Prepared By:
Dr. Gaurav Sehgal
Associate Professor
School of Business Studies
Department of HRM & OB
Central University of Jammu, Jammu, J&K State

UNIT – II
Topics Covered

- ❖ **Research Design : Definition & Types**
- ❖ **Scales**
- ❖ **Rating Scales / Scaling Techniques**
- ❖ **Measurement - Validity & Reliability**

Prepared By:
Dr. Gaurav Sehgal
Associate Professor
School of Business Studies
Department of HRM & OB
Central University of Jammu, Jammu, J&K State

2.1 Research Design

The design is the structure of any scientific work. It gives direction and systematizes the research. Different types of research designs have different advantages and disadvantages. The method you choose will affect your results and how you conclude the findings. Most scientists are interested in getting reliable observations that can help the understanding of a phenomenon.

There are two main approaches to a research problem, viz, Quantitative Research and Qualitative Research.

- ***Qualitative Research Design***

  Qualitative research design is a research method used extensively by scientists and researchers studying human behavior and habits. It is also very useful for product designers who want to make a product that will sell. For example, a designer generating some ideas for a new product might want to study people's habits and preferences, to make sure that the product is commercially viable. Quantitative research is then used to assess whether the completed design is popular or not.

  Qualitative research is often regarded as a precursor to quantitative research, in that it is often used to generate possible leads and ideas which can be used to formulate a realistic and testable hypothesis. This hypothesis can then be comprehensively tested and mathematically analyzed, with standard quantitative research methods. For these reasons, these qualitative methods are often closely allied with interviews, survey design techniques and individual case studies, as a way to reinforce and evaluate findings over a broader scale. A study completed before the experiment was performed would reveal which of the multitude of brands were the most popular. The quantitative experiment could then be constructed around only these brands, saving a lot of time, money and resources.

  Qualitative methods are probably the oldest of all scientific techniques, with Ancient Greek philosophers qualitatively observing the world around them and trying to come up with answers which explained what they saw.

  - *Design:* The design of qualitative research is probably the most flexible of the various experimental techniques, encompassing a variety of accepted methods and structures. From an individual case study to an extensive interview, this type of study still needs to be carefully constructed and designed, but there is no standardized structure. Case studies, interviews and survey designs are the most commonly used methods.
  - *Advantages:* Qualitative techniques are extremely useful when a subject is too complex be answered by a simple yes or no hypothesis. These types of designs are much easier to plan and carry out. They are also useful when budgetary decisions have to be taken into account. The broader scope covered by these designs ensures that some useful data is always generated, whereas an unproved hypothesis in a quantitative experiment can mean that a lot of time has been wasted. Qualitative research methods are not as dependent upon sample sizes as quantitative methods; a case study, for example, can generate meaningful results with a small sample group.
  - *Disadvantages:* Whilst not as time or resource consuming as quantitative experiments, qualitative methods still require a lot of careful thought and planning, to ensure that the results obtained are as accurate as possible. Qualitative data cannot be mathematically analyzed in the same comprehensive way as quantitative results, so can only give a guide to general trends. It is a lot more open to personal opinion and judgment, and so can only ever give observations rather than results. Any qualitative research design is

usually unique and cannot be exactly recreated, meaning that they do lack the ability to be replicated.

- ***Quantitative Research Design***

  Quantitative research design is the standard experimental method of most scientific disciplines. These experiments are sometimes referred to as true science, and use traditional mathematical and statistical means to measure results conclusively. They are most commonly used by physical scientists, although social sciences, education and economics have been known to use this type of research. It is the opposite of qualitative research. Quantitative experiments all use a standard format, with a few minor inter-disciplinary differences, of generating a hypothesis to be proved or disproved. This hypothesis must be provable by mathematical and statistical means, and is the basis around which the whole experiment is designed. Randomization of any study groups is essential, and a control group should be included, wherever possible. A sound quantitative design should only manipulate one variable at a time, or statistical analysis becomes cumbersome and open to question. Ideally, the research should be constructed in a manner that allows others to repeat the experiment and obtain similar results.

  - *Advantages:* Quantitative research design is an excellent way of finalizing results and proving or disproving a hypothesis. The structure has not changed for centuries, so is standard across many scientific fields and disciplines. After statistical analysis of the results, a comprehensive answer is reached, and the results can be legitimately discussed and published. Quantitative experiments also filter out external factors, if properly designed, and so the results gained can be seen as real and unbiased. Quantitative experiments are useful for testing the results gained by a series of qualitative experiments, leading to a final answer, and a narrowing down of possible directions for follow up research to take.
  - *Disadvantages:* Quantitative experiments can be difficult and expensive and require a lot of time to perform. They must be carefully planned to ensure that there is complete randomization and correct designation of control groups. Quantitative studies usually require extensive statistical analysis, which can be difficult, due to most scientists not being statisticians. The field of statistical study is a whole scientific discipline and can be difficult for non-mathematicians. In addition, the requirements for the successful statistical confirmation of results are very stringent, with very few experiments comprehensively proving a hypothesis; there is usually some ambiguity, which requires retesting and refinement to the design. This means another investment of time and resources must be committed to fine-tune the results. Quantitative research design also tends to generate only proved or unproven results, with there being very little room for grey areas and uncertainty. For the social sciences, education, anthropology and psychology, human nature is a lot more complex than just a simple yes or no response.

2.2 Different Research Designs / Methods

There are various designs which are used in research, all with specific advantages and disadvantages. Which one the scientist uses, depends on the aims of the study and the nature of the phenomenon:

(A) Descriptive Designs

The aim of this design includes: “Observe” and “Describe”. These are classified as: Descriptive Research*; Case Study; Naturalistic Observation; and Survey.

(B) Correlational Studies

The aim of this design includes: “Predict”. These are classified as:
Case Control Study; Observational Study; Cohort Study; Longitudinal Study*; Cross Sectional Study*; and Correlational Studies (in general).

(C) Semi-Experimental Designs

The aim of this design includes: “Determine Causes”. These are classified as:
Field Experiment; Quasi-Experimental Design; and Twin Studies.

(D) Experimental Designs

The aim of this design includes: “Determine Causes”. These are classified as:
True Experimental Design; and Double-Blind Experiment.

(E) Reviewing Other Research

The aim of this design includes: “Explain”. These are classified as:
Literature Review; Meta-analysis; and Systematic Reviews.

(F) Test Study Before Conducting a Full-Scale Study

The aim of this design includes: “Does the Design Work?”. These are classified as:
Pilot Study

(G) Typical Experimental Designs

(i) Simple Experimental Techniques

Pretest-Posttest Design; Control Group; Randomization; Randomized Controlled Trials; Between Subjects Design; and Within Subject Design.

(ii) Complex Experimental Designs

Factorial Design; Solomon Four-Group Design; Repeated Measures Design; Counterbalanced Measures Design; Matched Subjects Design; and Bayesian Probability

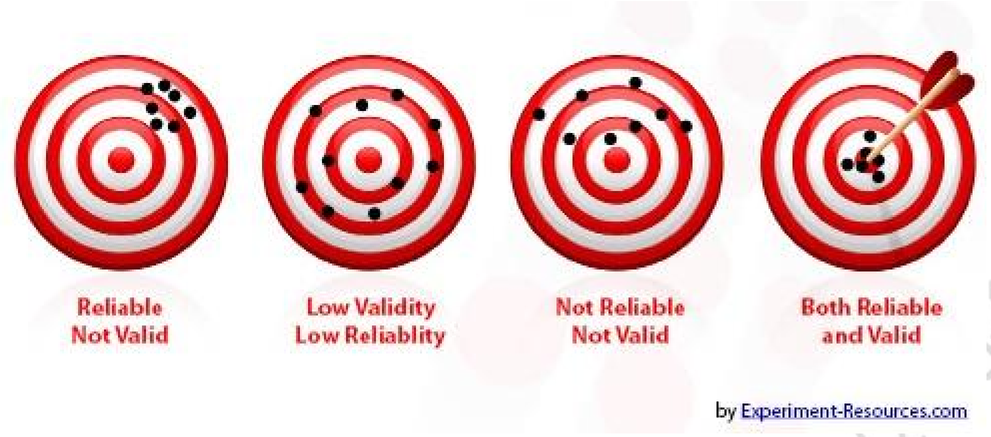
* Detailed Explanations for selected designs as per syllabus has been left as self-study for the students.

2.3 Validity

"Any research can be affected by different kinds of factors which, while extraneous to the concerns of the research, can invalidate the findings" (Seliger & Shohamy 1989, 95).

Validity refers to what degree the research reflects the given research problem, while Reliability refers to how consistent a set of measurements are.

The figure presents a description of the same,



2.4 Types of Validity

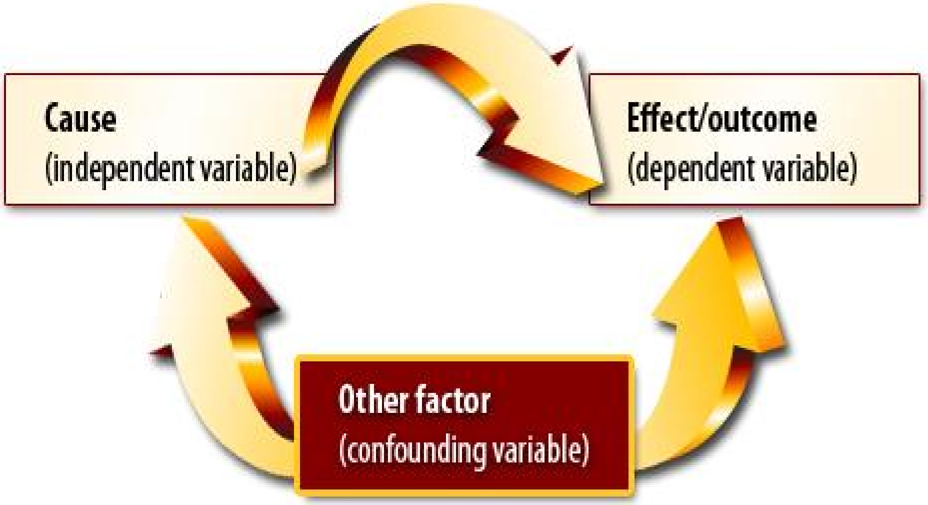
- ***External Validity***

  External validity is about generalization: To what extent can an effect in research, be generalized to populations, settings, treatment variables, and measurement variables?
  External validity is usually split into two distinct types, population validity and ecological validity and they are both essential elements in judging the strength of an experimental design.

  - *Population Validity* is a type of external validity which describes how well the sample used can be extrapolated to a population as a whole. It evaluates whether the sample population represents the entire population, and also whether the sampling method is acceptable. For example, an educational study that looked at a single school could not be generalized to cover children at every Indian School. On the other hand, a MHRD appointed study, that tested every pupil of a certain age group, will have exceptionally strong population validity. Due to time and cost restraints, most studies lie somewhere between these two extremes, and researchers pay extreme attention to their sampling techniques. Experienced scientists ensure that their sample groups are as representative as possible, striving to use random selection rather than convenience sampling.
  - *Ecological Validity* is a type of external validity which looks at the testing environment and determines how much it influences behavior. In the school test example, if the pupils are used to regular testing, then the ecological validity is high because the testing process is unlikely to affect behavior. On the other hand, taking each child out of class and testing them individually, in an isolated room, will dramatically lower ecological validity. The child may be nervous, ill at ease and is unlikely to perform in the same way as they would in a classroom. Generalization becomes difficult, as the experiment does not resemble the real world situation.

- ***Internal Validity***

  Internal validity is a measure which ensures that a researcher's experiment design closely follows the principle of cause and effect. Looking at some extreme examples, a physics experiment into the effect of heat on the conductivity of a metal has a high internal validity. The researcher can eliminate almost all of the potential confounding variables and set up strong controls to isolate other factors. At the other end of the scale, a study into the correlation between income level and the likelihood of smoking has a far lower internal validity. A researcher may find that there is a link between low-income groups and smoking, but cannot be certain that one causes the other. Social status, profession, ethnicity, education, parental smoking, and exposure to targeted advertising are all variables that may have an effect. They are difficult to eliminate, and social research can be a statistical minefield for the unwary.



- ***Test Validity***
  Test validity is an indicator of how much meaning can be placed upon a set of test results. In psychological and educational testing, where the importance and accuracy of tests is paramount, test validity is crucial. Test validity incorporates a number of different validity types, including criterion validity, content validity and construct validity. If a research project scores highly in these areas, then the overall test validity is high.

- ***Criterion Validity***
  Criterion Validity assesses whether a test reflects a certain set of abilities.
  - *Concurrent validity* measures the test against a benchmark test and high correlation indicates that the test has strong criterion validity.
  - *Predictive validity* is a measure of how well a test predicts abilities. It involves testing a group of subjects for a certain construct and then comparing them with results obtained at some point in the future.

- ***Content Validity***
  Content validity is the estimate of how much a measure represents every single element of a construct. It is sometimes called logical or rational validity, is the estimate of how much a measure represents every single element of a construct. For example, an educational test with strong content validity will represent the subjects actually taught to students, rather than asking unrelated questions.
  Content validity is often seen as a prerequisite to criterion validity, because it is a good indicator of whether the desired trait is measured. If elements of the test are irrelevant to

the main construct, then they are measuring something else completely, creating potential bias. In addition, criterion validity derives quantitative correlations from test scores.

Content validity is qualitative in nature, and asks whether a specific element enhances or detracts from a test or research program.

Content validity is related to face validity, but differs wildly in how it is evaluated. Face validity requires a personal judgment, such as asking participants whether they thought that a test was well constructed and useful. Content validity arrives at the same answers, but uses an approach based in statistics, ensuring that it is regarded as a strong type of validity. For surveys and tests, each question is given to a panel of expert analysts, and they rate it. They give their opinion about whether the question is essential, useful or irrelevant to measuring the construct under study. Their results are statistically analyzed and the test modified to improve the rational validity.

Let us now look for an example for Low Content Validity,

Take the example of from employment, where content validity is often used.

A school wants to hire a new science teacher, and a panel of governors begins to look through the various candidates. They draw up a shortlist and then set a test, picking the candidate with the best score. Sadly, he proves to be an extremely poor science teacher. After looking at the test, the education board begins to see where they went wrong. The vast majority of the questions were about physics so, of course, the school found the most talented physics teacher.

However, this particular job expected the science teacher to teach biology, chemistry and psychology. The content validity of test was poor and did not fully represent the construct of 'being a good science teacher.' Suitably embarrassed, the school redesigned the test and submitted it to a panel of educational experts. After asking the candidates to sit the revised test, the school found another teacher, and she proved to be an excellent and well-rounded science teacher. This test had a much higher rational validity and fully represented every element of the construct.

- ***Construct Validity***

Construct validity defines how well a test or experiment measures up to its claims. It refers to whether the operational definition of a variable actually reflect the true theoretical meaning of a concept. The simple way of thinking about it is as a test of generalization, like external validity, but it assesses whether the variable that you are testing for is addressed by the experiment. Construct validity is a device used almost exclusively in social sciences, psychology and education. For example, you might design whether an educational program increases artistic ability amongst pre-school children. Construct validity is a measure of whether your research actually measures artistic ability, a slightly abstract label.

In order words, Construct validity defines how well a test or experiment measures up to its claims. A test designed to measure depression must only measure that particular construct, not closely related ideals such as anxiety or stress.

For major and extensive research, especially in education and language studies, most researchers test the construct validity before the main research. These pilot studies establish the strength of their research and allow them to make any adjustments. Using an educational example, such a pre-test might involve a differential groups study, where researchers obtain test results for two different groups, one with the construct and one without.

The other option is an intervention study, where a group with low scores in the construct is tested, taught the construct, and then re-measured. If there is a significant difference pre and post-test, usually analyzed with simple statistical tests, then this proves good construct validity. There were attempts, after the war, to devise statistical methods to test construct validity, but they were so long and complicated that they proved to be unworkable. Establishing good construct validity is a matter of experience and judgment, building up as much supporting evidence as possible. A whole battery of statistical tools and coefficients are used to prove strong construct validity, and researchers continue until they feel that they have found the balance between proving validity and practicality.

- *Convergent validity* tests that constructs that are expected to be related are, in fact, related.
- *Discriminant validity* tests that constructs that should have no relationship do, in fact, not have any relationship. (also referred to as divergent validity)

- ***Face Validity***

Face validity is a measure of how representative a research project is ‘at face value,’ and whether it appears to be a good project. It is built upon the principle of reading through the plans and assessing the viability of the research, with little objective measurement. Whilst face validity, sometime referred to as representation validity, is a weak measure of validity, its importance cannot be underestimated.

In many ways, face validity offers a contrast to content validity, which attempts to measure how accurately an experiment represents what it is trying to measure. The difference is that content validity is carefully evaluated, whereas face validity is a more general measure and the subjects often have input.

An example could be, after a group of students sat a test, you asked for feedback, specifically if they thought that the test was a good one. This enables refinements for the next research project and adds another dimension to establishing validity.

Face validity is classed as 'weak evidence' supporting construct validity, but that does not mean that it is incorrect, only that caution is necessary.

For example, imagine a research paper about Global Warming. A layperson could read through it and think that it was a solid experiment, highlighting the processes behind Global Warming. On the other hand, a distinguished climatology professor could read through it and find the paper, and the reasoning behind the techniques, to be very poor. This example shows the importance of face validity as useful filter for eliminating shoddy research from the field of science, through peer review.

2.5 Reliability

A definition of reliability may be "Yielding the same or compatible results in different clinical experiments or statistical trials" (the free dictionary). Research methodology lacking reliability cannot be trusted. Replication studies are a way to test reliability.

Note that, both validity and reliability are important aspects of the research methodology to get better explanations of the world.

2.6 Types of Reliability

- ***Test-Retest Reliability***

The test-retest reliability method is one of the simplest ways of testing the stability and reliability of an instrument over time. For example, if a group of students takes a test, you would expect them to show very similar results if they take the same test a few months later. This definition relies upon there being no confounding factor during the intervening time interval. Instruments such as IQ tests and surveys are prime candidates for test-retest methodology, because there is little chance of people experiencing a sudden jump in IQ or suddenly changing their opinions. On the other hand, educational tests are often not suitable, because students will learn much more information over the intervening period and show better results in the second test.

Example-1: If a group of students take a geography test just before the end of semester and one when they return to school at the beginning of the next, the tests should produce broadly the same results.

If, on the other hand, the test and retest are taken at the beginning and at the end of the semester, it can be assumed that the intervening lessons will have improved the ability of the students. Thus, test-retest reliability will be compromised and other methods, such as split testing, are better.

Even if a test-retest reliability process is applied with no sign of intervening factors, there will always be some degree of error. There is a strong chance that subjects will remember some of the questions from the previous test and perform better.

Some subjects might just have had a bad day the first time around or they may not have taken the test seriously. For these reasons, students facing retakes of exams can expect to face different questions and a slightly tougher standard of marking to compensate.

Even in surveys, it is quite conceivable that there may be a big change in opinion. People may have been asked about their favorite type of bread. In the intervening period, if a bread company mounts a long and expansive advertising campaign, this is likely to influence opinion in favor of that brand. This will jeopardize the test-retest reliability and so the analysis that must be handled with caution.

Example-2: To give an element of quantification to the test-retest reliability, statistical tests factor this into the analysis and generate a number between zero and one, with 1 being a perfect correlation between the test and the retest. Perfection is impossible and most researchers accept a lower level, either 0.7, 0.8 or 0.9, depending upon the particular field of research. However, this cannot remove confounding factors completely, and a researcher must anticipate and address these during the research design to maintain test-retest reliability. To dampen down the chances of a few subjects skewing the results, for whatever reason, the test for correlation is much more accurate with large subject groups, drowning out the extremes and providing a more accurate result.

- ***Inter-rater Reliability***

For any research program that requires qualitative rating by different researchers, it is important to establish a good level of inter-rater reliability, also known as inter-observer reliability. This ensures that the generated results meet the accepted criteria defining reliability, by quantitatively defining the degree of agreement between two or more observers.

For example, watching any sport using judges, such as Olympics ice skating or a dog show, relies upon human observers maintaining a great degree of consistency between observers. If even one of the judges is erratic in their scoring system, this can jeopardize the entire system and deny a participant their rightful prize. Outside the world of sport and hobbies, inter-rater reliability has some far more important connotations and can directly influence your life. Examiners marking school and university exams are assessed on a regular basis, to ensure that they all adhere to the same standards. This is the most important example of inter-observer reliability - it would be extremely unfair to fail an exam because the observer was having a bad day. For most examination boards, appeals are usually rare, showing that the inter-rater reliability process is fairly robust.

- ***Internal Consistency Reliability***

Internal consistency reliability defines the consistency of the results delivered in a test, ensuring that the various items measuring the different constructs deliver consistent scores. For example, an English test is divided into vocabulary, spelling, punctuation and grammar. The internal consistency reliability test provides a measure that each of these particular aptitudes is measured correctly and reliably.

One way of testing this is by using a test-retest method, where the same test is administered some after the initial test and the results compared. However, this creates some problems and so many researchers prefer to measure internal consistency by including two versions of the same instrument within the same test. Our example of the English test might include two very similar questions about comma use, two about spelling and so on. The basic principle is that the student should give the same answer to both - if they do not know how to use commas, they will get both questions wrong. A few nifty statistical manipulations will give the internal consistency reliability and allow the researcher to evaluate the reliability of the test. There are three main techniques for measuring the internal consistency reliability, depending upon the degree, complexity and scope of the test. They all check that the results and constructs measured by a test are correct, and the exact type used is dictated by subject, size of the data set and resources.

- *Split-Halves Test:* The split halves test for internal consistency reliability is the easiest type, and involves dividing a test into two halves. For example, a questionnaire to measure extroversion could be divided into odd and even questions. The results from both halves are statistically analyzed, and if there is weak correlation between the two, then there is a reliability problem with the test. The division of the question into two sets must be random. Split halves testing was a popular way to measure reliability, because of its simplicity and speed. However, in an age where computers can take over the laborious number crunching, scientists tend to use much more powerful tests.
- *Kuder-Richardson Test:* The Kuder-Richardson test for internal consistency reliability is a more advanced, and slightly more complex, version of the split halves test. In this version, the test works out the average correlation for all the possible split half combinations in a test. The Kuder-Richardson test also generates a correlation of between zero and one, with a more accurate result than the split halves test. The weakness of this approach, as with split-halves, is that the answer for each question

must be a simple right or wrong answer, zero or one. For multi-scale responses, sophisticated techniques are needed to measure internal consistency reliability.

- *Cronbach's Alpha Test:* The Cronbach's Alpha test not only averages the correlation between every possible combination of split halves, but it allows multi-level responses. For example, a series of questions might ask the subjects to rate their response between one and five. Cronbach's Alpha gives a score of between zero and one, with 0.7 generally accepted as a sign of acceptable reliability. The test also takes into account both the size of the sample and the number of potential responses. A 40-question test with possible ratings of 1 - 5 is seen as having more accuracy than a ten-question test with three possible levels of response. Of course, even with Cronbach's clever methodology, which makes calculation much simpler than crunching through every possible permutation, this is still a test best left to computers and statistics spreadsheet programmes.

- ***Instrument Reliability***

A researcher will always test the instrument reliability of weighing scales with a set of calibration weights, ensuring that the results given are within an acceptable margin of error. Some of the highly accurate balances can give false results if they are not placed upon a completely level surface, so this calibration process is the best way to avoid this. In the non-physical sciences, the definition of an instrument is much broader, encompassing everything from a set of survey questions to an intelligence test. A survey to measure reading ability in children must produce reliable and consistent results if it is to be taken seriously. Political opinion polls, on the other hand, are notorious for producing inaccurate results and delivering a near unworkable margin of error. In the physical sciences, it is possible to isolate a measuring instrument from external factors, such as environmental conditions and temporal factors. In the social sciences, this is much more difficult, so any instrument must be tested with a reasonable range of reliability.

Any test of instrument reliability must test how stable the test is over time, ensuring that the same test performed upon the same individual gives exactly the same results. The test-retest method is one way of ensuring that any instrument is stable over time. Of course, there is no such thing as perfection and there will be always be some disparity and potential for regression, so statistical methods are used to determine whether the stability of the instrument is within acceptable limits.

- ***Statistical Reliability***

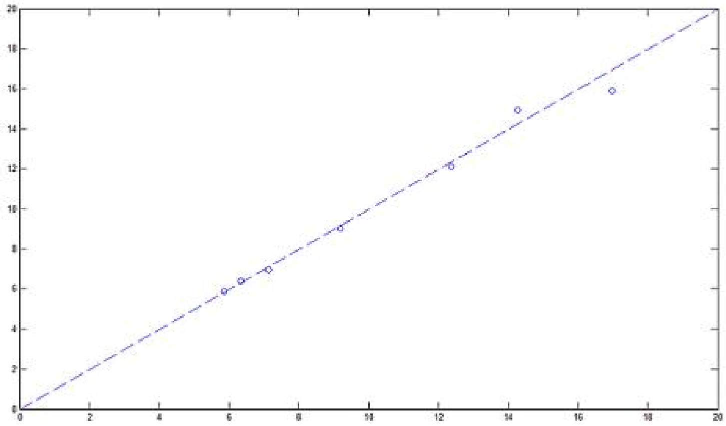
Statistical reliability is needed in order to ensure the validity and precision of the statistical analysis. It refers to the ability to reproduce the results again and again as required. This is essential as it builds trust in the statistical analysis and the results obtained. For example, suppose you are studying the effect of a new drug on the blood pressure in mice. You would want to do a number of tests and if the results are found to be good in controlling blood pressure, you might want to try it out in humans too. The statistical reliability is said to be low if you measure a certain level of control at one point and a significantly different value when you perform the experiment at another time. However, if the reliability is low, this means that the experiment that you have performed is difficult to be reproduced with

similar results then the validity of the experiment decreases. This means that people will not trust in the abilities of the drug based on the statistical results you have obtained.
In many cases, you can improve the reliability by taking in more number of tests and subjects. Simply put, reliability is a measure of consistency. Reliability can be measured and quantified using a number of methods.

Consider the previous example, where a drug is used that lowers the blood pressure in mice. Depending on various initial conditions, the following table is obtained for the percentage reduction in the blood pressure level in two tests. (Note that, this is just an illustrative example - no test has actually been conducted)

Time after injection	Test 1	Test 2
1 min	5.86	5.89
2 min	6.35	6.41
3 min	7.12	6.95
4 min	9.18	9.01
5 min	12.36	12.13
6 min	14.26	14.93
7 min	16.96	15.89

Ideally, the two tests should yield the same values, in which case the statistical reliability will be 100%. However, this doesn't happen in practice, and the results are shown in the figure below. The dotted line indicates the ideal value where the values in Test 1 and Test 2 coincide (see figure below).



Inference to Statistical Reliability: Using the above data, one can use the change in mean, study the types of errors in the experimentation including Type-I and Type-II errors or using retest correlation to quantify the reliability. The use of statistical reliability is extensive in psychological studies, and therefore there is a special way to quantify this in such cases, using Cronbach's Alpha. This gives a measure of reliability or consistency. With an increase in correlation between the items, the value of Cronbach's Alpha increases, and therefore in psychological tests and psychometric studies, this is used to study relationship between parameters and rule out chance processes.

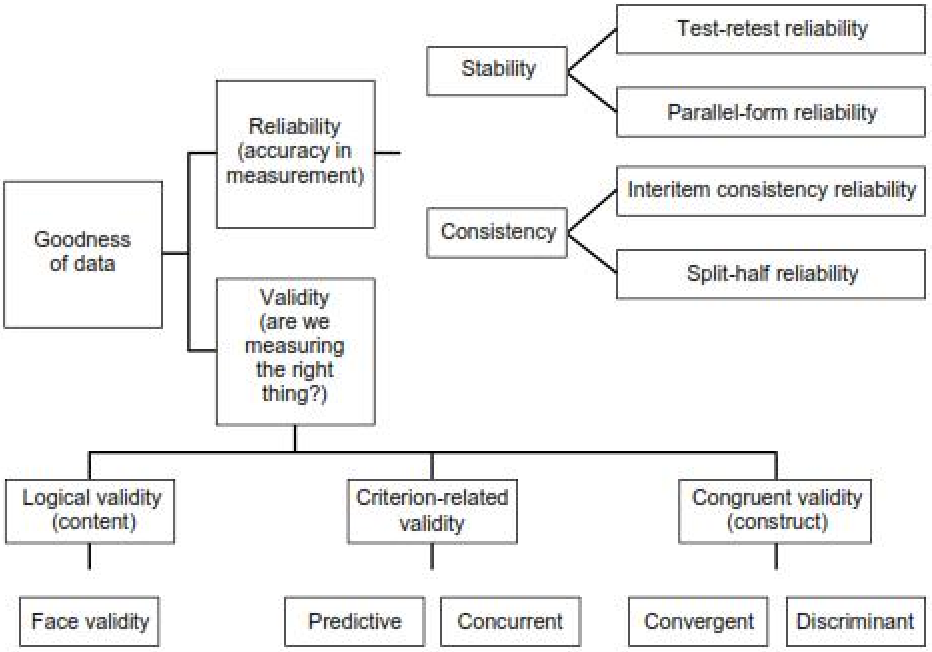
- ***Reproducibility***

  Reproducibility is regarded as one of the foundations of the entire scientific method, a benchmark upon which the reliability of an experiment can be tested. The basic principle is that, for any research program, an independent researcher should be able to replicate the experiment, under the same conditions, and achieve the same results. This gives a good guide to whether there were any inherent flaws within the experiment and ensures that the researcher paid due diligence to the process of experimental design. A replication study ensures that the researcher constructs a valid and reliable methodology and analysis.

  - *Reproducibility vs. Repeatability*

    Reproducibility is different to repeatability, where the researchers repeat their experiment to test and verify their results. Reproducibility is tested by a replication study, which must be completely independent and generate identical findings known as commensurate results. Ideally, the replication study should utilize slightly different instruments and approaches, to ensure that there was no equipment malfunction. If a type of measuring device has a design flaw, then it is likely that this artefact will be apparent in all models.

2.7 Flow-diagram for understanding Validity and Reliability


Goodness of data
Reliability (accuracy in measurement)
Stability
Test-retest reliability
Parallel-form reliability
Consistency
Interitem consistency reliability
Split-half reliability
Validity (are we measuring the right thing?)
Logical validity (content)
Face validity
Criterion-related validity
Predictive
Concurrent
Congruent validity (construct)
Convergent
Discriminant

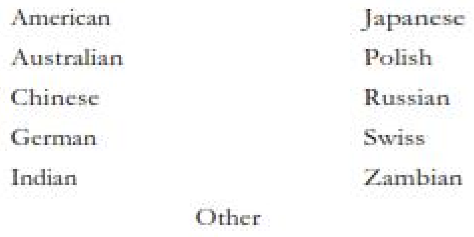
2.8 Scales

A scale is a tool or mechanism by which individuals are distinguished as to how they differ from one another on the variables of interest to our study. The scale or tool could be a gross one in the sense that it would only broadly categorize individuals on certain variables, or it could be a fine-tuned tool that would differentiate individuals on the variables with varying degrees of sophistication.

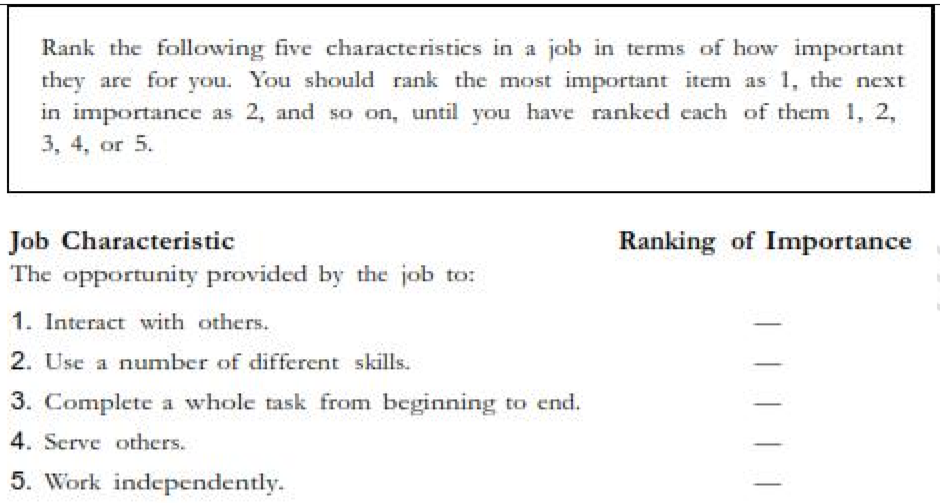
There are four basic types of scales: Nominal, Ordinal, Interval and Ratio. The degree of sophistication to which the scales are fine-tuned increases progressively as we move from the nominal to the ratio scale. That is, information on the variables can be obtained in greater detail when we employ an interval or a ratio scale than the other two scales. As the calibration or fine-tuning of the scale increases in sophistication, so does the power of the scale. With more powerful scales, increasingly sophisticated data analyses can be performed, which, in turn, means that more meaningful answers can be found to our research questions. However, certain variables lend themselves with greater ease to more powerful scaling than others.

Let us now examine each of these four scales:

Scale: Nominal	
Description	A nominal scale is one that allows the researcher to assign subjects to certain categories or groups. For example, with respect to the variable of gender, respondents can be grouped into two categories - male and female. These two groups can be assigned code numbers 1 and 2. These numbers serve as simple and convenient category labels with no intrinsic value, other than to assign respondents to one of two non-overlapping or *mutually exclusive* categories. Note that the categories are also *collectively exhaustive*. In other words, there is no third category into which respondents would normally fall. Thus, nominal scales categorize individuals or objects into mutually exclusive and collectively exhaustive groups. The information that can be generated from nominal scaling is to calculate the percentage (or frequency) of males and females in our sample of respondents. For example, if we had interviewed 200 people, and assigned code number 1 to all male respondents and number 2 to all female respondents, then computer analysis of the data at the end of the survey may show that 98 of the respondents are men and 102 are women. This frequency distribution tells us that 49% of the survey‘s respondents are men and 51% women. Other than this marginal information, such scaling tells us nothing more about the two groups. Thus the nominal scale gives some basic, categorical, gross information.
Example	Let us take a look at another variable that lends itself to nominal scaling – the nationality of individuals. We could nominally scale this variable in the following mutually exclusive and collectively exhaustive categories.

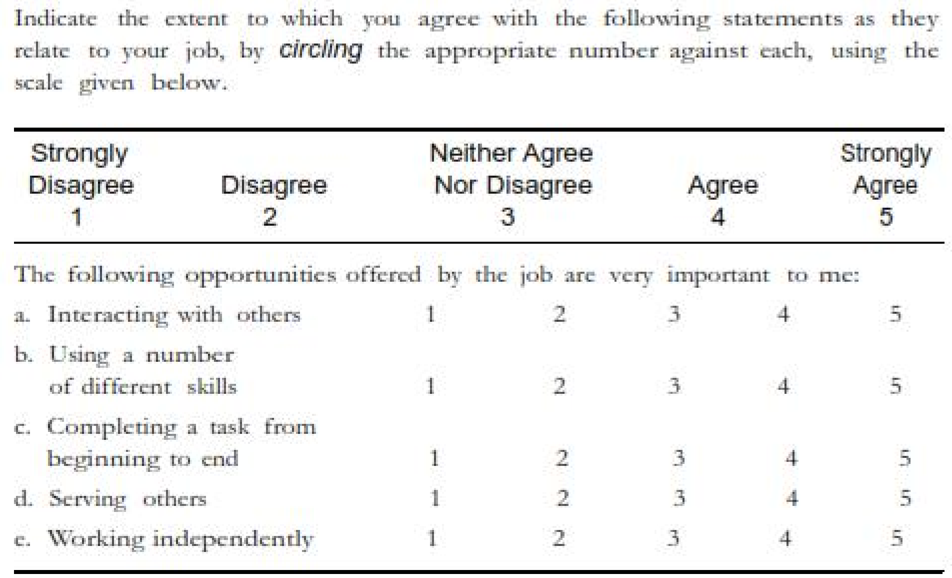
	 American Australian Chinese German Indian Japanese Polish Russian Swiss Zambian Other Note that every respondent has to fit into one of the above eleven categories and that the scale will allow computation of the numbers and percentage of respondents that fit into them.

Scale: Ordinal	
Description	An ordinal scale not only categorizes the variables in such a way as to denote differences among the various categories, it also rank-orders the categories in some meaningful way. With any variable for which the categories are to be ordered according to some preference, the ordinal scale would be used. The preference would be ranked (e.g., from best to worst; first to last) and numbered 1, 2, and so on. For example, respondents might be asked to indicate their preferences by ranking the importance they attach to five distinct characteristics in a job that the researcher might be interested in studying. Such a question might take the form as given in the example below. The ordinal scale helps the researcher to determine the percentage of respondents who consider interaction with others as most important, those who consider using a number of different skills as most important, and so on. Such knowledge might help in designing jobs that would be seen as most enriched by the majority of the employees. We can now see that the ordinal scale provides more information than the nominal scale. The ordinal scale goes beyond differentiating the categories to providing information on how respondents distinguish them by rank-ordering them. Note, however, that the ordinal scale does not give any indication of the magnitude of the differences among the ranks. For instance, in the job characteristics example, the first-ranked job characteristics might be only marginally preferred over the second-ranked characteristic, whereas the characteristic that is ranked third might be preferred in a much larger degree than the one ranked fourth. Thus, in ordinal scaling, even though differences in the ranking of objects, persons, or events investigated are clearly known, we do not know their magnitude. This deficiency is overcome by interval scaling, which is discussed next.

Example	 Rank the following five characteristics in a job in terms of how important they are for you. You should rank the most important item as 1, the next in importance as 2, and so on, until you have ranked each of them 1, 2, 3, 4, or 5. **Job Characteristic** **Ranking of Importance** The opportunity provided by the job to: 1. Interact with others. — 2. Use a number of different skills. — 3. Complete a whole task from beginning to end. — 4. Serve others. — 5. Work independently. —

Scale: Interval	
Description	An interval scale allows us to perform certain arithmetical operations on the data collected from the respondents. Whereas the nominal scale allows us only to qualitatively distinguish groups by categorizing them into mutually exclusive and collectively exhaustive sets, and the ordinal scale to rank-order the preferences, the interval scale lets us measure the distance between any two points on the scale. This helps us to compute the means and the standard deviations of the responses on the variables. In other words, the interval scale not only groups individuals according to certain categories and taps the order of these groups, it also measures the magnitude of the differences in the preferences among the individuals. If, for instance, employees think that: (1) it is more important for them to have a variety of skills in their jobs than to complete a task from beginning to end, and (2) it is more important for them to serve people than to work independently on the job, then the interval scale would indicate whether the first preference is to the same extent, a lesser extent, or a greater extent than the second. This can be done by now changing the scale from the ranking type in Example 8.5 to make it appear as if there were several points on a scale that would represent the extent or magnitude of the importance of each of the five job characteristics. Such a scale could be indicated for the job design example, as follows.

Example



Indicate the extent to which you agree with the following statements as they relate to your job, by *circling* the appropriate number against each, using the scale given below.

Strongly Disagree 1	Disagree 2	Neither Agree Nor Disagree 3	Agree 4	Strongly Agree 5

The following opportunities offered by the job are very important to me:

a. Interacting with others	1	2	3	4	5
b. Using a number of different skills	1	2	3	4	5
c. Completing a task from beginning to end	1	2	3	4	5
d. Serving others	1	2	3	4	5
e. Working independently	1	2	3	4	5

Let us illustrate how the interval scale establishes the equality of the magnitude of differences in the scale points. Let us suppose that employees circle the numbers 3, 1, 2, 4, and 5 for the five items in Example above. They then indicate to us that the extent of their preference for skill utilization over doing the task from beginning to end is the same as the extent of their preference for serving customers over working independently. That is, the magnitude of difference represented by the space between points 1 and 2 on the scale is the same as the magnitude of difference represented by the space between points 4 and 5, or between any other two points. Any number can be added to or subtracted from the numbers on the scale, still retaining the magnitude of the difference. For instance, if we add 6 to all five points on the scale, the interval scale will have the numbers 7 to 11 (instead of 1 to 5). The magnitude of the difference between 7 and 8 is still the same as the magnitude of the difference between 9 and 10. Thus, the origin, or the starting point, could be any *arbitrary number*. The clinical thermometer is a good example of an interval-scaled instrument; it has an arbitrary origin and the magnitude of the difference between 98.6 degrees (supposed to be the normal body temperature) and 99.6 degrees is the same as the magnitude of the difference between 104 and 105 degrees. Note, however, that one may not be seriously concerned if one‘s temperature rises from 98.6 to 99.6, but is likely to be so when the temperature goes up from 104 to 105 degrees! The interval scale, then, taps the differences, the order, and the equality of the magnitude of the differences in the variable. As such, it is a more powerful scale than the nominal and ordinal scales, and has for its measure of central tendency the arithmetic mean. Its measures of dispersion are the range, the standard deviation, and the variance.

Scale: Ratio

Description	The ratio scale overcomes the disadvantage of the arbitrary origin point of the interval scale, in that it has an *absolute* (in contrast to an *arbitrary*) zero point, which is a meaningful measurement point. Thus the ratio scale not only measures the magnitude of the differences between points on the scale but also taps the proportions in the differences. It is the most powerful of the four scales because it has a unique zero origin (not an arbitrary origin) and subsumes all the properties of the other three scales. The weighing balance is a good example of a ratio scale. It has an absolute (and not arbitrary) zero origin calibrated on it, which allows us to calculate the ratio of the weights of two individuals. For instance, a person weighing 250 pounds is *twice* as heavy as one who weighs 125 pounds. Note that multiplying or dividing both of these numbers (250 and 125) by any given number will preserve the ratio of 2:1. The measure of central tendency of the ratio scale could be either the arithmetic or the geometric mean and the measure of dispersion could be either the standard deviation, or variance, or the coefficient of variation.
Example	Some examples of ratio scales are those pertaining to actual age, income, and the number of organizations individuals have worked for.

2.9 Rating Scales / Scaling Techniques

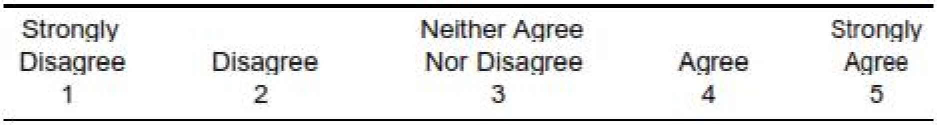
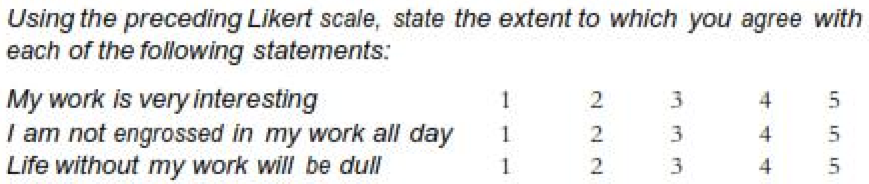
The following rating scales (scaling techniques) are often used in organizational research:
Dichotomous scale; Category scale; Likert scale; Numerical scales; Semantic differential scale; Itemized rating scale; Fixed or constant sum rating scale; Stapel scale; Graphic rating scale; Consensus scale.

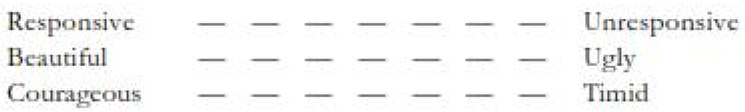
Other scales, such as, the Thurstone Equal Appearing Interval Scale, and the Multidimensional Scale are less frequently used.

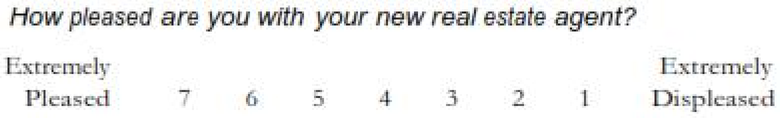
This section will briefly describe each of the above attitudinal scales.

Dichotomous Scale	
Description	The dichotomous scale is used to elicit a Yes or No answer, as in the example below. Note that a nominal scale is used to elicit the response.
Example	*Do you own a car?* Yes No

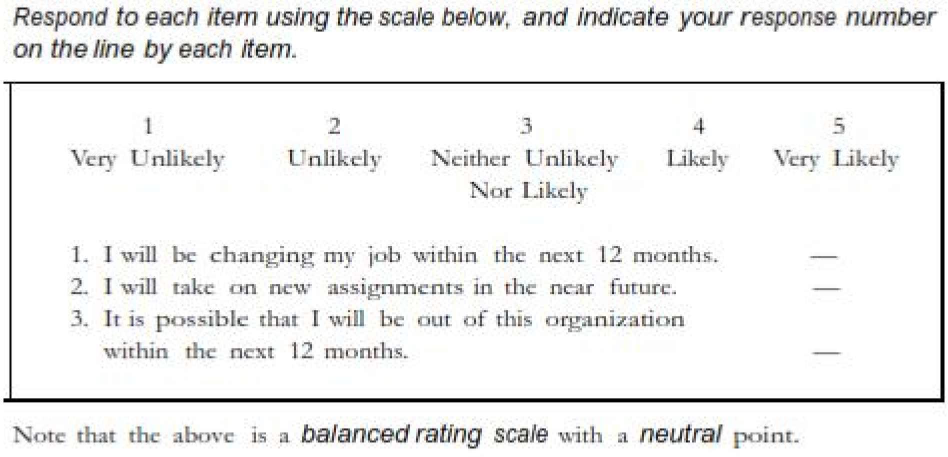
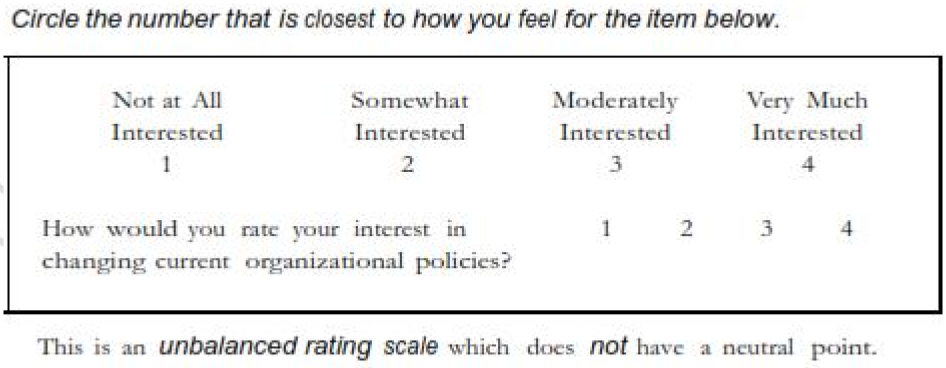
Category Scale	
Description	The category scale uses multiple items to elicit single response as per the following example. This also uses the nominal scale.
Example	Where in northern India do you reside? Himachal Pradesh Punjab Jammu & Kashmir Haryana Delhi

Likert Scale	
Description	The Likert scale is designed to examine how strongly subjects agree or disagree with statements on a 5-point scale with the following anchors:  Strongly Disagree 1 · Disagree 2 · Neither Agree Nor Disagree 3 · Agree 4 · Strongly Agree 5 The responses over a number of items tapping a particular concept or variable (as per the following example) are then summated for every respondent. This is an interval scale and the differences in the responses between any two points on the scale remain the same.
Example	 *Using the preceding Likert scale, state the extent to which you agree with each of the following statements:* *My work is very interesting* 1 2 3 4 5 *I am not engrossed in my work all day* 1 2 3 4 5 *Life without my work will be dull* 1 2 3 4 5

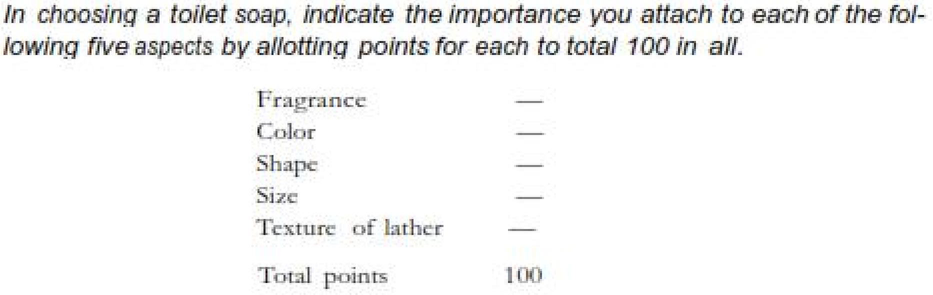
Semantic Differential Scale	
Description	Several bipolar attributes are identified at the extremes of the scale, and respondents are asked to indicate their attitudes, on what may be called a semantic space, toward a particular individual, object, or event on each of the attributes. The bipolar adjectives used, for instance, would employ such terms as Good–Bad; Strong–Weak; Hot–Cold. The semantic differential scale is used to assess respondents‗ attitudes toward a particular brand, advertisement, object, or individual. The responses can be plotted to obtain a good idea of their perceptions. This is treated as an interval scale. An example of the semantic differential scale follows.
Example	 Responsive — — — — — — — Unresponsive Beautiful — — — — — — — Ugly Courageous — — — — — — — Timid

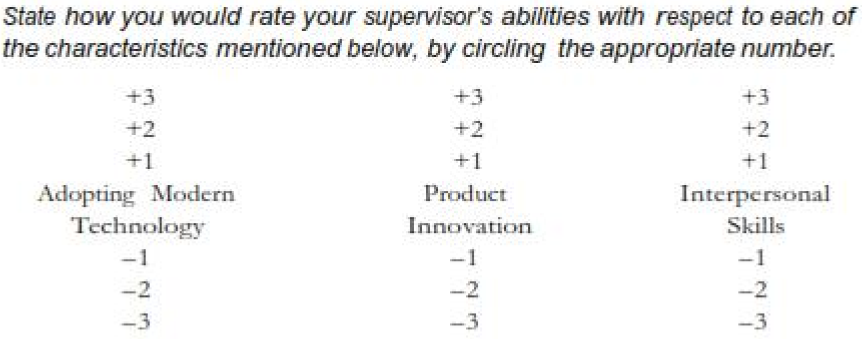
Numerical Scale	
Description	The numerical scale is similar to the semantic differential scale, with the difference that numbers on a 5-point or 7-point scale are provided, with bipolar adjectives at both ends, as illustrated below. This is also an interval scale.
Example	 *How pleased are you with your new real estate agent?* Extremely Pleased 7 6 5 4 3 2 1 Extremely Displeased

Itemized Rating Scale

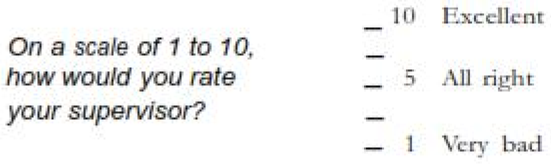
Description	A 5-point or 7-point scale with anchors, as needed, is provided for each item and the respondent states the appropriate number on the side of each item, or circles the relevant number against each item, as per the examples that follow. The responses to the items are then summated. This uses an interval scale. The itemized rating scale provides the flexibility to use as many points in the scale as considered necessary (4, 5, 7, 9, or whatever), and it is also possible to use different anchors (e.g., Very Unimportant to Very Important; Extremely Low to Extremely High). When a neutral point is provided, it is a balanced rating scale, and when it is not, it is an unbalanced rating scale. Research indicates that a 5-point scale is just as good as any, and that an increase from 5 to 7 or 9 points on a rating scale does not improve the reliability of the ratings (Elmore & Beggs, 975). The itemized rating scale is frequently used in business research, since it adapts itself to the number of points desired to be used, as well as the nomenclature of the anchors, as is considered necessary to accommodate the needs of the researcher for tapping the variable.
Example-1	 *Respond to each item using the scale below, and indicate your response number on the line by each item.* 1 Very Unlikely — 2 Unlikely — 3 Neither Unlikely Nor Likely — 4 Likely — 5 Very Likely 1. I will be changing my job within the next 12 months. — 2. I will take on new assignments in the near future. — 3. It is possible that I will be out of this organization within the next 12 months. — Note that the above is a *balanced rating scale* with a *neutral* point.
Example-2	 *Circle the number that is closest to how you feel for the item below.* Not at All Interested 1 — Somewhat Interested 2 — Moderately Interested 3 — Very Much Interested 4 How would you rate your interest in changing current organizational policies? 1 2 3 4 This is an *unbalanced rating scale* which does *not* have a neutral point.

Fixed or Constant Sum Scale	
Description	The respondents are here asked to distribute a given number of points across various items as per the example below. This is more in the nature of an ordinal scale.

Example	 *In choosing a toilet soap, indicate the importance you attach to each of the following five aspects by allotting points for each to total 100 in all.* Fragrance — Color — Shape — Size — Texture of lather — Total points 100

Stapel Scale	
Description	This scale simultaneously measures both the direction and intensity of the attitude toward the items under study. The characteristic of interest to the study is placed at the center and a numerical scale ranging, say, from + 3 to – 3, on either side of the item as illustrated below. This gives an idea of how close or distant the individual response to the stimulus is, as shown in the example below. Since this does not have an absolute zero point, this is an interval scale.
Example	 *State how you would rate your supervisor's abilities with respect to each of the characteristics mentioned below, by circling the appropriate number.* +3 +3 +3 +2 +2 +2 +1 +1 +1 Adopting Modern Technology / Product Innovation / Interpersonal Skills −1 −1 −1 −2 −2 −2 −3 −3 −3

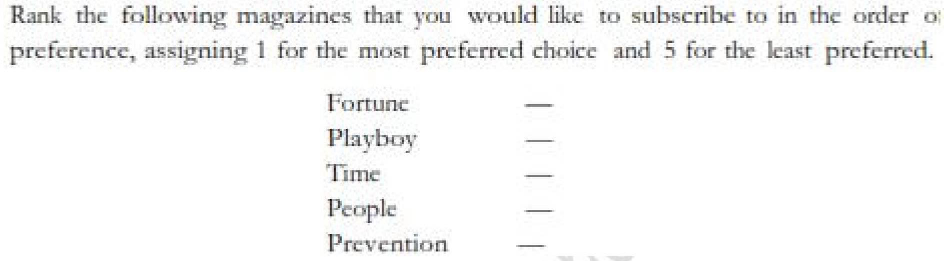
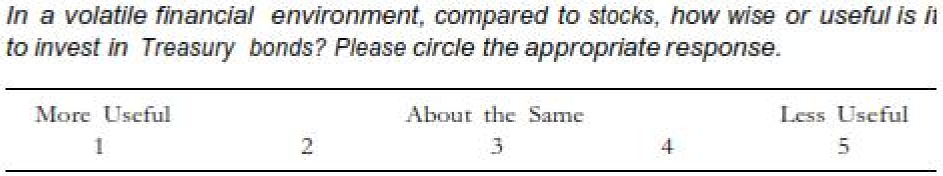
Graphic Rating Scale	
Description	A graphical representation helps the respondents to indicate on this scale their answers to a particular question by placing a mark at the appropriate point on the line, as in the following example. This is an ordinal scale, though the following example might appear to make it look like an interval scale. This scale is easy to respond to. The brief descriptions on the scale points are meant to serve as a guide in locating the rating rather than represent discrete categories. The **faces scale,** which depicts faces ranging from smiling to sad, is also a graphic rating scale. This scale is used to obtain responses regarding people‘s feelings with respect to some aspect—say, how they feel about their jobs.

Example	 On a scale of 1 to 10, how would you rate your supervisor? _ 10 Excellent _ _ 5 All right _ _ 1 Very bad

Consensus Scale	
Description	Scales are also developed by consensus, where a panel f judges selects certain items, which in its view measure the relevant concept. The items are chosen particularly based on their pertinence or relevance to the concept. Such a consensus scale is developed after the selected items are examined and tested for their validity and reliability. One such consensus scale is the **Thurstone Equal Appearing Interval Scale,** where a concept is measured by a complex process followed by a panel of judges. Using a pile of cards containing several descriptions of the concept, a panel of judges offers inputs to indicate how close or not the statements are to the concept under study. The scale is then developed based on the consensus reached. However, this scale is rarely used for measuring organizational concepts because of the time necessary to develop it.

Other Scales
There are also some advanced scaling methods such as **multidimensional scaling,** where objects, people, or both, are visually scaled, and a conjoint analysis is performed. This provides a visual image of the relationships in space among the dimensions of a construct. It is to be noted that usually the Likert or some form of numerical scale is usually the one most frequently used to measure attitudes and behaviors in organizational research.
Ranking Scales
As already mentioned, **ranking scales** are used to tap preferences between two or among more objects or items (ordinal in nature). However, such ranking may not give definitive clues to some of the answers sought. For instance, let us say there are four product lines and the manager seeks information that would help decide which product line should get the most attention. Let us also assume that 35% of the respondents choose the first product, 25% the second, and 20% choose each of products three and four as of importance to them. The manager cannot then conclude that the first product is the most preferred since 65% of the respondents did not choose that product! Alternative methods used are the *Paired comparisons*, *Forced choice*, and the *Comparative Scale*, which are discussed below.

Paired Comparison	
Description	The **paired comparison** scale is used when, among a small number of objects, respondents are asked to choose between two objects at a time. This helps to assess preferences. If, for instance, in the previous example, during the paired comparisons, respondents consistently show a preference for product one over products two, three, and four, the manager reliably understands which product

<table>
<tr><td></td><td>line demands his utmost attention. However, as the number of objects to be compared increases, so does the number of paired comparisons. The paired choices for *n* objects will be: [(*n*)*(*n*–1)/2]. The greater the number of objects or stimuli, the greater the number of paired comparisons presented to the respondents, and the greater the respondent fatigue. Hence paired comparison is a good method if the number of stimuli presented is small.</td></tr>
<tr><td colspan="2"><b>Forced Choice</b></td></tr>
<tr><td><b>Description</b></td><td>The <b>forced choice</b> enables respondents to rank objects relative to one another, among the alternatives provided. This is easier for the respondents, particularly if the number of choices to be ranked is limited in number.</td></tr>
<tr><td><b>Example</b></td><td>
Rank the following magazines that you would like to subscribe to in the order o
preference, assigning 1 for the most preferred choice and 5 for the least preferred.
Fortune —
Playboy —
Time —
People —
Prevention —</td></tr>
<tr><td colspan="2"><b>Comparative Scale</b></td></tr>
<tr><td><b>Description</b></td><td>The <b>comparative scale</b> provides a benchmark or a point of reference to assess attitudes toward the current object, event, or situation under study. An example of the use of comparative scale follows.</td></tr>
<tr><td><b>Example</b></td><td>
<i>In a volatile financial environment, compared to stocks, how wise or useful is i</i>
<i>to invest in Treasury bonds? Please circle the appropriate response.</i>
<table><tr><td>More Useful</td><td></td><td>About the Same</td><td></td><td>Less Useful</td></tr><tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr></table></td></tr>
</table>