Essentials of Biostatistics

For Paramedical and Allied Health Sciences

Special Features

- First book on Biostatistics specifically designed for Paramedical and Allied Health Sciences
- **100+** illustrations including flowcharts, tables, and graphs
- **100+** Solved Examples
- **10+** Appendices containing information on the applicability of biostatistics in research
- Practical boxes with applicable statistical tips
- Exclusively included statistical tools and their applications in the biostatistics

CBS Publishers & Distributors Pvt. Ltd.

Anju Dhir

16

Scaling to Improve Reliability and Validity

"Good design is like a refrigerator—when it works, no one notices, but when it doesn't, it sure stinks."

—Irene Au

LEARNING OBJECTIVES

After the completion of the chapter, the readers will be able to:

- Understand scaling and application of its technique.
- Know about T-score in psychology and education.
- Interpret reliability of test scores.

CHAPTER OUTLINE

- Introduction
- Statistics as a Tool
- Applications of Statistical Knowledge
- Importance of Statistics for Students of Psychology and Education
- Scaling
- Measurement
- Standard Score and T-Score
- Reliability and Validity of Test Scores

INTRODUCTION

Knowledge of statistics is important for the students of psychology and education. Reasons are as follows:

- These courses deal with theories and research studies which are based on statistical analysis.
- Students have to undertake research where they have to handle, analyze and interpret data.

STATISTICS AS A TOOL

Statistics is a useful tool for the students. It is a means of communicating knowledge which is required to read and evaluate surveys, experiments, and other practical problems in the field of psychology and education. It is used in research starting with planning a study, analyzing the data, and interpreting the results. The courses on statistics are necessary in the first and second year of the courses of psychology and education. The background of high school statistics is nil for the majority of fresh students. There is no developed problem-solving and analyzing skills in these students. Analysis revealed that students have difficulties with:

- Graphical representation

- Working on methodology based on mathematics.
- Manipulation of summation symbol Z, etc.

Most students while dealing with statistics:

- Wait until somebody tells them how to tackle the problem.
- Prefer verbal expressions to mathematical ones.
- Memorize without trying to understand.

Although the availability of statistical computer packages has changed statistics courses both in contents and in methods used. Mathematical competence is not in focus while using these softwares.

APPLICATIONS OF STATISTICAL KNOWLEDGE

Competence in Reading and Evaluating Research

- **Competence in reading:**
  - Knowing and understanding the techniques used, their area of application, and their assumptions.
  - Knowing and understanding the decisions taken concerning methodological aspects.
- **Competence in evaluating:**
  - Competence in evaluating the decisions against other competing decisions.
  - Evaluating subprocedures used during the statistical techniques.
  - Evaluating the interpretation given by the researcher versus alternative interpretations.
  - Competence to characterize features of the study.

Competence in Doing Research

Because of the availability of statistical computer softwares, the use of complex statistical techniques is not difficult. Mathematical abilities are no longer core of the matter for using statistics. But these are of little help at the planning stage of the study, while choosing the appropriate technique, and in interpreting the results. A good statistical understanding is still required which includes a certain level of mathematical and problem-solving skills.

The task of research is to establish causal relationships and to explain this relationship. Researchers in the field of psychology and education have different backgrounds and rely in general on one model that is humanistic.

IMPORTANCE OF STATISTICS FOR STUDENTS OF PSYCHOLOGY AND EDUCATION

Statistics is a process to collect and analyze the data for interpreting the results. A student deals with many questions which need to be analyzed via statistics, e.g., researcher may ask the subjects (thousands) to rate their favorite movie from 1 to 5 on a rating scale.

Issues Faced in Attitude Measurement by Psychiatry Student

When a researcher is interested in measuring the attitudes, feelings or opinions of respondents he/she should be cleared about the following:

- What is to be measured?
- Who is to be measured?
- The choices available in data collection techniques.

He/she must know which type of variables will be used like nominal, or ordinal or interval or ratio level, etc. Therefore, statistics can help a student in following ways:

- **Organize the data in a better way:** In the field of psychology and education, student will come across a large data which is not easy to handle. But by using graphs and pie charts, etc., data can be presented in an easy way.
- **Describing data gets easy:** One can easily describe the data collected during its collection. In the field of psychology, it is called descriptive statistics. For example, how many men and women are employed in an area along with other problems can be described with statistics. Researcher can make conclusions based upon the data after its collection and treatment.
- **Makes teaching and learning easy:** Statistics is important in education and psychology because of the different types of problems that are come across. Data involved is always huge. Therefore, statistics helps in making it concise. Statistics enables us to study the scores of data, objectively.
- **Helps teacher to provide exact description:** For example, the marks obtained in a class test can indicate about the effect of teaching method. The scores obtained by students show their perception toward a topic. Therefore, the data can be accurately described with the help of statistics.
- **Makes the teacher exact and precise in procedures and thinking:** It may become vague to describe a student's performance if knowledge of the statistics is lacking. With statistics methods, it becomes easy.
- **Summarization of results becomes easy with statistics:** Statistics arranges the data in an ordered manner. The data can be expressed in an understandable and meaningful way.
- **Enables a user to draw general conclusions:** It not only enables a user to draw general conclusions but helps to extract the inferences, step by step.
- **Future performance can be predicted:** Statistics enables its users to predict what will be the outcome under specific situations. Hence, the decisions can be taken accordingly. Although some margin of error is always there.
- **Statistics enables to analyze some of the causal factors underlying complex and otherwise confusing events:** Behavioral outcome is result of causal factors. Therefore, the cause of performance of a student or behavior of a patient can be studied by keeping extraneous variables as constant.

SCALING

Scaling is used to:

- Assigning numbers or other symbols to the characteristics of objects according to the certain prespecified rules.
- The measurement of physical properties is not a complex deal, whereas measurement of psychological properties requires a careful attention of a researcher.

Basic Characteristics of Scales

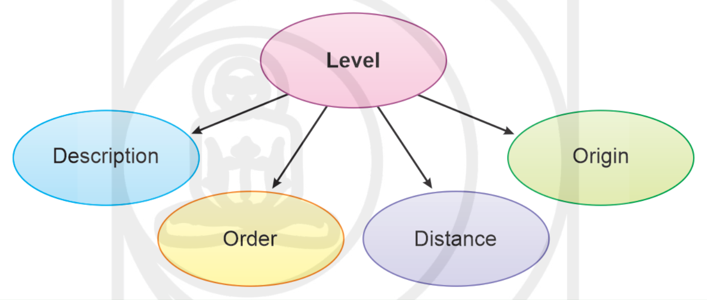
All scales used in scaling techniques can be explained in terms of four basic characteristics which are: description, order, distance, and origin**.** These characteristics collectively define the level of measurements of scale (Refer to Unit I). The level of measurement indicates that what properties of an object are measured or not measured by the scale.

Scaling is a procedure of measuring and assigning the objects to the numbers according to specified rules. In other words, the process of finding the measured objects in a range having a continuous sequence of numbers, to which the objects are assigned, is called scaling.

The data consists of quantitative variables, like height, weight, income, sales, etc., and qualitative variables like knowledge, performance, character, etc. For further analysis, the qualitative information must be converted into numerical form. This is possible through measurement and scaling techniques. Most commonly, feature of survey-based research is to have respondent's feelings, attitudes, opinions, etc. in some measurable form.

Practical Tips

Levels of measurement define the relationship among the values assigned to the characteristics of an object. We can say that what properties of an object, the scale is measuring or not measuring, is denoted by the levels of measurement.



1. **Description:** The description means a particular unique label and descriptor which is used to designate the values of the scale. For example, we have the descriptors 1. Male, 2. Female. Here, male and female are unique descriptors denoting values 1 and 2 on gender scale. All scales have unique descriptors or labels which are used to define the values of the scale and the response options.
2. **Order:** The order means the relative size and position of the descriptor. Here, the order is associated with only relative values and no absolute values. Thus, the order is denoted by descriptors like "less than", "greater than", "equal to". For example, optician's preference for three brands of lenses is shown in the order given below with the most preferred brand listed first and the least preferred on the last.
   - Essilor 360 DS
   - Bosch and Lomb
   - Softens 59

   This shows that the preference for Essilor 360 DS is greater than the preference for Bosch and Lomb and likewise, the preference for Softens 59 is less than the preference for Bosch and Lomb.

   It is important to note that all the scale does not possess order characteristic. Such as gender scale (1. Female, 2. Male) does not possess order as one cannot determine whether a female is greater than or less than male.
3. **Distance:** Distance means that the absolute differences between the descriptors on a scale are known and can be expressed in units. For example, a five-person room has one patient more than a four-person room and likewise a four-person room has one patient more than the three-person room.

   It is to be noted that, the scale that has the distance characteristic also has the order. As we know that five-person classroom is greater than the four-person classroom in terms of a number of persons in class. Thus, we can say that distance implies order, but the reverse is not true, i.e. order does not necessarily imply distance.

Contd...

4. **Origin:** The origin shows that scale has a unique or fixed starting or true zero point. A scale having origin characteristic also has the distance, order, and description. Many scales used in the marketing research do not have any fixed origin.
   For example, in case of unfavorable-favorable scale,
   1 = extremely unfavorable
   2 = unfavorable
   3 = neither unfavorable nor favorable
   4 = favorable
   5 = extremely favorable

   Here, 1 is an arbitrary origin or starting point. This scale could have started with 0 = extremely unfavorable and 4 = extremely favorable. Likewise, it can also be started with −2, where −2 = extremely unfavorable and 2 = extremely favorable. Thus, this scale does not have any fixed origin and hence does not possess the origin characteristic.

   You must have observed that description, order, distance, and origin depict successively higher level characteristics. Origin being the highest level characteristic while the description being the most basic characteristic. If the scale has order characteristic, then it will also have the description, and likewise the scale with distance characteristic has both the description and order. The scale with origin characteristic has all that is distance, order and description.

   It means that the higher level characteristics possess, the lower level characteristics, however, the lower level characteristics may not necessarily possess the higher-level characteristics.

Scaling—as the Extension of Measurement

Scaling is a process of placing respondents on a continuum or range with respect to their preference for the object. Generally, in research, the numbers are assigned to the qualitative traits of the object because the quantitative data helps in statistical analysis of the data and further facilitates the communication of the measurement rules and results.

The measurement is a process of assigning numbers or symbols to the characteristics of the object as per the specified rules. The researcher assigns numbers, not to the object, but to its characteristics like perceptions, attitudes, preferences, and other relevant traits.

Measurement is a process of observing and recording the observations that are collected as part of research. It may be in terms of numbers or other symbols to characteristics of objects according to certain prescribed rules. The respondent's, characteristics are feelings, attitudes, opinions, etc. The most important aspect of measurement is the specification of rules for assigning numbers to characteristics. The rules for assigning numbers are standardized and applied uniformly. They do not change over time or objects.

Scaling is the assignment of objects to numbers or semantics according to a rule. In scaling, the objects are text statements, usually statements of attitude, opinion, or feeling.

Applications of Scaling Technique

- Explain the concepts of measurement and scaling.
- Classify and discuss different scaling techniques.
- Discuss four levels of measurement scales.
- Select an appropriate attitude measurement scale for a research problem.

Types of Scaling Techniques

The scaling techniques vary and are used as per the requirement of a researcher (Fig. 16.1).

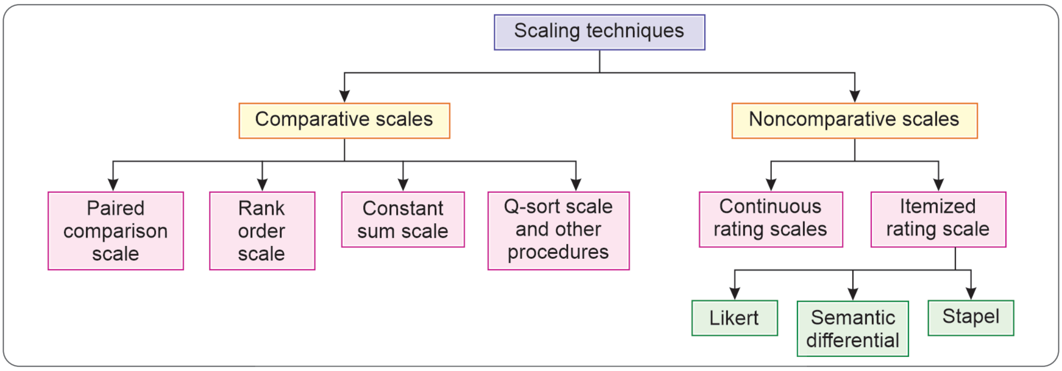


Figure 16.1: Types of scaling techniques

Comparative Scales

In comparative scaling, the respondent is asked to compare one object with another.

The comparative scales can further be divided into the following four types of scaling techniques:

Paired Comparison Scale

This is a comparative scaling technique in which a respondent is presented with two objects at a time and asked to select one object according to some criterion. The data obtained are ordinal in nature. For example, there are four opticians A, B, C, and D. The respondents can prefer A to B or B to C, etc., (Table 16.1).

TABLE 16.1: Preference of optician using paired comparison scale

Preference	Optician	A	B	C	D
A–B	A	–	#	–	–
A–C	B	–	–	–	–
A–D	C	#	#	–	–
B–C	D	#	#	#	
Total		**2**	**3**	**1**	**0**

Rank Order Scale

This is another type of comparative scaling technique in which respondents are presented with several items simultaneously and asked to rank them in the order of priority. This is an ordinal scale that describes the favored and unfavored objects, but does not reveal the distance between the objects. The resultant data in rank order is ordinal data. This yields better results when direct comparison is required between the given objects. The major disadvantage of this technique is that only ordinal data can be generated.

Example: Rank the following opticians in order of preference.

Solution: Begin by picking out the optician you like most and assign it a number 1, then find second most preferred one, and assign it number 2. Continue this procedure until all opticians are marked in order of preference. Least preferred should be assigned rank 4. Also remember that no two opticians receive the same rank (Table 16.2).

TABLE 16.2: Preference of optician using rank order

Optician	A	B	C	D
Rank	3	1	2	4

Constant Sum Scale

In this scale, the respondents are asked to allocate a constant sum of units like points or rupees among a set of stimulus objects with respect to some criterion. For example, researcher wants to determine how important the attributes of price, fragrance, packaging, cleaning power, and lather of a disinfectant are to consumers. Respondents might be asked to divide a constant sum to indicate the relative importance of the attributes. The advantage of this technique is that time is saved. However, main disadvantages are as follows:

- The respondents may allocate more or fewer points than those specified.
- The second problem is respondents might be confused.

Example: Between the attributes of detergent please allocate 100 points among the attributes so that your allocation reflects the relative importance you attach to each attribute.

Solution: The more points an attribute receives, the more important attribute is. If an attribute is not important, assign it zero point. If an attribute is important, it should receive twice as many points Table. 16.3.

TABLE 16.3: Importance of disinfectant attributes using a constant sum scale

Attribute	Price	Fragrance	Packaging	Cleaning power	Lather	**Total score**
Points	50	5	10	30	5	**100**

Q-sort Scale and other Procedures

This is a comparative scale that uses a rank order procedure to sort objects based on similarity with respect to some criterion. The important characteristic of this methodology is that it is more important to make comparisons among different responses of a respondent than the responses between different respondents. Therefore, it is a comparative method of scaling rather than an absolute rating scale. In this method, the respondent is given statements in large number for describing the characteristics of a product or a large number of brands of a product.

Example: The packet given to you contains pictures of 90 journals. Please choose 10 journals you 'Prefer most', 20 journals you 'Like', 30 journals you 'Dislike' and 10 journals you 'Prefer least'. Please list the scored journals names in the respective columns of the form provided to you (Table. 16.4).

Solution: The data is sorted according to instructions in the Table 16.4.

Noncomparative Scales

In noncomparative scaling, respondents need only to evaluate a single object. Their evaluation is independent of the other object or objects, which the researcher is studying. The noncomparative scaling techniques can be further divided into:

TABLE 16.4: Preference of journals according to Q-sort scale

Prefer most	Like	Neutral	Dislike	Prefer least
.... 10	 20	 30	 20	 10

Continuous Rating Scales

Continuous rating scale is very simple and highly useful. In continuous rating scale, the respondents rate the objects by placing a mark at the appropriate position on a continuous line that runs from one extreme of the criterion variable to the other (Fig. 16.2).

Example: How would you rate TV advertisement as a guide for buying a product?

Solution: Shown in Figure 16.2 according to continuous rating scale.

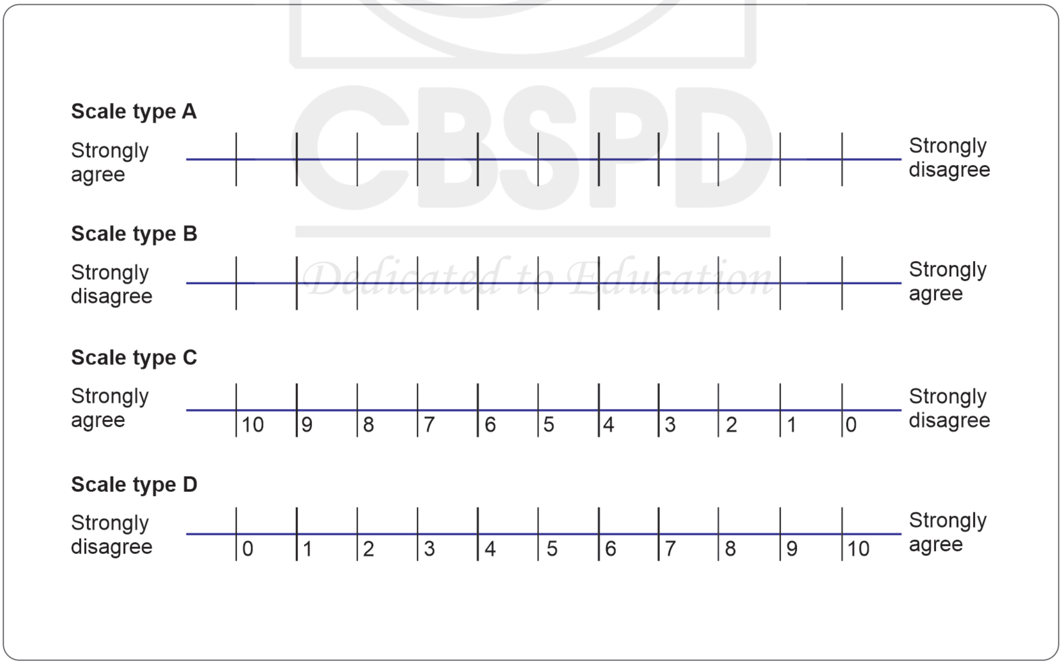


Figure 16.2: Continuous rating scale

Itemized Rating Scales

Itemized rating scale is a scale having numbers or brief descriptions associated with each category. The categories are ordered in terms of scale position and the respondents are required to select one of the limited number of categories that best describes the product, brand, company, or product attribute being rated. Itemized rating scales are widely used in marketing research. Itemized rating scales is further divided into three parts:

1. Likert scale
2. Semantic differential scale
3. Staple scale

The itemized rating scales can be graphic, verbal or numeric (Fig. 16.3):

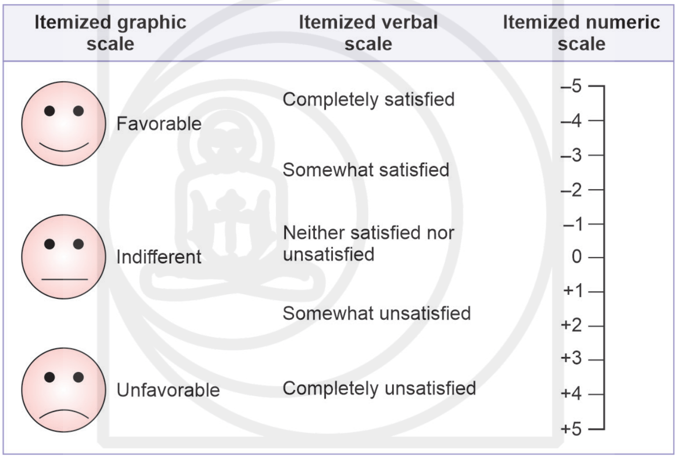


Figure 16.3: Types of itemized rating scales

Some common words for categories used in itemized rating scales are shown into Table 16.5:

TABLE 16.5: Common words for categories used in itemized rating scales

Quality: Excellent Very Good	Good Good	Not decided Neither good nor bad	Poor Fair	Worst Poor
Importance: Very Important	Fairly important	Neutral	Not so important	Not at all important
Interest: Very interested	Somewhat interested	Neither interested nor disinterested	Somewhat uninterested	Not very interested

Contd...

Satisfaction: Completely satisfied	Somewhat satisfied	Neither satisfied nor unsatisfied	Somewhat unsatisfied	Completely unsatisfied
Frequency: All of the time Very often	Very often Often	Often Sometimes	Sometimes Rarely	Hardly even Never
Truth: Very true	Somewhat true	Not very true	Not at all true	
Purchase Interest: Definitely will buy	Probably will buy	Probably will not buy	Definitely will not buy	
Level of Agreement: Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
Dependability: Completely dependable	Somewhat dependable	Not very dependable	Not at all dependable	
Style: Very stylish	Somewhat stylish	Not very stylish	Completely unstylish	
Cost: Extremely expensive	Expensive	Neither expensive nor inexpensive	Slightly inexpensive	Very inexpensive
Ease of use: Very case to use	Somewhat easy to use	Not very easy to use	Difficult to use	
Modernity: Very modern	Somewhat modern	Neither modern nor old-fashioned	Somewhat old fashioned	Very old fashioned
Alert: Very alert	Alert	Not alert	Not at all alert	

Likert Scale

Likert is extremely popular for measuring attitudes, because the method is simple to administer. With the Likert scale, the respondents indicate their own attitudes by checking how strongly they agree or disagree with carefully worded statements that range from very positive to very negative toward the attitudinal object. Respondents generally choose from five alternatives (like strongly agree, agree, neither agree nor disagree, disagree, strongly disagree). A Likert scale may include a number of items or statements.

Disadvantage of Likert Scale: It takes longer time to complete than other itemized rating scales because respondents have to read each statement. Despite this disadvantage, this scale has several advantages. It is easy to construct, administer and use (Table 16.6).

TABLE 16.6: Example of Likert scale

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
If the price of raw materials fall, firms too should reduce the price of the food products	1	2	3	4	5
There should be uniform price through out the country for food products	1	2	3	4	5
The food companies should concentrate more on keeping hygiene while manufacturing food products	1	2	3	4	5
The expiry dates should be printed on the food products before they are delivered to consumers in the market	1	2	3	4	5
There should be government regulations on the firms in keeping acceptable quality and on the prices	1	2	3	4	5
Now-a-days most food companies are concerned only with profit making rather than taking care of quality	1	2	3	4	5

Semantic Differential Scale

This is a seven-point rating scale with end points associated with bipolar labels (such as good and bad, complex and simple) that have semantic meaning (Table 16.7).

- It has been widely used in comparing brands, products and company images.
- It has also been used to develop advertising and promotion strategies and in a new product development study.
- It can be used to find whether a respondent has a positive or negative attitude toward an object.

Example: A data of some examples semantic differential scales is given in Table 16.7. Show the trends followed by the respondents.

TABLE 16.7: Examples of semantic differential scales

Examples of semantic differential scales								
Modern	–	–	–	–	–	–	–	Old-fashioned
Good	–	–	–	–	–	–	–	Bad
Clean	–	–	–	–	–	–	–	Dirty
Important	–	–	–	–	–	–	–	Unimportant
Expensive	–	–	–	–	–	–	–	Inexpensive
Useful	–	–	–	–	–	–	–	Useless
Strong	–	–	–	–	–	–	–	Weak

Contd...

Examples of semantic differential scales								
Quick	–	–	–	–	–	–	–	Slow
	+3	+2	+1	0	–1	–2	–3	
Useful	–	–	–	–	–	–	–	Useless
Attractive	–	–	–	–	–	–	–	Unattractive
Passive	–	–	–	–	–	–	–	Active
Beneficial	–	–	–	–	–	–	–	Harmful
Interesting	–	–	–	–	–	–	–	Boring
Dull	–	–	–	–	–	–	–	Sharp
Pleasant	–	–	–	–	–	–	–	Unpleasant
Cold	–	–	–	–	–	–	–	Hot
Good	–	–	–	–	–	–	–	Bad
Likable	–	–	–	–	–	–	–	Unlikable

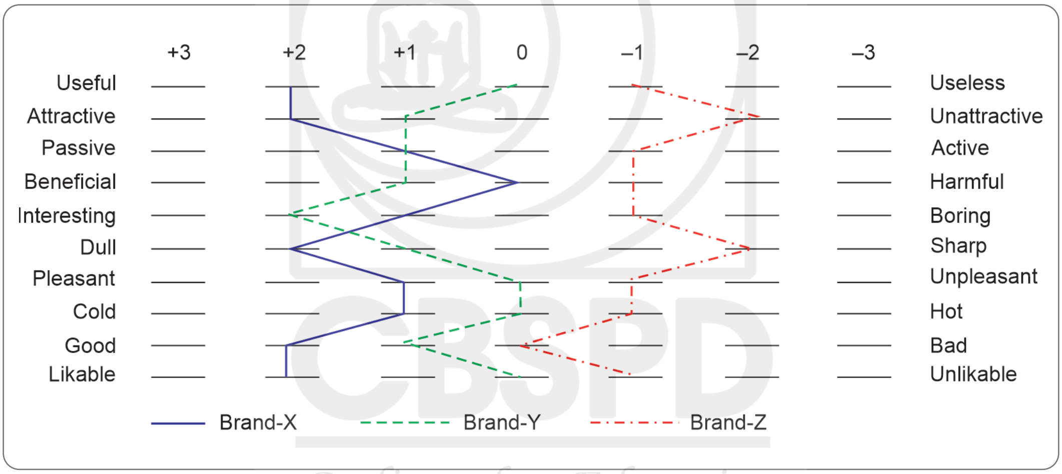


Figure 16.4: Trends followed by respondents

Solution: In example given in Table 16.7, the broken lines in the format (Fig. 16.4) clearly are showing the trends followed among the respondents toward an object.

Staple Scale

The staple scale (Fig. 16.5) was originally developed to measure the direction and intensity of an attitude simultaneously. Modern versions of the staple scale place a single adjective as a substitute for the semantic differential when it is difficult to create pairs of bipolar adjectives. The modified staple scale places a single adjective in the center of an even number of numerical values.

Example: Select a plus number for words that you think describe personnel banking of a bank accurately. The more accurately you think the word describes the bank, the larger the plus number you should choose. Select a minus number for words you think do not describe the bank accurately. The less accurately you think the word describes the bank, the larger the minus number you should choose.

	+4	+3	+2	+1	–1	–2	–3	–4
Fast service								
Friendly								
Honest								
Convenient location								
Convenient hours								
Dull								
Good services								
High saving rates								

Figure 16.5: Staple scale

Solution: The response (Fig. 16.6) by respondents is recorded as follows:

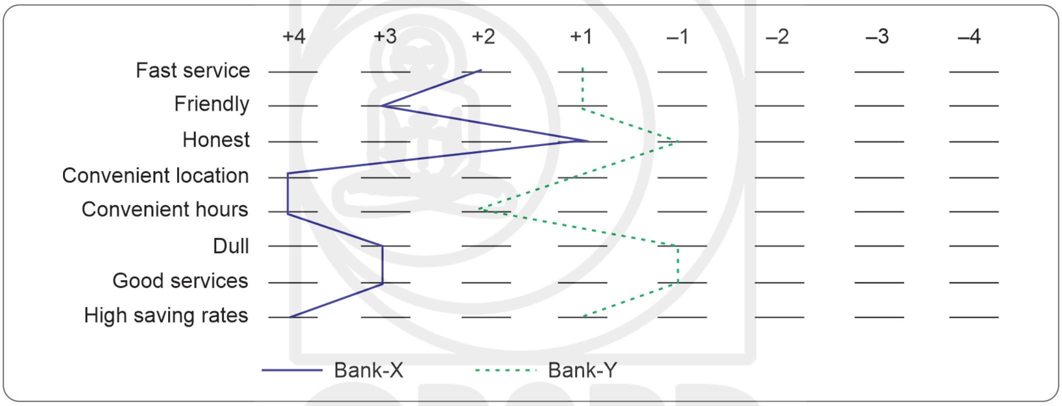


Figure 16.6: Response in staple scale

Practical Tips

Selection of an appropriate scaling technique:

A number of issues decide the choice of scaling technique. Some significant issues are as follows:

- Problem definition and statistical analysis
- The choice between comparative and noncomparative scales
- Type of category labels
- Number of categories
- Balanced versus unbalanced scale
- Forced versus nonforced categories

Therefore, one can say that number of scaling techniques are available for measurement of attitudes although there is no unique way to choose a particular scaling technique for research study.

MEASUREMENT

The level of measurement refers to the relationship among the values that are assigned to the attributes, feelings or opinions for a variable. *(Refer Unit 1, Chapter 3 for details on Levels of Measurements).*

STANDARD SCORE AND T SCORE

Z-score

Standard scores, or "Z-score" measures the relation between each score and its distribution.

Z-score is useful to simplify many problems. One use of Z-score is to convert any normal distribution to the standard normal distribution (also discussed in Unit IV, Chapter 9).

The equation for standard score calculation is:

$$Zx_i = \frac{X_i - \overline{X}}{s_x}$$

Example:

- Suppose the Mean is 100 and the standard deviation is 15.
- Suppose $X_i = 70$, find Z-score.
- Suppose $X_i = 115$, find Z-score of this value.

Solution:

To find a Z-score, subtract the mean and divide by the standard deviation. In this example, we subtract 100, and divide the difference by 15:

$$z = (70 - 100)/15 = -30/15 = -2$$

$$z = (115 - 100)/15 = 15/15 = 1$$

Example: We might know the Z-score and need to solve for the "raw" score; that is, we know *z* and we find *X* if the mean is 100 and S_X is 15, $z = 2$; find X_i.

Solution:

$$Z = \frac{X_i - Mean}{Sx}$$

$$2 = \frac{X_i - 100}{15}$$

$$30 = X_i - 100$$

Add 100 to both sides

$$100 + 30 = X_i - \cancel{100} + \cancel{100}$$

$$X_i = 130$$

Must Know

Properties of Z-score

- Z-score always have a mean of zero.
- Z-score always have a variance *and* standard deviation of 1.
- If *X* is above the mean, its Z-score is positive; if *X* is below its mean, its Z-score is negative.
- In biostatistics, comparing scores to the mean is both useful and easy. All you have to do is calculate a **Z-score.**

Applications of Z-score

Comparing with the Z-score: The Z-score basically converts raw scores into new scores that shows how they can be compared to the mean. Once we have a Z-score for one value, it becomes easier to compare it to other values with the help of formula:

$$Z = \frac{X - \overline{X}}{SD}$$

- The numerator tells us how much the raw score differs from the mean and is sometimes called the deviant score (it is most often used for calculation of standard deviation).
- The standard deviation is in the denominator and indicates that researcher is converting the deviant score into units of standard deviations.
- Any distribution of raw scores can be converted to a distribution of Z-scores. This means that instead of the raw score units, like test percentage points, along the x-axis, we will have units of standard deviations and it is useful for measuring.

Measuring with the Z-score: A Z-score just measures how much a score deviates from the mean in terms of standard deviations. It means that we can compare two raw scores by putting them both in terms of standard deviations. So, a Z-score allows to compare raw scores, even from different distributions. This is because they are in terms of standard deviations instead of other units. Therefore, we can say that Z-scores are a strong statistical tool worth having and using.

For example, suppose the national average MAT score is 1002 with a standard deviation of 194 points. With the help of formula, the Z-score is calculated as +1.48. It means that the student's score is 1.48 standard deviations above the mean.

T-score

A T-score is a form of standardized test statistic besides the Z-score. The T-score formula enables to take an individual score and transform it into a standardized form which further helps a researcher to compare scores. For example, T-score shows how much your bone density is higher or lower than the bone density of a healthy 30-year-old adult. A healthcare provider looks at the lowest T-score to diagnosis osteoporosis. According to the World Health Organization (WHO) "T-score of –1.0 or above is normal bone density". To compare T-score with Z-score, we can say that T-score is a comparison of a person's bone density with that of a healthy 30-year-old of the same sex. Whereas, Z-score is a comparison of a person's bone density with that of an average person of the same age and sex.

Features of T-score

- Very similar to Z-scores.
  - A bunch of T-scores form a T-distribution.
  - Provides way of judging how extreme a sample mean is.
- T-test is done when SD (σ) is unknown.
- Used for hypothesis testing:

  For example: You wonder if college students really get 8 hours of sleep
  - H_0: μ = 8 (College students do get eight hours of sleep)
  - H_a: μ ≠ 8 (College students do not get eight hours of sleep)

- T-distribution provides foundation for T-test:
  - Can be done on SPSS.
  - Can be done by hand.
- Key difference: T-test is done when σ is unknown.
- T-score is calculated as:

$$t = \frac{\overline{x} - \mu}{\frac{S}{\sqrt{n}}}$$

t = T-score

$\overline{x}$ = sample mean

μ = population mean

S = sample standard deviation

n = sample size

If you have only one item in sample, the square root in the denominator becomes $\sqrt{1}$. This means the formula becomes:

$$t = \frac{\overline{x} - \mu_0}{s}$$

Must Know

The larger the t score, the larger the difference is between the groups under test. It is influenced by many factors including:

- How many items are in your sample
- The means of your sample
- The mean of the population from which your sample is drawn
- The standard deviation of your sample

Example: A pharmaceutical institute claims its graduates earn an average of rupees 300 per hour. A sample of 15 graduates is selected and found to have a mean salary of rupees 280 with a SD of rupees 50. Assuming the claim of pharmaceutical institute true, what is the probability that the average salary of graduates will be no >₹280.

Solution:

Step 1: Put the information into the formula and solve:

$\overline{x}$ = sample mean = 280

μ_0 = population mean = 300

s = standard deviation of sample = 50

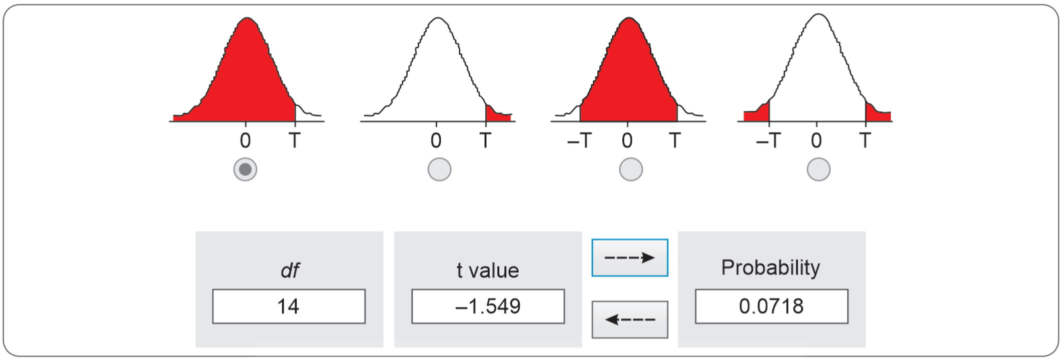
n = sample size = 15

Now,

$$t = (280 - 300)/\ (50/\sqrt{15}) = -20/12.909945 = -1.549.$$

Step 2: Subtract 1 from the sample size to get the degrees of freedom as 15 – 1 = 14. The degrees of freedom lets one to know which form of the T-distribution is to be used.

Step 3: Use an online calculator to find the probability using the obtained degrees of freedom. Here are the results (Fig. 16.7).



df: 14

t value: −1.549

Probability: 0.0718

Figure 16.7: Using an online calculator to find the probability

Note that button under the left tail has been selected, as we are looking for a result that is not <₹280: The probability is 0.0718, or 7.18%.

T-score in Psychology and Education

A T-score testing is a special term, which is not the same as a T-score that we get from a T-test.

Practical Tip

T-scores in T-test can be positive or negative but T-scores in psychology testing are always positive, with an average of 50.

A T-score is similar to a Z-score as it represents the number of SDs **from the mean**. While the Z-score returns values between −5 and 5 (generally, scores fall between −3 and 3) standard deviations from the mean, whereas the T-score has a greater value and results return between 0 and 100 (most scores fall between 20 and 80) (Table 16.8). Many people prefer T-scores because they lack negative numbers and are easier to work with. Moreover, there is a larger range so decimals are almost eliminated. The table here shows Z-scores with their equivalent T-scores.

TABLE 16.8: Z-scores with their equivalent T-scores

Z-score	−5	−4	−3	−2	−1	0	1	2	3	4	5
T-score	0	10	20	30	40	50	60	70	80	90	100

T-score Conversion in Psychology

Calculating a T-score is just a conversion from a Z-score to a T-score, like conversion of Celsius to Fahrenheit. The formula is:

$$\text{T-score} = (Z \times 10) + 50$$

Example: A candidate takes a written test for a job where the average score is 1026 and the standard deviation is 209. The candidate scores 1100. Calculate the T-score for this candidate.

Note: If the Z-score for a question is given, jump to Step 2.

Solution:

Step 1: Calculate the Z-score. The Z-score for the data in this sample question is 0.354.

Step 2: Multiply the Z-score from Step 1 by 10:

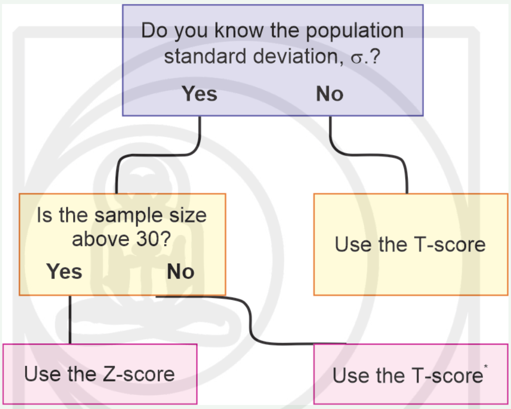
$$10 \times 0.354 = 3.54$$

Step 3: Add 50 to the result from Step 2:

$$3.54 + 50 = 53.54$$

Practical Tips

Decision-Making: T-score Versus Z-score


Do you know the population standard deviation, σ.?
Yes
No
Is the sample size above 30?
Yes
No
Use the T-score
Use the Z-score
Use the T-score*

- **Note that** Z-scores and T-scores both represent standard deviations from the mean, but while "0" on a Z-score is 0 standard deviations from the mean, a "50" on a T-score represents the same thing. It is because T-scores use a mean of 50 and Z-scores use a mean of 0.
- A T-score of >50 is above average; below 50 is below average. In general, a T-score of above 60 means that the score is in the top one-sixth of the distribution; above 63, the top one-tenth. A T-score below 40 indicates a lowest one-sixth position; below 37, the bottom one-tenth.

T-score versus Z-score—WHO diagnostic

T-scores	Z-scores
• WHO diagnostic classification in postmenopausal women and men age 50 and older • WHO classification with T-score cannot be applied to healthy premenopausal women, men under age 50, and children	• For use in reporting BMD in healthy premenopausal women, men under age 50, and children • Z-score –2.0 or less is defined as "below the expected range for age" • Z-score above –2.0 is "within the expected range for age"

RELIABILITY AND VALIDITY OF TEST SCORES

Before discussing reliability, we have to know why we need it. The errors come in the research due to systemic errors or they may occur randomly.

Reliability of test is connected with validity. Accuracy followed in all the steps of research from inception till inference, gives accurate, reliable and valid results.

Validity

Validity refers to a test's accuracy. A test is valid when it measures what it is intended to measure. The intended uses for most tests fall into one of three categories, and each category is associated with a different method for establishing validity:

- **Content validity:** The test is used to obtain information about an examinee's familiarity with a particular content or behavior domain.
- **Construct validity:** The test is administered to determine the extent to which an examinee possesses a particular hypothetical trait
- **Criterion-related validity:** The test is used to estimate or predict an examinee's standing or performance on an external criterion.

Content Validity

A test has content validity (Fig. 16.8) to the extent that it adequately samples the content or behavior domain that it is designed to measure.

- If test items are not a good sample, results of testing will be misleading.
- Although content validation is sometimes used to establish the validity of personality, aptitude, and attitude tests, it is most associated with achievement-type tests that measure knowledge of one or more content domains and with tests designed to assess a well-defined behavior domain.
- Adequate content validity would be important for a statistics test and for a work (job) sample test.
- The degree to which the measured variable appears to have adequately.
- Content validity is usually "built into" a test as it is constructed through a systematic, logical, and qualitative process that involves clearly identifying the content or behavior domain to be sampled and then writing or selecting items that represent that domain.
- Once a test has been developed, the establishment of content validity relies primarily on the judgment of subject matter experts.
- If experts agree that test items are an adequate and representative sample of the target domain, then the test is said to have content validity.

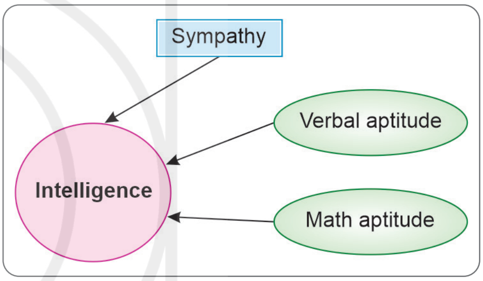


Figure 16.8: Content validity

Although content validation depends mainly on the judgment of experts, supplemental quantitative evidence can be obtained.

Characteristics of Content Validity

If a test has adequate content validity:

- A coefficient of internal consistency will be large.
- The test will correlate highly with other tests that purport to measure the same domain; and
- Pre-/post-test evaluations of a program designed to increase familiarity with the domain will indicate appropriate changes.

Construct Validity

When a test has been found to measure the hypothetical trait (construct) it is intended to measure, the test is said to have construct validity. A construct is an abstract characteristic that cannot be observed directly but must be inferred by observing its effects. Intelligence, mechanical aptitude, self-esteem, and neuroticism are all constructs. When a test has been found to measure the hypothetical trait (construct) it is intended to measure, the test is said to have construct validity. A construct is an abstract characteristic that cannot be observed directly but must be inferred by observing its effects.

Methods to Establish Construct Validity

There is no single way to establish a test's construct validity. Instead, construct validation entails a systematic accumulation of evidence showing that the test actually measures the construct it was designed to measure. Various methods used to establish this type of validity each answer a slightly different question about the construct and include the following:

- **Assessing the test's internal consistency:** Do scores on individual test items correlate highly with the total test score; i.e., are all of the test items measuring the same construct?
- **Studying group differences:** Do scores on the test accurately distinguish between people who are known to have different levels of the construct?
- **Conducting research to test hypotheses about the construct:** Do test scores change, following an experimental manipulation, in the direction predicted by the theory underlying the construct?
- **Assessing the test's convergent and discriminant validity:** Does the test have high correlations with measures of the same trait (convergent validity) and low correlations with measures of unrelated traits (discriminant validity)?
- **Assessing the test's factorial validity:** Does the test have the factorial composition it would be expected to have; i.e., does it have factorial validity?

Must Know

Construct validity is said to be the most theory-laden of the methods of test validation. The developer of a test designed to measure a construct begins with a theory about the nature of the construct, which then guides the test developer in selecting test items and in choosing the methods for establishing the test's validity.

Face Validity

The extent to which the measured variable appears to be an adequate measure of the conceptual variables. For example, liking for Japanese was assessed on a scale:

Strongly disagree 1 2 3 4 5 6 7 8 Strongly agree.

The results obtained are shown in Figure 16.9 as measured variable among all the subjects that is liking for Japanese and the result was represented by conceptual variable.

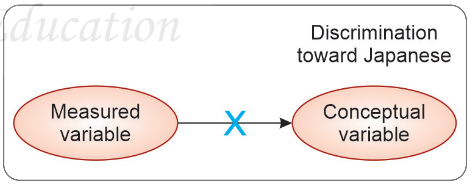


Figure 16.9: Results of liking for Japanese

Must Know

Difference between Content Validity and Face Validity

Content validity refers to the systematic evaluation of a test by experts who determine whether or not test items adequately sample the relevant domain, while face validity refers simply to whether or not a test "looks like" it measures what it is intended to measure.

Although face validity is not an actual type of validity, it is a desirable feature for many tests. If a test lacks face validity, examinees may not be motivated to respond to items in an honest or accurate manner. A high degree of face validity does not, however, indicate that a test has content validity.

Criterion-Related Validity

- **Convergent validity:** The extent to which a measured variable is found to be related to other measured variables designed to measure the same conceptual variable.
- **Discriminant validity:** The extent to which a measured variable is found to be unrelated to other measured variables designed to measure the different conceptual variables.
- **Criterion validity:** The extent to which a self-report measure correlates with a behavioral measured variable.
- **Predictive validity:** The extent to which the scores can predict the participants' future performance.
- **Concurrent validity:** The extent to which the self-report measure correlates with the behavioral measure that is assessed at the same time.

Practical Tips

To improve reliability and validity of test:

- Conduct a pilot test, trying out a questionnaire or other research instruments on a small group.
- Use multiple measures.
- Ensure variability that is in your measures (that can be controlled by a researcher).
- Write good items.
- Request the respondents to take questions seriously
- Make your items nonreactive.
- Be certain to consider face and content validity by choosing reasonable terms and cover a broad range of issues reflecting the conceptual variables.
- Use existing measures.

Reliability

Reliability is one of the important characteristics of any consistency test. It refers to the precision or accuracy of the measurement of score. Reliability means the stability consistency of a test measure or protocol (Fig. 16.10).

- Rosenthal in 1991 described reliability as a major concern when a psychological test is used to measure some attribute or behavior.
- Anastasi in 1968 said that reliability refers to the consistency of scores that are obtained by the same individuals when:
  - Re-examined with test on different occasions.
  - Or with different sets of equivalent items.
  - Or under other variable examining conditions.

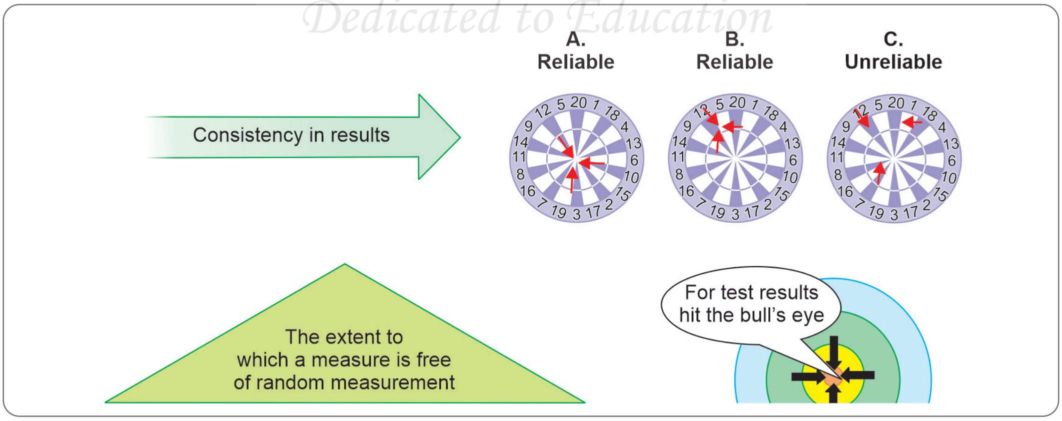

A.
Reliable
B.
Reliable
C.
Unreliable
Consistency in results
The extent to which a measure is free of random measurement
For test results hit the bull's eye

Figure 16.10: Consistency is reliability

Source of Errors in Reliability

Errors are caused by:

- **Examinee-specific factors** such as concentration, motivation, fatigue, boredom, momentary lapses of memory, carelessness in marking answers, and luck in guessing.
- **Test-specific factors** like the specific set of questions selected for a test, ambiguous or tricky items, and poor directions.
- **Scoring-specific factors** like carelessness, no uniform scoring guidelines, and counting or computational errors.

These errors are random. Therefore, it is desirable to use tests with good measures of reliability so as to ensure that the test scores reflect more than just random error.

Tests of Reliability

Test reliability means the consistency of scores that subjects will receive on alternate forms of the same. It provides a measure of the extent to which an examinee's score reflects random measurement error. When a test is reliable, it provides dependable consistent results and, for this reason, the term consistency is often given as a synonym for reliability.

Reliability is a precursor to test validity. That is, if test scores cannot be assigned consistently, it is impossible to conclude that they accurately measure the domain of interest. Validity as earlier explained means the extent to which the inferences made from a test are justified and accurate. Further, validity is the psychometric property about which we are most concerned. However, formally assessing the validity of a specific use of a test is laborious and time-consuming. Therefore, reliability analysis is often a first-step in the test validation process.

If the test is unreliable, one can stop investigating. If the test has adequate reliability, then a validation study is important.

Reliability is also internal consistency. It is the extent to which the scores on the items correlate with each other and thus are all measuring the true score rather than reflecting random error (Fig. 16.11).

Norm referenced tests:

- It is frequently used for norm referenced tests (NRTs). It is a measure of how well the items on the test measure the same construct or idea. This method has an advantage that it is capable to be conducted using a single form given in a single administration.

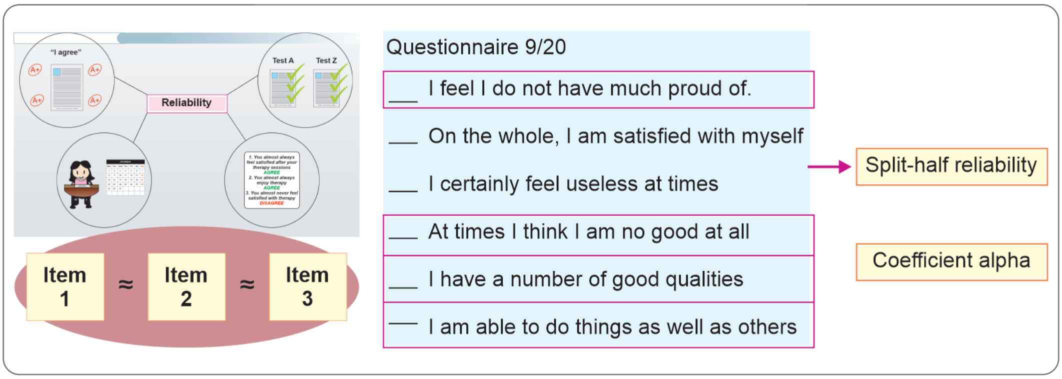

"I agree"
A+
A+
A+
A+
Reliability
Test A
Test Z
1. You almost always feel satisfied after your therapy sessions
AGREE
2. You almost always enjoy therapy
AGREE
3. You almost never feel satisfied with therapy
DISAGREE
Item 1
≈
Item 2
≈
Item 3
Questionnaire 9/20
___ I feel I do not have much proud of.
___ On the whole, I am satisfied with myself
___ I certainly feel useless at times
___ At times I think I am no good at all
___ I have a number of good qualities
— I am able to do things as well as others
Split-half reliability
Coefficient alpha

Figure 16.11: Measuring internal consistency

- The internal consistency method estimates how well the set of items on a test correlate with one another; that is, how similar the items on a test form are to one another.
- Many test analysis software programs produce this reliability estimate automatically.
- **Split-half reliability and coefficient alpha** are two methods for evaluating internal consistency. Both involve administering the test once to a single group of subjects, and both yield a reliability coefficient that is also known as the coefficient of internal consistency.

Necessity of Reliability

We need reliability to know the truth of research outcomes. We need it to have effective results. Reliability gives statistical power to the results. The relationship between variables can be assured with the help of reliability (Fig. 16.12).

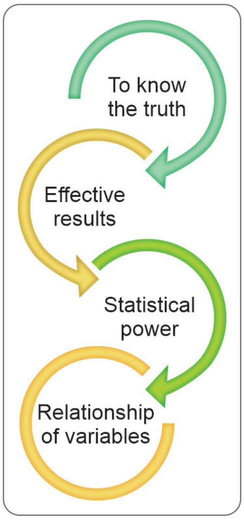


Figure 16.12: Necessity of reliability

Reliability Coefficient

Most methods for estimating reliability produce a reliability coefficient, which is a correlation coefficient that ranges in value from 0.0 to + 1.0. When a test's reliability coefficient is 0.0, this means that all variability in obtained test scores is due to measurement error. Conversely, when a test's reliability coefficient is +1.0, this indicates that all variability in scores reflects true score variability.

The reliability coefficient is symbolized with the letter "r" and a subscript that contains two of the same letters or numbers (e.g., "r_{xx}"). The subscript indicates that the correlation coefficient was calculated by correlating a test with itself rather than with some other measure.

Regardless of the method used to calculate a reliability coefficient, it is interpreted directly as the proportion of variability in obtained test scores that reflects true score variability.

For example, a reliability coefficient of 0.84 indicates that 84% of variability in scores is due to true score differences among examinees, while the remaining 16% (1.00 – 0.84) is due to measurement error.

True score variability (84%)	Error (16%)

It must be noted that a reliability coefficient does not provide any information about what is actually being measured by a test. It only indicates whether the attribute measured by the test—whatever it is being assessed in a consistent and precise way. Whether the test is actually assessing what it was designed to measure, is addressed by an analysis of the test's validity.

Factors Affecting Reliability Coefficient

The magnitude of the reliability coefficient is affected not only by the sources of error, but also by the length of the test, the range of the test scores, and the probability that the correct response to items can be selected by guessing.

- Test length
- Range of T-scores
- Guessing

Test length: The longer the test, the larger the test's reliability coefficient.

Range of test scores: The range is directly affected by the degree of similarity of subjects with regard to the attribute measured by the test. When examinees are heterogeneous, the range of scores is maximized. The range is also affected by the difficulty level of the test items. When all items are either very difficult or very easy, all examinees will obtain either low or high scores, resulting in a restricted range. Therefore, the best strategy is to choose items so that the average difficulty level is in the mid-range ($r = 0.50$).

Guessing: As the probability of correctly guessing answers increases, the reliability coefficient decreases. All other things being equal, a true/false test will have a lower reliability coefficient than a four-alternative multiple-choice test which, in turn, will have a lower reliability coefficient than a free recall test.

Remember that in contrast to other correlation coefficients, the reliability coefficient is never squared to interpret it but is interpreted directly as a measure of true score variability. A reliability coefficient of 0.89 means that 89% of variability in obtained scores is true score variability.

The selection of a method for estimating reliability depends on the nature of the test. Each method not only entails different procedures but is also affected by different sources of error. For many tests, more than one method should be used.

Types of Reliability

It is of the following types (Fig. 16.13):

- Test–retest reliability
- Split-half or internal consistency reliability
- Parallel-forms reliability or equivalent-forms
- Alternate form reliability
- Inter-rater reliability

Test-retest Reliability

The test-retest method for estimating reliability involves administering the same test to the same group of examinees on two different occasions and then correlating the two sets of scores. When using this method (Fig. 16.14), the reliability coefficient indicates the degree of stability (consistency) of examinees' scores over time and is also known as the coefficient of stability.

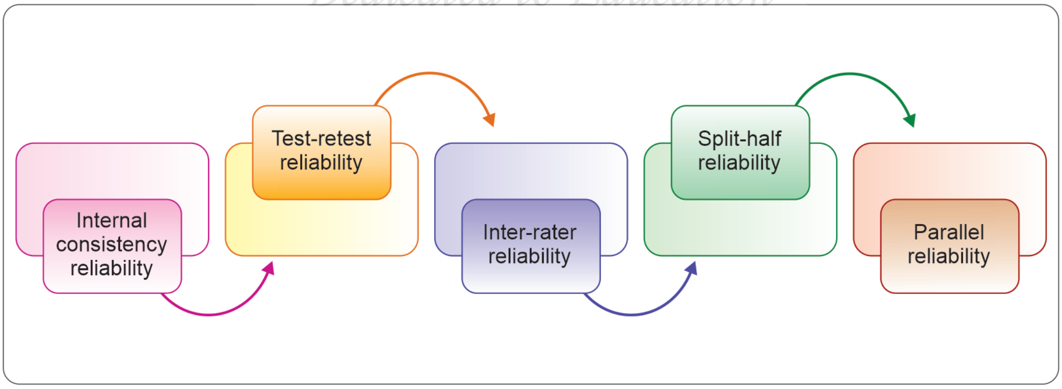


Figure 16.13: Types of reliability

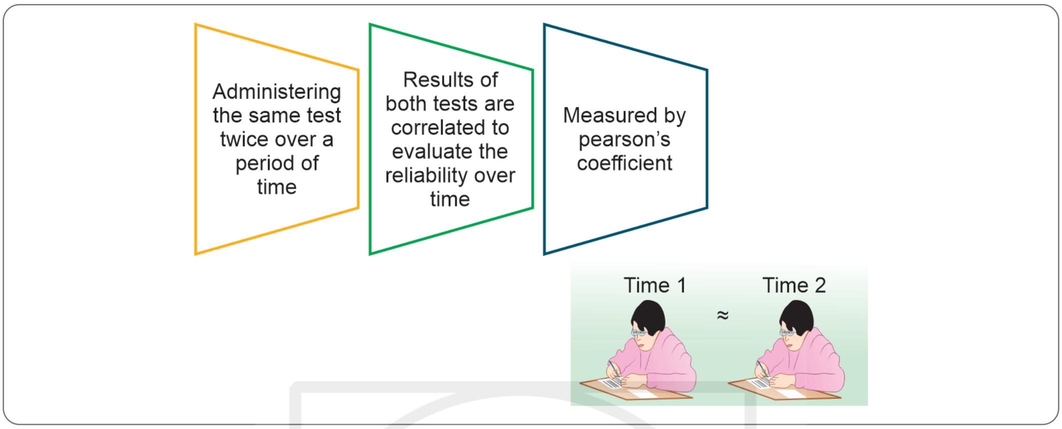

Administering the same test twice over a period of time
Results of both tests are correlated to evaluate the reliability over time
Measured by pearson's coefficient
Time 1
≈
Time 2

Figure 16.14: Process of test-retest reliability

The two sets, when correlated, give the value of the reliability coefficient. The test-retest reliability is an instrument that measures at two times for multiple persons. It computes correlation between the two measures. It assumes that there is no change in the underlying trait between Time 1 and Time 2. The scores from Time 1 to Time 2 can then be correlated in order to evaluate the test for stability, over time. For example, it can determine how much data a student memorizes on a test. The extent to which scores on the same measured variable correlates with each other on two different times is shown in the Table 16.9 as:

Factors contributing to test-retest reliability:

- Clear instructions for administrators, research participants, and raters.
- Unambiguously phrased tasks/questions.
- Tasks/questions in participants' first language or target language at appropriate level of difficulty.
- Reliability has subtypes that must be satisfied before a test or assessment is carried out.

Primary sources of measurement error:

- Any random factors related to the time that passes between the two administrations of the test. These time sampling factors include:
  - Random fluctuations in examinees over time (e.g., changes in anxiety or motivation) and random variations in the testing situation.
  - Memory and practice also contribute to error when they have random carryover effects; i.e., when they affect many or all examinees but not in the same way.

TABLE 16.9: Retesting effects of test-retest reliability

Score at one time	Score at another time
4 I feel I do not have much proud of	4 I feel I do not have much proud of
3 On the whole, I am satisfied with myself	4 On the whole, I am satisfied with myself
2 I certainly feel useless at times	1 I certainly feel useless at times
1 At times I think I am no good at all	1 At times I think I am no good at all
4 I have a number of good qualities	4 I have a number of good qualities
3 I am able to do things as well as others	4 I am able to do things as well as others

Use:

- Test-retest reliability is appropriate for determining the reliability of tests designed to measure attributes that are relatively stable over time and that are not affected by repeated measurement.
- It would be appropriate for a test of aptitude, which is a stable characteristic, but not for a test of mood, since mood fluctuates over time, or a test of creativity, which might be affected by previous exposure to test items.

Split-half Reliability

Other name for split-half reliability is internal consistency reliability. It indicates the homogeneity of the test. In this method, the test is divided into two equal or nearly equal halves. Commonly, the odd-even method is used here. It measures the extent to which the scores on the items correlate with each other and thus measures the true score rather than reflecting random error (Fig. 16.15).

It is a measure of consistency where a test is split in two and the scores for each half of the test are compared with one another. A test is split into two—odds and evens. If the two scores for the two tests are similar then the test is reliable. It measures internal consistency. It tells how well the test components contribute to the construct that is being measured.

In split-half reliability, a test for a single area of knowledge, is split into two parts and then both parts are given to one group of students at the same time. Then scores from both parts of the test are correlated. A reliable test will show high correlation, indicating that a student would perform equally well or poorly on both halves of the test.

It is most commonly used for multiple choice tests that one can theoretically use for any type of test—even tests with essay questions.

Steps to calculate split-half reliability:

1. Administer the test to a large group students (generally, over 30).
2. Randomly divide the test questions into two parts. For example, separate even questions from odd questions.
3. Score each half of the test for each student.
4. Find correlation coefficient for the two halves.

Drawbacks:

It only works for a large set of questions (100-point test is required), which measures the same construct/area of knowledge.

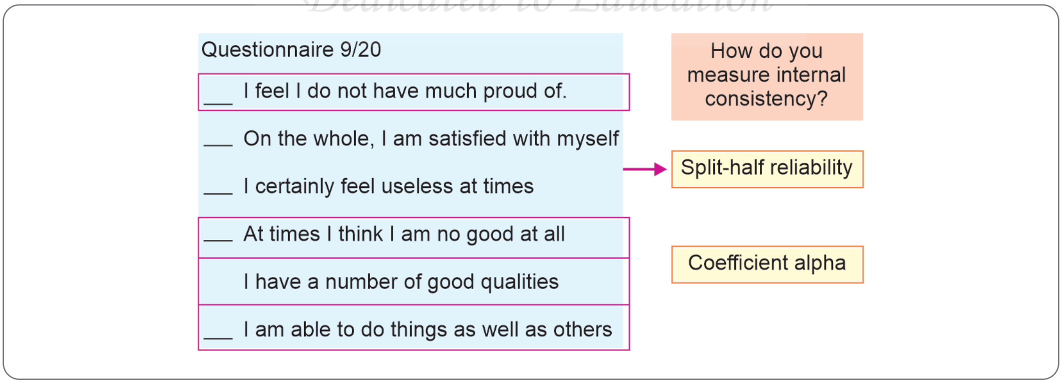

Questionnaire 9/20
___ I feel I do not have much proud of.
___ On the whole, I am satisfied with myself
___ I certainly feel useless at times
___ At times I think I am no good at all
I have a number of good qualities
___ I am able to do things as well as others
How do you measure internal consistency?
Split-half reliability
Coefficient alpha

Figure 16.15: Split-half reliability

Parallel-forms Reliability

Parallel-forms reliability is a measure of reliability obtained by administering different versions of an assessment tool (both versions must contain items that probe the same construct, skill, knowledge base, etc.) to the same group of individuals. The scores from the two versions can then be correlated in order to evaluate the consistency of results across alternate versions.

Many programs develop multiple and parallel-forms of an examination to provide test security like Series A, Series B and so on. These are constructed to match the test blueprint or originality, and the parallel test forms are constructed to be similar in difficulty level of average item. In other words, it uses one set of questions divided into two sets or forms which measure the same knowledge or skill.

Parallel-forms reliability is estimated by administering both patterns of the exam to the same group of examinees. While the time between the two test administrations should be short, it does need to be long enough as it may affect the scores of examinees due to fatigue. The examinees' scores on the two test forms are correlated in order to determine how similarly the two-test pattern function. This reliability estimate is a measure of how consistent examinees' scores can be expected to be across test patterns. For example, the test patterns in GRE, SAT, GMAT, TOEFL, etc.

To assess a test's alternate forms reliability, two equivalent forms of the test are administered to the same group of examinees and the two sets of scores are correlated.

Must Know

Difference between split-half and parallel-forms reliability tests:

- Split-half reliability is similar to parallel form reliability that uses one set of questions divided into two equivalent sets. The sets are given to the same students, usually within a fixed time frame, for example, one set of test questions is given on Monday and another set on Saturday. With split-half reliability, the two tests are given to one group of students who sit the test at the same time.
- The two tests in parallel-forms reliability are equivalent and are independent of each other. This is not required in split-half reliability. Here, the two sets do not have to be equivalent ("parallel").

Alternate form Reliability

Indicates the consistency of responding to different item samples (the two test forms) and, when the forms are administered at different times, the consistency of responding over time.

- The alternate forms reliability coefficient is also called the coefficient of equivalence when the two forms are administered at about the same time;
- The coefficient of equivalence and stability when a relatively long period of time separates administration of the two forms.

Source of error: The primary source of measurement error for alternate forms reliability is content sampling, or error introduced by an interaction between different subjects' knowledge and the different content assessed by the items included in the two forms (e.g., Form A and Form B).

The items in Form A might be a better match of one subject's knowledge than items in Form B, while the opposite is true for another subject. In this situation, the two scores obtained by each examinee will differ, which will lower the alternate form reliability coefficient.

When administration of the two forms is separated by a period of time, time sampling factors also contribute to error.

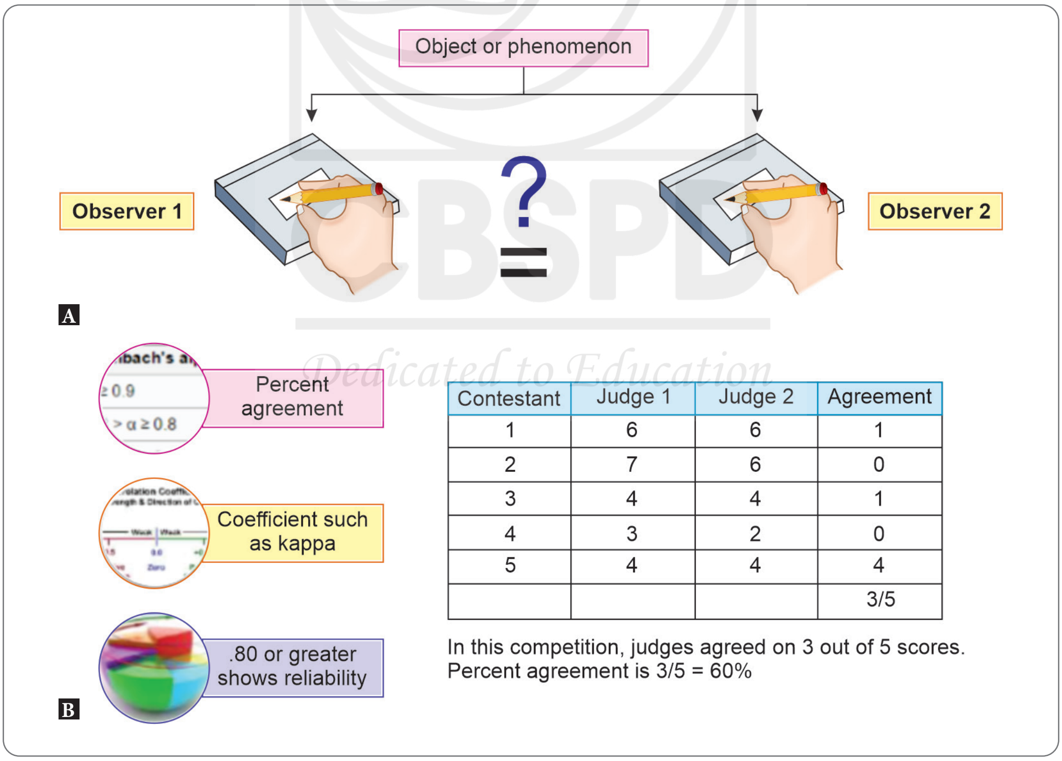
Limitations: Like test-retest reliability, alternate form reliability is not appropriate when the attribute measured by the test is likely to fluctuate over time (and the forms will be administered at different times) or when scores are likely to be affected by repeated measurement.

- If the same strategies required to solve problems on Form A are used to solve problems on Form B, even if the problems on the two forms are not identical, there are likely to show practice effects.
- When these effects differ for different subjects (i.e., the effects are random), practice will serve as a source of measurement error.
- Although alternate form reliability is considered by some experts to be the most rigorous (and best) method for estimating reliability, it is not often assessed due to the difficulty in developing forms that are truly equivalent.

Inter-Rater Reliability

All of the methods for estimating reliability discussed are intended to be used for objective tests. When a test like performance task is there, or other items, which need to be scored by human raters, then the reliability of those raters has to be estimated. This reliability method asks the question, “If multiple raters scored a single subject or examinee’s performance, would the examinee receive the same score”. Inter-rater reliability measures (Figs 16.16A and B).

- The dependability or consistency of scores that might be expected across raters.



Contestant	Judge 1	Judge 2	Agreement
1	6	6	1
2	7	6	0
3	4	4	1
4	3	2	0
5	4	4	4
			3/5

In this competition, judges agreed on 3 out of 5 scores.
Percent agreement is 3/5 = 60%

Figures 16.16A and B: **A.** Inter-rater reliability; **B.** Measurement of inter-rater reliability

- It tells about the degree of agreement between raters.
- It gives a score that tells how much similarity or consensus is there in the ratings given by judges or raters.

Inter-rater reliability is of concern whenever test scores depend on a rater's judgment.

A test constructor may want to make sure that an essay test, a behavioral observation scale, or a projective personality test have adequate inter-rater reliability. This type of reliability is assessed either by calculating a correlation coefficient (e.g., a kappa coefficient or coefficient of concordance) or by determining the percent agreement between two or more raters.

Sources of error for inter-rater reliability include factors related to the raters such as lack of motivation and rater biases and characteristics of the measuring device.

An inter-rater reliability coefficient is likely to be low, for instance, when rating categories are not exhaustive (i.e., do not include all possible responses or behaviors) and/or are not mutually exclusive.

Practical Tips

To determine a test's split-half reliability, the test is split into equal halves so that each subject has two scores (one for each half of the test).

Scores on the two halves are then correlated. Tests can be split in several ways, but probably the most common way is to divide the test on the basis of odd- versus even-numbered items.

A problem with the split-half method is that it produces a reliability coefficient that is based on test scores that were derived from one-half of the entire length of the test.

If a test contains 30 items, each score is based on 15 items. Because reliability tends to decrease as the length of a test decreases, the split-half reliability coefficient usually underestimates a test's true reliability. For this reason, the split-half reliability coefficient is ordinarily corrected using the Spearman-Brown prophecy formula, which provides an estimate of what the reliability coefficient would have been had it been based on the full length of the test.

Cronbach's coefficient alpha: *Cronbach's coefficient alpha* also involves administering the test once to a single group of subjects. However, rather than splitting the test in half, a special formula is used to determine the average degree of inter-item consistency.

One way to interpret coefficient alpha is as the average reliability that would be obtained from all possible splits of the test. Coefficient alpha tends to be conservative and can be considered the lower boundary of a test's reliability.

When test items are scored dichotomously (right or wrong), a variation of coefficient alpha known as the Kuder-Richardson Formula 20 (KR-20) can be used.

Note that content sampling is a source of error for both split-half reliability and coefficient alpha.

- For split-half reliability, content sampling refers to *the error resulting from differences between the content of the two halves of the test* (i.e., the items included in one half may better fit the knowledge of some examinees than items in the other half); for coefficient alpha, content (item) sampling refers to *differences between individual test items rather than between test halves*. Coefficient alpha also has as a source of error, *the heterogeneity of the content domain.*
- A test is heterogeneous with regard to content domain when its items measure several different domains of knowledge or behavior.
- The greater the heterogeneity of the content domain, the lower the inter-item correlations and the lower the magnitude of coefficient alpha.
- The methods for assessing internal consistency reliability are useful when a test is designed to measure a single characteristic, when the characteristic measured by the test fluctuates over time, or when scores are likely to be affected by repeated exposure to the test.
- They are not appropriate for assessing the reliability of speed tests because, for these tests, they tend to produce spuriously high coefficients. (For speed tests, alternate form reliability is usually the best choice.)

Methods to Improve Reliability and Validity

- Consensual observer drift can be eliminated by having raters work independently or by alternating raters.
- Rating accuracy is also improved when raters are told that their ratings will be checked.
- Overall, the best way to improve both inter- and intra-rater accuracy is to provide raters with training that emphasizes the distinction between observation and interpretation.

Must Know

Remember the Spearman-Brown formula is related to split-half reliability and KR-20 is related to the coefficient alpha. Also know that alternate form reliability is the most thorough method for estimating reliability and that internal consistency reliability is not appropriate for speed tests.

Interpretation of Reliability

The interpretation of a test's reliability entails considering its effects on the scores achieved by a group of examinees as well as the score obtained by a single examinee.

Reliability Coefficient

A reliability coefficient is interpreted directly as the proportion of variability in a set of test scores that is attributable to true score variability.

A reliability coefficient of 0.84 indicates that 84% of variability in test scores is due to true score differences among examinees, while the remaining 16% is due to measurement error.

While different types of tests can be expected to have different levels of reliability, for most tests in the social sciences, reliability coefficients of 0.80 or larger are considered acceptable.

When interpreting a reliability coefficient, it is important to keep in mind that there is no single index of reliability for a given test.

Instead, a test's reliability coefficient can vary from situation to situation and sample to sample. Ability tests, for example, typically have different reliability coefficients for groups of individuals of different ages or ability levels.

Confidence Interval

A common practice when interpreting obtained score is to construct a confidence interval around that score.

The confidence interval helps a test user estimating the range within which an examinee's true score is likely to fall given him or her obtained score.

Standard Error of Measurement

The range is calculated using the standard error of measurement, which is an index of the amount of error that can be expected in obtained scores due to the unreliability of the test. (When raw scores have been converted to percentile ranks, the confidence interval is referred to as a percentile band). The following formula is used to estimate the standard error of measurement:

Standard Error of Measurement or

$$SE_{\text{meas}} = SD_{\text{x}} \times (1 - r_{xx})^{1/2}$$

Where,

SE_{meas} = Standard error of measurement

SD_{x} = Standard deviation of test scores

r_{xx} = Reliability coefficient

The magnitude of the standard error is affected by two factors:

1. Standard deviation of the test scores (SD_{x})
2. The test's reliability coefficient (r_{xx}).

The lower the test's standard deviation and the higher its reliability coefficient, the smaller the standard error of measurement (and vice versa).

Practical Tips

- The standard error is a type of standard deviation, it can be interpreted in terms of the areas under the normal curve.
- With regard to confidence intervals, this means that a 68% confidence interval is constructed by adding and subtracting one standard error to a subject's obtained score; a 95% confidence interval is constructed by adding and subtracting two standard errors; and a 99% confidence interval is constructed by adding and subtracting three standard errors.
- Due to the effects of measurement error, obtained test scores tend to be biased (inaccurate) estimates of true scores. More specifically, scores above the mean of a distribution tend to overestimate true scores, while scores below the mean tend to underestimate true scores.
- The farther are the mean and obtained score, the greater is the bias.

STUDENT ASSIGNMENT

LONG ANSWER QUESTIONS

1. What do you understand by scaling? Describe it.
2. What is Z-score and T-score?
3. What is reliability of test scores? What are its reasons and is reliability similar to validity? Discuss.

SHORT ANSWER QUESTIONS

1. Write a short note on test-retest method of reliability.
2. Write about parallel-forms
3. Write a short note on split-half method.

MULTIPLE CHOICE QUESTIONS

1. The measurement of variability which we use as a unit of the scale of measurement in a normal distribution is:

a Average deviation — b Standard deviation
c Range — d Quartile deviation

2. The most stable index of variability is:

a Average deviation — b Standard deviation
c Range — d Median

3. When the scores are distributed symmetrically around a central point and the distribution is not badly skewed, we generally compute:

a Mean — b Median
c Mode — d None of these

4. The formula for finding out Mode from a frequency distribution is:

a 3 median – 2 mean — b 2 median – 3 mean
c 2 mean – 3 median — d 3 mean – 2 median

5. In Psychology and Education, we come across measurement data heavily dependent upon:

a Nominal scale — b Ordinal scale
c Interval scale — d Ratio scale

6. Sex, nationality, occupation, religion, marital status are examples of:

a Quantitative variable — b Qualitative variable
c Discontinuous variable — d Continuous variable

7. The numerical quantities which characterize a population are called:

a Parameters — b Statistics
c Data — d Scores

8. All the important characteristics of a population can be specified in terms of a few:

a Parameters b Scores
c Data d Statistics

9. Statistical inference is concerned with derivation of Scientific inference about generalization of results from:

a The study of a few particular cases
b The study of population as a whole
c The study of a random group
d The study of the entire population of the world.

10. The branch which deals with collection, analysis and interpretation of data obtained by conducting a survey or an experimental study is known as:

a Psychology b Statistics
c Sociology d Mathematics

11. A psychologist wants to make a statement about the mean IQ in the complete population of students in a particular university from a knowledge of the mean completed on the sample of 100 and to estimate the error involved in this statement. For this purpose, he will use procedures from:

a Mathematics b Geometry
c Geography d Inferential statistics

ANSWER KEY

1. b **2.** b **3.** a **4.** a **5.** c **6.** b **7.** a **8.** a
9. a **10.** c **11.** d

Essentials of

Biostatistics

For Paramedical and Allied Health Sciences

Salient Features

- The text is enriched with a variety of formulas, ensuring a deeper practical understanding of statistical analysis when applied to real-world scenarios.
- Simplified solutions are provided in the form of solved examples, making it easier to grasp the complex concepts and reinforcing understanding of each respective topic.
- Divided into 10 Units and 18 chapters, the book covers a wide range of biostatistical concepts—from basic principles to more advanced, complex topics—offering a thorough exploration of the subject.
- Numerous practical examples have been included with step-by-step solutions to illustrate the application of statistical procedures in real-time research and data analysis.
- Practical Tips boxes throughout the book, provide valuable insights and actionable advice, helping the students in the practical implementation of statistical methods.
- Must Know boxes with valuable facts are strategically placed to highlight critical information, ensuring readers are well-informed of key concepts and important details.

Learning Objectives enlist what the students will learn after studying the entire chapter.

LEARNING OBJECTIVES

After the completion of the chapter, the readers will be able to:

- Understand concepts of central tendencies.
- Know about the relation between the measures of central tendencies.

Chapter Outline provides a quick glance of the entire chapter in one go.

CHAPTER OUTLINE

- Introduction
- Central Tendency
- Mean
- Mode

Must Know boxes covering valuable facts are strategically placed to highlight critical information, ensuring readers are well-informed of key concepts and important details.

Must Know

Modifying a distribution by dumping scores or by addition of new scores will generally change the value of the mean and it will affect: Number of scores; Sum of the scores. If a constant value is added to every score in a distribution, then the same constant value is added to the mean. Also, if every score is multiplied by a constant value, then the mean is also multiplied by the same constant value.

Practical Tips boxes throughout the book, provide valuable insights and actionable advice, helping the students in the practical implementation of statistical methods.

Practical Tips

- In both the above given formulae '$n - 1$' is used instead of n in the denominator, because it gives a more accurate estimate of population SD.

Illustrations and Tables are used to make learning easy for students.

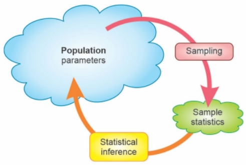


Figure 2.1: The wholesome picture of statistic and parameter

Table 2.1: Comparison of statistic and parameter

Characteristics	Statistic	Parameter
Definition	A characteristic of a small part of the population, i.e., sample.	A fixed measure that describes the target population.
Nature	A variable and known number that depend on the sample of the population.	Parameter is fixed and unknown numerical value.

Important and summarized facts of respective topic are covered under **Takeaway** boxes

Takeaway

$$\textit{Mean} = \frac{\textit{Sum of all values}}{\textit{Total number of values}}$$

Median =Middle value (when the data are arranged in order)

Mode = Most common value

- Central tendency: A score which indicates a position where the center of a distribution tends to be located
- Mean is sum of all scores divided by the number of items

Detailed **Student Assignment** in the form of exercises in each and every chapter will facilitate structured learning and revision of the material provided in the respective chapters.

STUDENT ASSIGNMENT

LONG ANSWER QUESTIONS

1. What is the significance and scope of statistics?
2. How does statistics help in epidemiological studies?

SHORT ANSWER QUESTIONS

1. Write a short note on variables.
2. What are the applications of statistics in medical field?

About the Author

Anju Dhir, *PhD Microbiology,* is presently working as Senior Product Manager and Developmental Editor in Health Sciences division. She is a former Lecturer, Department of Microbiology at Shivalik Institute of Nursing, Shimla, Himachal Pradesh. She is a Gold Medalist in Microbiology. She had been in teaching profession for the last 25 years. Her thesis and research papers are published in national and international journals.

CBSPD
Dedicated to Education

CBS Publishers & Distributors Pvt. Ltd.
4819/XI, Prahlad Street, 24 Ansari Road, Daryaganj, New Delhi 110 002, India
E-mail: feedback@cbspd.com, **Website:** www.cbspd.com
New Delhi | Bengaluru | Chennai | Kochi | Kolkata | Lucknow | Mumbai
Hyderabad | Jharkhand | Nagpur | Patna | Pune | Uttarakhand

ISBN: 978-81-977500-5-2
9 788197 750052