al-Ma'arif]. (1422 AH). [*Regulations for Special Education Institutes and Programs*] [In Arabic]. <https://moe.gov.sa/en/aboutus/aboutministry/RPRLibrary/القواعد%20التنظيمية%20لمعاهد%20وبرامج%20التربية%20الخاصة%201422هـ.pdf>
- Alasmari, M., Alduais, A., Qasem, F., Almaghlouth, S., & AlAmri, L. (2024). Predicting language, cognition, and social skills in individuals with autism spectrum disorder using the Arabic validation of Gilliam Autism Rating Scale–Third Edition. *Applied Neuropsychology: Child*. Advance online publication. <https://doi.org/10.1080/21622965.2024.2417812>

- Alotaibi, B. M., & Alotaibi, A. A. (2021). Exploratory and confirmatory factor analyses of the Arabic version of the Childhood Autism Rating Scale. *Research in Autism Spectrum Disorders*, 86, Article 101827. <https://doi.org/10.1016/j.rasd.2021.101827>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct* (2002, amended effective June 1, 2010, and January 1, 2017). <https://www.apa.org/ethics/code>
- Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment*, 5(4), 313–326.
- Bratsberg, B., & Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. *Proceedings of the National Academy of Sciences*, 115(26), 6674–6678.
- Council for Exceptional Children. (2015). *What every special educator must know: Professional ethics and standards* (7th ed.). CEC.
- Ejbara, M. A. (2024). Psychometric characteristics of a Saudi version of Wechsler Intelligence Scale for Preschool and Young Children. *Dirasat: Educational Sciences*, 51(2), 35–52. <https://doi.org/10.35516/edu.v51i2.6661>
- Emam, M. M., Al-Sulaimani, H., Omara, E., & Al-Nabhany, R. (2020). Assessment of adaptive behaviour in children with intellectual disability in Oman: An examination of ABAS-3 factor structure and validation in the Arab context. *International Journal of Developmental Disabilities*, 66(4), 317–326. <https://doi.org/10.1080/20473869.2019.1587939>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.
- Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, 95(1), 29–51.
- International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, 1(2), 93–114.
- International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). <https://www.intestcom.org>
- Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (2016). *Intelligent testing with the WISC-V*. Wiley.

- Lohman, D. F., & Korb, K. A. (2006). Gifted today but not tomorrow? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted*, 29(4), 451–484.
- McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2018). Cognitive profile analysis in school psychology: History, issues, and continued concerns. *Journal of School Psychology*, 71, 108–121.
- National Association of School Psychologists. (2020). *The professional standards of the National Association of School Psychologists*. NASP.
- Salvia, J., Ysseldyke, J. E., & Witmer, S. (2017). *Assessment in special and inclusive education* (13th ed.). Cengage Learning.
- Sattler, J. M. (2018). *Assessment of children: Cognitive foundations and applications* (6th ed.). Jerome M. Sattler, Publisher.
- Schalock, R. L., Luckasson, R., & Tassé, M. J. (2021). *Intellectual disability: Definition, diagnosis, classification, and systems of supports* (12th ed.). AAIDD.
- Seif Eldin, A., Habib, D., Noufal, A., Farrag, S., Bazaid, K., Al-Sharbati, M., Badr, H., Moussa, S., Essali, A., & Gaddour, N. (2008). Use of M-CHAT for a multinational screening of young children with autism in the Arab countries. *International Review of Psychiatry*, 20(3), 281–289. <https://doi.org/10.1080/09540260801990324>
- Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn effect: A meta-analysis. *Psychological Bulletin*, 140(5), 1332–1360.
- Wechsler, D. (2014). *Wechsler Intelligence Scale for Children* (5th ed.). Pearson.

When is it sound to conclude that a student has a disability, a disorder, giftedness or twice-exceptionality, and on what grounds?

This book traces the stages of diagnosis and identification systematically, from referral and the gathering of evidence to the interpretation of that evidence and the decisions built on it. It goes beyond defining the categories and setting out their criteria to examine the nature of measurement instruments, the procedures for administering them, the qualifications required of those who administer them, the principles for interpreting their results, and the point at which the evidence is sufficient to support a judgment.

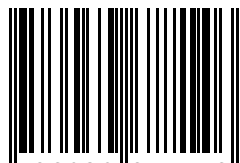
The book brings together two fields that the literature usually treats separately: disabilities and disorders, and giftedness and twice-exceptionality. It reads international standards and the American experience in light of published Saudi regulations, and describes the status of the Arabic version of each instrument it presents.

Its fifteen chapters address the foundations of measurement; formal, informal and alternative methods of gathering evidence; and identification models; then take up the categories of exceptionality one by one, supported by decision pathways, instrument profiles and applied cases. They close with fairness in diagnosis, report writing, and the development of the individualized education program.

Diagnosis and Identification of Exceptional Learners

Dr. Omar A. Alsamani

ISBN 978-603-06-7714-6



9 786030 677146