04.

Factors affecting test administration

1. **Relationship between examiner and test taker:**
   The way an examiner behaves and their relationship with the test taker can affect test scores. One study with 1st- through 7th-grade children used the Wechsler Intelligence Scale for Children (WISC) under two conditions. In the enhanced rapport condition, examiners were friendly and gave encouragement. In the neutral rapport condition, examiners were not friendly and gave no reinforcement. For younger children (through 3rd grade), the examiner's behavior had little effect. But for 5th- to 9th-grade students, those with friendly examiners scored higher (average IQ 122) than those with neutral examiners (average IQ 109), a difference of almost one standard deviation.
   Another study looked at examiner comments. Children whose examiners made disapproving comments scored lower than those with neutral or approving comments. Younger children also did better when tested by a familiar examiner.
2. **The race of the tester:**

   Some people worry that a tester's race might affect a child's test scores. Research shows that a tester's race usually does not affect IQ scores. IQ tests are highly standardized, and well-trained examiners should follow the same procedures. Differences in scores are more likely due to examiner attitude, nonverbal cues, or cultural differences.

   Small race effects have been found in some cases, especially when examiners are less trained or have more discretion. In one study, white examiners got higher scores from white children than African American children, but African American examiners got similar scores from both groups.
3. **Language of the test taker:**

   Tests can disadvantage non-English speakers, even if they don't require talking, because instructions assume English understanding. Translations may not be as valid or reliable as the original. Bilingual test takers should use their strongest language, and interpreters should be used cautiously to avoid bias.
4. **Training of test administrators:**

   Different tests require different levels of training. Some behavioral assessments need training but not a formal degree. Psychiatric diagnoses, like those using the Structured Clinical Interview for DSM-V (SCID), are usually done by licensed psychiatrists or psychologists who receive extra training on the test.
   Complex tests like the WAIS-R need practice—students usually need at least 10 administrations to score accurately.
5. **Expectancy effects:**
   Sometimes, an experimenter's expectations can influence the results of a study. This is called an expectancy effect or Rosenthal effect. In classic experiments, if students were told someone would do well, they scored them higher; if told they would do poorly, they scored them lower. The effect is usually small, but it can occur in both humans and animals.
   In intelligence testing, expectancy can affect scoring. For example, students scoring ambiguous IQ responses tended to give higher credit to responses they thought came from "bright" people
   Even though results are inconsistent, examiners should always be aware of their expectations and try to eliminate bias to maintain objectivity.
6. **Effects of Reinforcing Responses:**

   Incentives like money or tokens can improve performance for some children, but the effect varies by age, gender, income, and ethnicity. For example, praise can be as effective as money or candy, and process-focused praise ("you worked hard") works better than person-focused praise ("you are clever"). In surveys, approval from the interviewer can increase the number of reported responses

As different test takers respond differently to reinforcement, most test manuals prohibit giving feedback. Some exceptions exist, such as tests for the blind or the elderly, but even then, standardized methods are developed.

7. **Computer-Assisted Test Administration**

   Computer technology has greatly influenced testing, making it easier to administer, score, and standardize tests. Interactive testing uses computers to present test items, record responses automatically, and provide instructions when needed.

   Many major tests, including the SAT, GRE, and psychological assessments, are now computer-based. Studies show scores are similar between computer and paper tests, but computer testing reduces errors, allows better experimental control, and may make test takers more comfortable and honest, especially when answering sensitive questions.

   Advantages of computer-assisted testing include:

   - Excellent standardization and consistency
   - Tailored item presentation for each test taker
   - Precise timing of responses
   - Freeing human testers for other duties
   - Patience (test takers aren't rushed)
   - Reduced bias
   - Randomized item order for each test taker
   - Time limits for each item
   - Prevention of skipping ahead or revisiting completed sections

   Cautions and limitations:

   - Computer reports cannot replace a psychologist's judgment.
   - Software errors, unvalidated scoring routines, and outdated databases can cause problems.
   - Inexperienced users may misinterpret results.

8. **Mode of administration**

   The way a test or questionnaire is given—self-administered, interviewer-administered, or computer-based—can affect results:

   - **Interviewer-administered** measures often show people in better health than self-completed forms. Telephone interviews also tend to produce higher health scores than paper questionnaires.
   - **Computer-administered tests** usually give more accurate results. Men, in particular, may respond randomly on paper tests.
   - In **psychiatric assessments**, self-completed questionnaires may show more distress or disability, especially in younger people, compared to interviewer-administered questionnaires.
   - In some cases, such as reporting on cancer screening, the mode of administration has little impact.
   - In **educational testing**, studies show that computer-based and paper-based reading tests for K–12 students produce similar scores.

9. **Subject Variables**

   Test Anxiety: Many college students experience test anxiety, which makes it hard to focus on test items. They may get distracted by thoughts like "I am not doing well" or "I am running out of time." Test anxiety has three components:

1. **Worry** – concern about failing or poor performance
2. **Emotionality** – feelings of nervousness or tension
3. **Lack of self-confidence** – doubting one's abilities

Health: Physical health also impacts performance. Being sick with a cold or flu can reduce concentration and thinking. Many variations in health status affect both behavior and cognitive functioning, which is why medical drugs are often evaluated for their effect on cognitive processes.

Special Populations: Certain groups may need special testing arrangements. For example, elderly individuals often perform better in individual testing sessions rather than in groups, even for tests that are normally group-administered.

Behavioral Assessment

1. **Reactivity -** Reactivity happens when people change their behavior because they know they're being watched. For example, if a student knows a teacher is observing them, they might behave better than usual. This makes it hard to measure their *true* behavior. Observers try to reduce this by watching secretly or making people forget they're being observed.
2. **Drift -** Drift means that observers slowly stop following the exact rules they were trained with. Over time, they might rate or judge behaviors differently — not because the behavior changed, but because *their standards* did. To prevent this, observers are retrained and their work is checked regularly to stay consistent.
3. **Expectancies -** Expectancies happen when an observer's *expectations* affect what they see or record. For instance, if an observer expects a student to do well, they might (without realizing it) record their behavior more positively. This bias can distort results, so observers are trained to remain objective and unaware of expected outcomes.
4. **Deception -** Deception involves purposely hiding the real purpose of a test or study so participants act naturally. For example, if people know they're being tested for honesty, they might act more honest. Researchers sometimes use deception to get genuine behavior — though ethically, participants must be told afterward what the study was really about.

Necessity of Test Administration

Test administration is the process of **conducting a test under specific conditions**. It is necessary because:

1. **Ensures Standardization -** All test-takers should receive the same instructions, environment, and procedures. Without standardization, test results may not be comparable, reducing the reliability and validity of the test.
2. **Reduces Bias and Error -** Controlled administration prevents differences caused by reinforcement, feedback, or the examiner's behavior. For example, praising or rewarding some students but not others can unfairly influence scores.
3. **Maintains Reliability and Validity -** A test only measures what it claims to measure if it is administered properly. Deviation from the manual can reduce both the reliability (consistency) and validity (accuracy) of the test.
4. **Controls Situational Variables -** Conditions like timing, environment, instructions, and even the examiner's tone can influence test performance. Test administration provides guidelines to control these variables.
5. **Accommodates Special Populations -** Some groups (e.g., children, elderly, visually impaired) may need special but standardized methods of administration. This ensures fairness while still following controlled procedures.
6. **Facilitates Fair Scoring and Interpretation -** Proper administration minimizes errors in scoring and ensures that the results truly reflect the test-taker's abilities, not external factors.

Test Interpretation

Test interpretation or score interpretation is a process of analyzing scores in a test and translating qualitative data into quantitative and grading into numerical.

Two methods of test interpretation — Criterion-Referenced and Norm-Referenced.

Criterion-Referenced Interpretation

Compares a student's score to a **specific standard or criterion**. Shows whether the student **can do a particular task or knows a specific skill**, not how they compare to others.

Example: A driving test where you need 70% to pass.

Advantages:

- Clear goal: shows what the student can or cannot do.
- Useful for mastery learning and instruction.
- Helps in grading or certifying skills.
- Encourages cooperation rather than competition.

Disadvantages:

- Doesn't show how the student performs compared to peers.
- Hard to set an appropriate cut-off score.
- Can't easily show relative strengths or weaknesses in a group.

Norm-Referenced Interpretation

Compares a student's score to the **scores of a reference group (norm group)**. Shows whether the student is **above, below, or at the average**.

Example: IQ tests or standardized achievement tests.

Advantages:

- Shows relative performance compared to peers.
- Useful for ranking or selection.
- Can identify strengths and weaknesses in comparison to a group.
- Easy to apply on large groups (40+ students).

Disadvantages:

- Doesn't tell whether the student has mastered specific skills.
- Scores depend on the norm group, so changing the group changes interpretation.
- Can encourage competition rather than mastery.

Normal Distribution

A statistical model of a bell shaped curve used to describe how scores are spread out in a population.
Characteristics:

1. **Symmetrical Shape -** Left and right sides are like mirror images. The curve is even on both sides.
2. **Central Tendency -** The mean, median, and mode all have the same value, located at the center of the curve. This represents the average performance and helps us describe an entire set of scores with one representative number.
3. **Predictable Spread -** Consistent pattern of standard deviation.

   68% fall within 1 standard deviation of the mean. 95% fall within 2. 99% fall within 3

Score

A **test score** is a summary of a person's responses to test items that measure a specific skill or trait. It can also be the **total or combined result** from all items on a single test or a set of tests.

Types - Raw score, Scaled score, Standard score

Raw Score: Simply the score a student gets on a test before any conversion or comparison is made. For example if an exam has 100 questions, and someone get 80 of them correct, their raw score is 80. Raw score itself has no meaning, it doesn't describe anything. They aren't useful for comparing student's performance with others because raw scores are often uneven and not normally distributed. Teachers need to see how a student is doing compared to others, not just how many answers were right. Hence, for fair comparison raw scores are transformed into derived scores (scaled score & standard score)

Scaled Score/Criterion Referenced Scores

It is a mathematical transformation of raw score, A representation of the total number of correct questions a participant has answered that has been converted into a consistent scale. It helps make scores from different test forms comparable.

Advantages:

- Makes comparisons fair across test forms.
- Easier to interpret than raw scores alone.
- Can reflect different aspects of performance (speed, quality, precision).

Limitations:

- Requires statistical conversion, which may confuse some users.
- May not reflect all nuances of raw performance.
- Unique scales may not be directly comparable to other scores.

Types:

1. **Number or Percent Correct:** Shows the raw score as a number or percentage of correct answers. Example**:** You answered 45 out of 50 questions correctly → **90%**.
2. **Speed of Performance:** Measures how quickly a student completes the test or task. Example**:** A typing test where you type 60 words per minute.
3. **Quality of Performance:** Focuses on how accurate or effective the responses are, not just completion. Example**:** An essay graded for clear arguments, proper grammar, and evidence.
4. **Precision of Performance:** Measures how careful or exact the work is. Example**:** A math test where partial credit is given for carefully shown steps, not just the final answer.
5. **Unique Score Scales:** Scores that don't fit into the standard categories above, often special scales used for particular tests. Example**:** IQ score (mean = 100, SD = 15) or a personality score on a 1–10 scale.

Standard Score

A type of score that shows how far a person's performance is from the average of the norm group using standard deviation units. These are raw scores transformed to fit normal distribution - a distribution with fixed, known characteristics.

Advantages

1. **Equal Units Across the Scale** – interpretation is consistent and precise because each unit represents the same amount everywhere on the scale.

2. **Independent of Test Format** – scores can be compared fairly across different tests, regardless of the number of items or scoring method.
3. **Clear Interpretation** – shows whether a student is below average, average, or above average, and how far their score is from the mean.
4. **Enables Comparisons Across Subjects or Tests** – allows meaningful comparison even when tests differ in content or difficulty.

Disadvantages of Standardized Scores

1. **Requires Statistical Transformation** – raw scores must be converted, which may be confusing to some users.
2. **Depends on Norm Group** – meaningful interpretation requires an appropriate and representative norm group.
3. **May Overlook Specific Skills** – a standardized score shows relative standing but doesn't always reflect detailed mastery of particular content.

Types of Norm referenced Scores

1. **Percentile Rank** (এইটা বাদে বাকি ৩টা স্ট্যান্ডার্ড স্কোরেরও টাইপস)**:** A single number that indicates a student's position relative to the norm group. It answers the question to "what percent of people scored lower than this score". It ranges from 1 to 99.

$$\text{Percentile Rank} = \frac{\text{Number of students who scored less}}{\text{Total number of students}} \times 100$$

- A percentile rank is not equal to the percentage of items answered correctly. For example, if someone ranked 9th out of 25 students with a score of 34 out of 45.

  Percentage: $\frac{34}{45} * 100 = 76$

  Percentile Rank: $\frac{16}{25} * 100 = 64$; 64% people scored lower than them.

Limitations: Percentile ranks are ordinal hence unequal interval between ranks. Also it can't be used for mathematical calculations, they can't be added, subtracted, divided, or multiplied.

2. **Developmental & Growth Scales:**
   - **Grade Equivalent Scores-** Grade equivalent scores show what grade level a student's performance is similar to in the norm group. Scores are expressed in years and months. Left digit = grade level, right digit = month (0–9). Example: 4.2 means the student's performance is similar to the average performance of students in grade 4, second month.
   - **Age Equivalent Scores-** Very similar to grade equivalent scores, but instead of comparing students by grade level they compare students by age level. It shows the average test performance of students at a certain age. Scores are expressed similarly as grade equivalent scores except months are 0 to 11.
3. **Linear Standard Scores**
   - **Z-Score:**

     A Z score shows how far a raw score is from the mean, measured in standard deviation units. It is a type of norm referenced and linear standard score, and considered as the most basic standard score.

$$Z = \frac{X - \text{mean}}{\text{standard deviation}}$$

     A positive z score → the score is above the mean. A negative z score → the score is below the mean. Z score=0 → the score is exactly the average.

Limitations:

- **Negative Scores:** Half the students will have negative z-scores because they scored below the mean. Example: z = −2.50 may be confusing for students and parents.
- **Difficult Interpretation:** Understanding z-scores requires knowledge of the mean, standard deviation, and norm-referencing.
- **Motivation Impact:** Seeing a negative score on an achievement test could negatively affect a student's motivation.

- **McCall's T-score:** Another way of expressing a Z score. A type of standard score designed to avoid negative scores and fractional values that occur with z-scores. Instead of having a mean of 0 and a standard deviation of 1, T scores have a mean of 50 and a standard deviation of 10. This makes the numbers easier to work with and avoids negative values.

Formula: $T=(Z*10)+50$

4. **Normalized Standard Scores**

- **Stanines:** A system of transforming raw scores into standardized scores (just like Z scores, T scores, quartiles, or deciles). It was developed by the U.S. Air Force during World War II.
  Stanine, short for "standard nine," is used in norm-referenced tests to show performance in **score bands** rather than exact values. Scores range from 1 to 9, where 1–3 is below average, 4–6 is average, and 7–9 is above average. The mean is 5 and the standard deviation is 2.

  Stanines are easy to interpret and give a clear picture of relative performance, though they are less precise than percentile ranks.
- **SAT/GRE Scores:** The scores for the SAT and GRE (also called CEEB scores, after the College Entrance Examination Board) use a special type of scale, though they convey the same basic information as other standardized test scores. With a mean = 500, SD = 100.

Formula: $(Z*100)+500$

- **Normal Curve Equivalent Scores:** NCE scores are a type of standardized score that normalizes performance across a group. They have a mean of 50 and a standard deviation of 21.06. This unusual value ensures that NCE scores exactly match percentile ranks at three points: 1, 50, and 99.

  Key Advantage: Unlike percentile ranks, NCE scores represent equal units across the entire scale. This makes it easier to compare differences in performance at any point on the scale.
- **Deviation IQ scores:** Deviation IQ scores are a type of standardized score used mainly for tests of mental ability. They show a person's performance relative to the average in a normal distribution. With a mean of 100 and SD of 15 or 16, depending on the test.

Summary on Interpreting Norm-Referenced Scores

- All norm-referenced scores (percentile ranks, z-scores, T-scores, stanines, SAT/GRE scores, etc.) essentially provide the **same information**:
  - They show the how does someone compare with others.
  - The only difference is the scale used.
  - It doesn't matter which norm-referenced score you use — all convey relative standing within the norm group.
- Percentile ranks are unequal units. Rest of the scales have equal units.

Necessity of test interpretation

1. **Transforming Raw Scores Into Meaningful Information -** Raw scores only show correct answers. Interpretation converts them into scaled or standard scores, making them meaningful.

2. **Making Comparisons Across Tests and Forms -** Scaled and standard scores allow fair comparison across different test forms or subjects.
3. **Comparing Performance Across Students or Groups -** Norm-referenced scores allow comparison within a group, showing who is above, below, or at average.
4. **Highlighting Relative Strengths and Weaknesses -** Norm-referenced scores identify which subtests or skills a student excels in or struggles with, which raw scores alone cannot show
5. **Providing Equal and Interpretable Units -** Standardized scores (z, T, NCE, IQ) give equal units, ensuring consistent interpretation; percentile ranks alone can be misleading.
6. **Supporting Instructional and Educational Decisions -** Criterion- and norm-referenced scores help determine skill mastery, growth, and relative standing for instructional decisions.
7. **To Prevent Misuse or Misunderstanding of Scores -** Without proper interpretation, test results can be misused—for example, assuming a norm score sets a standard every student must reach, which is incorrect.