Rating and Ranking Scale, Thurston, Likert and Semantic Differential Scaling, Paired Comparison, Reliability and Validity Scale

What is a Rating Scale?

Rating scale questions are a variation of multiple choices. They ask the respondent to assign a value to a particular object or subject. Rating scales are close-ended questions that can help you gain quantitative data – information you can measure, hard facts. Rating scales allow you to collect data in a way that is easier to analyze and use.

A Rating scale question is one that seeks respondent feedback in a comparative form for specific features, products or service – "on a scale of 1 to 7 where one means 'not at all likely' and seven means 'extremely likely,' how likely are you to purchase the product in the next 3 months?"

Say, for example, you want to determine how your customers feel about your new soda branding. Using a rating scale question, you can show them a photo of the newly branded soda bottle and ask how likely they are to choose your item: very likely, neutral, or not at all likely.

Rating scales can take on many different forms. A numerical scale gives respondents a series of numbers to choose from, each representing a different rating.

Comparison scales, on the other hand, ask respondents to compare two or more items, rating them according to a defined scale. For example, if you want to determine how the cost of your soda compares to a competing brand, you would ask respondents to determine if yours is priced higher, about the same, or lower. In this way, you are using a rating scale to compare two items.

Self Rating Scales

1. Graphic Rating Scale

The respondents rate the objects by placing a mark at the appropriate position on a line that runs from one extreme of the criterion variable to another. Example

0	1	5	7
(poor quality)	(bad quality)	(neither good nor bad)	(good quality)

BRAND 1

This is also known as continuous rating scale. The customer can occupy any position. Here one attribute is taken ex-quality of any brand of ice-cream.

Poor	good

BRAND 2

This line can be vertical or horizontal and scale points may be provided. No other indication is there on the continuous scale. A range is provided. To quantify the responses to question that "indicate your overall opinion about ice-ream Brand 2 by placing a tick mark at appropriate position on the line", we measure the physical distance between the left extreme position and the response position on the line.; the greater the distance, the more favourable is the response or attitude towards the brand.

Its limitation is that coding and analysis will require substantial amount of time, since we first have to measure the physical distances on the scale for each respondent.

2. Itemized Rating Scales

These scales are different from continuous rating scales. They have a number of brief descriptions associated with each category. They are widely used in Marketing Research. They

essentially take the form of the multiple category questions. The most common are - Likert, Sementic, Staple and Multiple Dimension. Others are - Thurston and Guttman.

A) Likert Scale: It was developed Rensis Likert. Here the respondents are asked to indicate a degree of agreement and disagreement with each of a series of statement. Each scale item has 5 response categories ranging from strongly agree and strongly disagree.

5 Strongly agree	4 Agree	3 Indifferent	2 Disagree	1 Strongly disagree

Each statement is assigned a numerical score ranging from 1 to 5. It can also be scaled as -2 to +2.

-2	-1	0	1	2

For example quality of Mother Diary ice-cream is poor then Not Good is a negative statement and Strongly Agree with this means the quality is not good.

Each degree of agreement is given a numerical score and the respondents total score is computed by summing these scores. This total score of respondent reveals the particular opinion of a person.

Likert Scale is of ordinal type, they enable one to rank attitudes, but not to measure the difference between attitudes. They take about the same amount of efforts to create as Thurston scale and are considered more discriminating and reliable because of the larger range of responses typically given in Likert scale.

A typical Likert scale has 20 - 30 statements. While designing a good Likert Scale, first a large pool of statements relevant to the measurement of attitude has to be generated and then from the pool statements, the statements which are vague and non-discriminating have to be eliminated.

Thus, likert scale is a five point scale ranging from 'strongly agreement'to 'strongly disagreement'. No judging gap is involved in this method.

B) Semantic Differential Scale

This is a seven point scale and the end points of the scale are associated with bipolar labels.

1 Unpleasant Submissive	2	3	4	5	6	7 Pleasant Dominant

Suppose we want to know personality of a particular person. We have options-

- Unpleasant/Submissive
- Pleasant/Dominant

Bi-polar means two opposite streams. Individual can score between 1 to 7 or -3 to 3. On the basis of these responses profiles are made. We can analyse for two or three products and by joining these profiles we get profile analysis. It could take any shape depending on the number of variables.

Profile Analysis

----------/----------
----------/----------
--------/----------

Mean and median are used for comparison. This scale helps to determine overall similarities and differences among objects.

When Semantic Differential Scale is used to develop an image profile, it provides a good basis for comparing images of two or more items. The big advantage of this scale is its simplicity, while producing results compared with those of the more complex scaling methods. The method is easy and fast to administer, but it is also sensitive to small differences in attitude, highly versatile, reliable and generally valid.

C) Paired Comparison Scaling

Definition: The Paired Comparison Scaling is a comparative scaling technique wherein the respondent is shown two objects at the same time and is asked to select one according to the defined criterion. The resulting data are ordinal in nature.

Paired comparison is a practical technique for comparing up to; say 10-15 items (ideas, options or criteria etc.) – i.e. too many to rank easily just by inspection, but not so many that the table size becomes unmanageable. However, if a larger comparison is necessary then you can use the same principle with computer aided methods such interpretive structural modelling.

This example matrix shows a personal choice amongst seven different fruit

	(A)Apple	(O)Orange	(M)Melon	(K)Kiwi	(B)Banana	(P)Pear	Total stars for each fruit over whole table
(C) Cherries	C ***	C *	C **	C *	C *	C *	Cherries get 9
(A)Apple		O ***	M *	A **	B *	P *	Apples get 2
(O)Orange			M *	O **	B *	P *	Oranges get 5
(M)Melon				M **	B *	M **	Melons get 6
(K)Kiwi					B *	K **	Kiwis get 2
(B)Banana						P *	Bananas get 4
(P)Pear							Pears get 3

- Arrange a matrix as show above, giving each item a unique one-letter abbreviation (e.g. O for Orange in the example).

- ❖ Mark each cell in the matrix to indicate which fruit you prefer of the two items it represents. You could also show how strong each preference is as the example illustrates. For instance, in the example

  - o ‘C ***’ means: Cherries very much preferred
  - o ‘B *’ means: Bananas slightly preferred’

- ❖ Now sum up the total number of preferences or ‘*’s each item has. For instance:

  - o There are 6 cells where Cherries are preferred (‘C’) which between them have 9‘*’s, thus Cherries get a total score of 9.
  - o Conversely there are only 2 cells where Oranges are preferred (‘O’) with 5 ‘*’s between them, so Oranges get a total score of 5.

- ❖ These total scores are shown in the right-hand column. Clearly, Cherries win by quite a wide margin, followed by Melons, Bananas and Pears.

D) **Thurston Scale**

Developed by Louis Leon Thurston in 1928, the Thurston scale was the first formal technique used to measure attitudes. At first, it was used to measure attitudes towards religion, but later it found its application in sociology and psychology.

As one of the leading scaling theorists of the times, Thurston actually came up with 3 different scales, but when we say Thurston scale, in most cases we mean the method of equal-appearing intervals, which is why the scale is often referred to as the equal-appearing interval scale.

The other two scales are based on the method of paired comparisons and the method of successive intervals but aren’t used as commonly due to the fact that they are a bit more complex to develop. Still, they rely on the same agree/disagree question type as the equal-appearing interval scale.

Basically, the Thurston scale consists of a set of statements about a certain issue, each of which has a numerical value stating how (un)favorable it is judged to be. The respondents then tick only those statements to which they agree. After they complete the survey, the mean score is calculated, indicating their attitude on the issue in question.

Ranking Scale

Ranking scales offer a different approach to gathering data—these questions ask respondents to compare items to one another, rather than rating them on a common scale. When trying to negotiate which items to remove from your dessert menu, for example, you might ask customers to rank the seven desserts you offer from their most favorite to least favorite, giving you insight into customer preferences.

Let's consider your hypothetical soda brand. Say you want to collect data on how people respond to several different branding options. You present respondents with five different images of your branding options and ask them to rank the images in order of most preferred to least preferred.

This style of question gives you insight into what your customers like or dislike and provides you with feedback of how your products (or branding, in this case) compares alongside others.

A rating scale question allows you to measure strength of response. A ranking scale question allows you to measure priority of options. Using the two in tandem can give you very powerful insights into consumer preferences.

For example, if you were to ask a member to rate five different product concepts and she rated three of the five with a score of "7," you do not know which of the three is most important. Following this question up with a ranking question of the five options would give you this critical information.

Reliability and validity

Reliability and validity are important aspects of selecting a survey instrument.

Meaning: Reliability refers to the extent that the instrument yields the same results over multiple trials.

Validity refers to the extent that the instrument measures what it was designed to measure. In research, there are three ways to approach validity and they include content validity, construct validity, and criterion-related validity.

- Content validity measures the extent to which the items that comprise the scale accurately represent or measure the information that is being assessed. Are the questions that are asked representative of the possible questions that could be asked?

- Construct validity measures what the calculated scores mean and if they can be generalized. Construct validity uses statistical analyses, such as correlations, to verify the relevance of the questions. Questions from an existing, similar instrument, that has been found reliable, can be correlated with questions from the instrument under examination to determine if construct validity is present. If the scores are highly correlated it is called convergent validity. If convergent validity exists, construct validity is supported.

- Criterion-related validity has to do with how well the scores from the instrument predict a known outcome they are expected to predict. Statistical analyses, such as correlations, are used to determine if criterion-related validity exists. Scores from the instrument in question should be correlated with an item they are known to predict. If a correlation of > .60 exists, criterion related validity exists as well.

Reliability can be assessed with the test-retest method, alternative form method, internal consistency method, the split-halves method, and inter-rater reliability.

- Test-retest is a method that administers the same instrument to the same sample at two different points in time, perhaps one year intervals. If the scores at both time periods are highly correlated, > .60, they can be considered reliable.

- The alternative form method requires two different instruments consisting of similar content. The same sample must take both instruments and the scores from both instruments must be correlated. If the correlations are high, the instrument is considered reliable.

- Internal consistency uses one instrument administered only once. The coefficient alpha (or cronbach's alpha) is used to assess the internal consistency of the item. If the alpha value is .70 or higher, the instrument is considered reliable.

- The split-halves method also requires one test administered once. The number of items in the scale is divided into halves and a correlation is taken to estimate the reliability of each half of the test. To estimate the reliability of the entire survey, the spearman-brown correction must be applied.

- Inter-rater reliability involves comparing the observations of two or more individuals and assessing the agreement of the observations. Kappa values can be calculated in this instance.