Block 7: Measurement and scaling - fundamentals, comparative and non-comparative scaling

(Activity solutions can be found at the end of the document.)

We introduce instruments of **measurement** to accurately fulfil research objectives. The concepts of **scaling** and measurement are explored. Comparative and non-comparative scaling techniques are examined, along with **reliability** and **validity**.

Learning Objectives

- Explain the characteristics of description, order, distance and origin and how they define the level of measurement in a scale
- Discuss the primary scales of measurement and differentiate nominal, ordinal, interval and ratio scales
- Classify and discuss scaling techniques as comparative and non-comparative and describe the comparative techniques of paired comparison, rank order, constant sum scaling
- Describe the non-comparative scaling techniques, distinguish between continuous and itemised rating scales and explain Likert, semantic differential and Stapel scales
- Discuss the decisions involved in constructing itemised rating scales.

Reading List

Malhotra, N.K., D. Nunan and D.F. Birks. Marketing Research: An Applied Approach. (Pearson, 2017) 5th edition [ISBN 9781292103129] Chapter 12.

7.1 Measurement and scaling - fundamentals, comparative and non-comparative scaling

For each section of *Measurement and scaling - fundamentals, comparative and non-comparative scaling*, use the LSE ELearning resources to test your knowledge with the Key terms and concepts flip cards.

Measurement and scaling

Measurement means assigning numbers or other symbols to characteristics of objects according to certain pre-specified rules. There must be a *one-to-one correspondence* between the numbers and the characteristics being measured. The rules for assigning numbers should be *standardised* and *applied uniformly*. The rules must not change over objects or time.

Scaling involves creating a continuum on which measured objects are located. For example, consider an *attitude scale* from 1 to 100. Each participant is assigned a number from 1 to 100, with 1 = extremely unfavourable, and 100 = extremely favourable. Measurement is the actual assignment of a number from 1 to 100 to each participant. Scaling is the process of placing the participants on a continuum, for example with respect to their attitude toward Formula One racing.

Primary scales of measurement

For a **(categorical) nominal scale**, the numbers serve only as labels or tags for identifying and classifying objects. When used for *identification*, there is a strict one-to-one correspondence between the numbers and the objects. The *numbers do not reflect the amount of the characteristic* possessed by the objects. The only permissible operation on the numbers in a nominal scale is *counting*. Only a *limited number of statistics*, all of which are based on frequency counts, are permissible such as percentages and the mode.

An **ordinal scale** is a *ranking scale* in which numbers are assigned to objects to indicate the relative extent to which the objects possess some characteristic. It is possible to determine *whether* an object has more or less of a characteristic than some other object, but not *how much* more or less. Any series of numbers can be assigned which *preserves the ordered relationships* between the objects. In addition to the counting operation allowable for nominal scale data, ordinal scales permit the use of statistics based on centiles such as percentiles, quartiles and the median.

For an **interval scale**, numerically equal distances on the scale represent equal values in the characteristic being measured. It *permits comparison of the differences* between objects. The location of the *zero point is not fixed*. Both the zero point and the units of measurement are arbitrary. Any positive linear transformation of the form y=a+bxy=a+bxwill preserve the properties of the scale. It is not meaningful to take ratios of scale values. Statistical techniques which may be used include all of those which can be applied to nominal and ordinal data. In addition, the arithmetic mean, standard deviation, and other statistics commonly used in market research are applicable.

A **ratio scale** possesses all the properties of the nominal, ordinal and interval scales. It has an *absolute zero point*. It is meaningful to compute ratios of scale values. Only proportionate transformations of the form y=bxy=bx, where bb is a positive constant, are allowed. *All statistical techniques can be applied* to ratio data.

Figure 12.1 of the textbook provides an illustration of the primary scales of measurement.

Scale	Basic characteristics	Common examples	Marketing examples	Descriptive statistics	Inferential statistics
Nominal	Numbers identify and classify objects	Student registration numbers, numbers on football players' shirts	Gender classification of retail outlet types	Percentages, mode	Chi-square, binomial test
Ordinal	Numbers indicate the relative positions of the objects but not the magnitude of differences between them	Rankings of the top four teams in the football World Cup	Ranking of service quality delivered by a number of shops; rank order of favourite TV programmes	Percentile, median	Rank order correlation, Friedman, ANOVA

Scale	Basic characteristics	Common examples	Marketing examples	Descriptive statistics	Inferential statistics
Interval	Differences between objects can be compared; zero point is arbitrary	Temperature (Fahrenheit, Celsius)	Attitudes, opinions, index numbers	Range, mean, standard deviation	Product moment correlations, *t*tests, ANOVA, regression, factor analysis
Ratio	Zero point is fixed; ratios of scale values can be computed	Length, weight	Age, income, costs, sales, market shares	Geometric mean, harmonic mean	Coefficient of variation

Primary scales of measurement

Comparative scales

Figure 12.2 of the textbook shows a classification of scaling techniques.

Comparative scales involve the direct comparison of stimulus objects. Comparative scale data must be interpreted in relative terms and have only ordinal or rank-order properties. In **non-comparative scales**, each object is scaled independently of the others in the stimulus set. The resulting data are generally assumed to be interval- or ratio-scaled.

Advantages:

- Small differences between stimulus objects can be detected.
- There are the same known reference points for all participants.
- Easily understood and can be applied.
- Involves fewer theoretical assumptions.
- Tends to reduce halo or carryover effects from one judgement to another.

Disadvantages:

- Ordinal nature of the data.
- Inability to generalise beyond the stimulus objects scaled.

Paired comparison scaling is the most widely-used comparative scaling technique. A participant is presented with *two objects and asked to select one* according to some criterion. The data obtained are ordinal in nature. With nnbrands, n(n−1)/2n(n−1)/2 *paired comparisons* are required. Under the **assumption of transitivity of preference**, it is possible to convert paired comparison data to a rank order. Figure 12.3 of the textbook provides an example for obtaining bottled-beer preferences using paired comparisons.

The most common method of *taste testing* is paired comparison. The consumer is asked to sample two different products and select the one with the most appealing taste. The test is done in private and a minimum of 1,000 responses is considered an *adequate sample size*. Example:

> A ***blind taste test*** for a soft drink, where imagery, self-perception and brand reputation are very important factors in the consumer's purchasing decision, may not be a good indicator of

performance in the marketplace. The introduction of New Coke illustrates this point. New Coke was heavily favoured in blind paired comparison taste tests, but its introduction was less than successful, because *image plays a major role* in the purchase of Coca-Cola.

For **rank-order scaling**, participants are presented with *several objects simultaneously* and asked to order or rank them according to some criterion. It is possible that the participant may dislike the brand ranked 1 in an absolute sense. Furthermore, rank-order scaling also results in ordinal data. Only n−1n−1 *scaling decisions* need to be made in rank-order scaling. Figure 12.4 of the textbook provides an example of the preference for film genres using rank order scaling.

For **constant sum scaling**, participants allocate a *constant sum of units*, such as 100 points, to attributes of a product to reflect their importance. If an attribute is unimportant, the participant assigns it zero points. If an attribute is twice as important as some other attribute, it receives twice as many points. The sum of all the points is 100, hence the name of the scale! Figure 12.5 of the textbook provides an example of the importance of bottled-beer attributes using a constant sum scale.

Non-comparative scales

In a **non-comparative scale**, participants evaluate only one object at a time, and for this reason non-comparative scales are often referred to as *monadic scales*. Non-comparative techniques consist of **continuous** and **itemised rating scales**.

Basic non-comparative scales

Scale	Basic characteristics	Examples	Advantages	Disadvantages
Continuous rating scale	Place a mark on a continuous line	Reaction to TV advertisements	Easy to construct	Scoring can be cumbersome unless computerised
Likert scale	Degree of agreement on a 1 (strongly disagree) to 5 (strongly agree) scale	Measurement of attitudes	Easy to construct, administer and understand	More time-consuming
Semantic differential scale	Seven-point scale with bipolar labels	Brand, product and company images	Versatile	Controversy as to whether the data are interval
Stapel scale	Unipolar 10-point scale, -5 to +5, without a neutral point (zero)	Measurement of attitudes and images	Easy to construct, administered over the phone	Confusing and difficult to apply

Figure 12.6 of the textbook provides an example of a continuous rating scale.

For an **itemised rating scale**, participants are provided with a scale which has a number or brief description associated with each category. The categories are *ordered in terms of scale position*, and the participants are required to select the specified category which best describes the object being

rated. The commonly-used itemised rating scales are the **Likert scale**, **semantic differential** and **Stapel scale**.

Figure 12.7 of the textbook provides an example of a Likert scale.

Figure 12.8 of the textbook provides an example of a semantic differential scale.

Figure 12.9 of the textbook provides an example of a Stapel scale.

Non-comparative itemised rating scale decisions

Decisions to be made:

- **The number of scale categories to use.** Although there is no single, optimal number, traditional guidelines suggest that there should be between five and nine categories.
- **Balanced versus unbalanced scale.** In general, the scale should be balanced to obtain the most objective data. Figure 12.10 of the textbook provides an example of balanced and unbalanced scales.
- **Odd or even number of categories.** If a neutral or indifferent scale response is possible from at least some of the participants, an odd number of categories should be used.
- **Forced versus unforced choice.** In situations where the participants are expected to have no opinion, the accuracy of the data may be improved by a non-forced scale.
- **Nature and degree of verbal description.** An argument can be made for labelling all or many scale categories. The category descriptions should be located as close to the response categories as possible.
- **Physical form of the scale.** A number of options should be tried and the best one selected. Figure 12.11 of the textbook provides an example of different rating scale configurations. Figure 12.12 of the textbook provides examples of some unique rating scale configurations.

Some commonly-used scales in marketing

Construct			**Scale descriptors**		
Attitude	Very bad	Bad	Neither bad nor good	Good	Very good
Importance	Not at all important	Not important	Neutral	Important	Very important
Satisfaction	Very dissatisfied	Dissatisfied	Neither dissatisfied nor satisfied	Satisfied	Very satisfied
Purchase intent	Definitely will not buy	Probably will not buy	Might or might not buy	Probably will buy	Definitely will buy

Some commonly-used scales in marketing

Construct			Scale descriptors		
Purchase frequency	Never	Rarely	Sometimes	Often	Very often

Figure 12.13 of the textbook shows the stages of development of a multi-item scale.

Measurement accuracy

The **true score model** provides a framework for understanding the **accuracy of measurement**, such that:

$$X_O = X_T + X_S + X_R$$

where:

- X_O = the observed score or measurement
- X_T = the true score of the characteristic
- X_S = the systematic error
- X_R = the random error.

Potential sources of error in measurement include:

- Other relatively stable characteristics of the individual which influence the test score, such as intelligence, social desirability and education
- Short-term or transient personal factors, such as health, emotions and fatigue
- Situational factors, such as the presence of other people, noise and distractions
- Sampling of items included in the scale: addition, deletion or changes in the scale items.
- Lack of clarity of the scale, including the instructions or the items themselves
- Mechanical factors, such as poor printing, overcrowding items in the questionnaire and poor design
- Administration of the scale, such as differences among interviewers
- Analysis factors, such as differences in scoring and statistical analysis.

Reliability can be defined as the extent to which measures are *free from random error*, X_R. If $X_R = 0$, the measure is perfectly reliable.

In **test-retest reliability**, participants are administered identical sets of scale items at two different times and the degree of similarity between the two measurements is determined. In **alternative-forms reliability**, two equivalent forms of the scale are constructed and the same participants are measured at two different times, with a different form being used each time.

Internal consistency reliability determines the extent to which different parts of a summated scale are consistent in what they indicate about the characteristic being measured. In **split-half reliability**, the items on the scale are divided into two halves and the resulting half scores are correlated. **Cronbach's alpha**, or the coefficient alpha, is the average of all possible split-half coefficients resulting from different ways of splitting the scale items. This coefficient varies from 0 to 1, and a value of 0.6 or less generally indicates unsatisfactory internal consistency reliability.

The **validity** of a scale may be defined as the extent to which differences in observed scale scores reflect true differences among objects on the characteristic being measured, rather than systematic or random error. *Perfect* **validity** requires that there be no measurement error ($X_O = X_T, X_R = 0 \; and \; X_S = 0$.

Content validity is a subjective but systematic evaluation of how well the content of a scale represents the measurement task at hand.

Criterion validity reflects whether a scale performs as expected in relation to other variables selected (criterion variables) as meaningful criteria.

Concurrent validity is assessed when the data on the scale being evaluated and on the criterion variables are collected at the same time.

Predictive validity is concerned with how well a scale can forecast a future criterion.

Construct validity addresses the question of which construct or characteristic the scale is, in fact, measuring. This includes convergent, discriminant and nomological validity.

Convergent validity is the extent to which the scale correlates positively with other measurements of the same construct.

Discriminant validity is the extent to which a measure does not correlate with other constructs from which it is supposed to differ.

Nomological validity is the extent to which the scale correlates in theoretically predicted ways with measures of different but related constructs.

If a measure is perfectly valid, it is also perfectly reliable. In this case $X_O = X_T, X_R = 0 \; and \; X_S = 0$. If a measure is unreliable, it cannot be perfectly valid, since at a minimum $X_O = X_T + X_R$ Furthermore, systematic error may also be present, i.e. $X_S \neq 0$. Therefore, *unreliability implies invalidity*. If a measure is perfectly reliable, it may or may not be perfectly valid, because systematic error may still be present ($X_O = X_T + X_S$).

Reliability is a necessary, but not sufficient, condition for validity. **Generalisability** refers to the extent to which one can generalise from the observations at hand to a universe of generalisations.

Discussion forum and case studies

To access the solutions to these questions and case study, click here to access the printable Word document or click here to go to LSE's Elearning resources.

Case study: Healthcare industry

You work in the market research department of a firm specialising in decision support systems for the health care industry. Your firm would like to measure the attitudes of hospital administrators toward decision support systems produced by your firm and its main competitors. The attitudes would be

measured using a telephone survey. You have been asked to develop an appropriate scale for this purpose. You have also been asked to explain and justify your reasoning in constructing this scale.

Case study: Renault

Design Likert scales to measure the usefulness of Renault's website.

Visit the site and rate it on the scales which you have developed. After your site visit, were there any aspects of usefulness which you had not considered in devising your scales? What were they? Why were they not apparent before you made your site visit?

Case study: Louis Vuitton Möet Hennessy

Design Likert scales to measure the usefulness of the Louis Vuitton Möet Hennessy website. Visit the site and rate it on the scales which you have developed. After your site visit, were there any aspects of usefulness which you had not considered in devising your scales, what were they and why were they not apparent before you made your site visit?

Learning outcomes checklist

Use this to assess your own understanding of the chapter. You can always go back and amend the checklist when it comes to revision!

- Discuss the concepts of product moment correlation and the partial correlation coefficient and show how they provide a foundation for regression analysis
- Explain the nature and methods of bivariate regression analysis and describe the general model, estimation of parameters, standardised regression coefficient, significance testing, prediction accuracy, residual analysis and model cross-validation
- Explain the nature and methods of multiple regression analysis and the meaning of partial regression coefficients
- Describe specialised techniques used in multiple regression analysis, particularly stepwise regression and regression with dummy variables.

Block 7: Measurement and scaling - fundamentals, comparative and non-comparative scaling

Commentary on Case Study: Healthcare industry

An itemised rating scale would need to be used since no comparison to a particular decision support system (DSS) is being made and the telephone mode of data collection rules out a continuous rating scale's use. In this instance, a Likert scale is most suitable since the procedure is easily understood over the telephone, statements about various aspects of DSS usage can be constructed easily, and evaluating results is straightforward.

Commentary on Case study: Renault

Likert scales can be developed to measure the usefulness of Renault's website. Scale items should include availability of information, visual search, price information availability, ease of navigation, dealer information and linkages to dealers and other relevant sites, service to customers, user groups and links to user groups and the overall visual attractiveness of the site.

Commentary on Case study: Louis Vuitton Möet Hennessy

Likert scales can be developed to measure the usefulness of Louis Vuitton Möet Hennessy's website. Scale items should include availability of information, visual search, information related to brands in the LVMH portfolio, ease of navigation and linkages to other relevant sites, service to customers, user groups and links to user groups and the visual attractiveness of the site.