Source: © Getty Images

12

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

> *When you can measure what you are speaking about and express it in numbers, you know something about it.*
>
> Lord Kelvin

Objectives

After reading this chapter, you should be able to:

1 introduce the concepts of measurement and scaling and show how scaling may be considered an extension of measurement;

2 discuss the primary scales of measurement and differentiate nominal, ordinal, interval and ratio scales;

3 classify and discuss scaling techniques as comparative and non-comparative and describe the comparative techniques of paired comparison, rank order, constant sum and Q-sort scaling;

4 explain the concept of verbal protocols and discuss how they could be employed to measure consumer response to advertising;

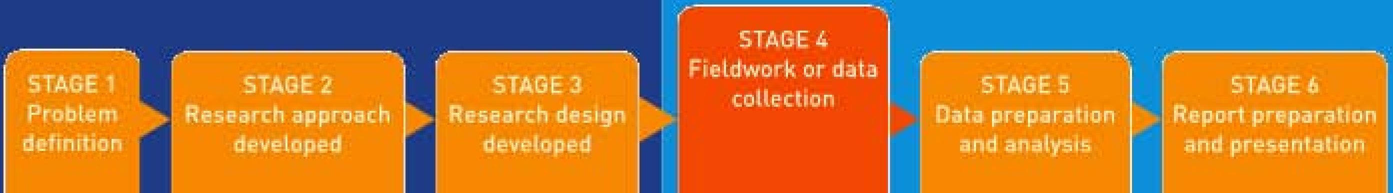
5 describe the non-comparative scaling techniques, distinguish between continuous and itemised rating scales, and explain Likert, semantic differential and Stapel scales;

6 discuss the decisions involved in constructing itemised rating scales;

7 discuss the criteria used for scale evaluation and explain how to assess reliability, validity and generalisability;

8 discuss the considerations involved in implementing the primary scales of measurement in an international setting;

9 understand the ethical issues involved in selecting scales of measurement.



STAGE 1 Problem definition

STAGE 2 Research approach developed

STAGE 3 Research design developed

STAGE 4 Fieldwork or data collection

STAGE 5 Data preparation and analysis

STAGE 6 Report preparation and presentation

Overview

Once marketing researchers have a clear understanding of what they wish to understand in their target respondents, they should consider the concepts of scaling and measurement. These concepts are vital in developing questionnaires or 'instruments of measurement' that will fulfil their research objectives in the most accurate manner. In this chapter we describes the concepts of scaling and measurement and discusses four primary scales of measurement: nominal, ordinal, interval and ratio. We describe and illustrate both comparative and non-comparative scaling techniques in detail. The comparative techniques, consisting of paired comparison, rank order, constant sum and Q-sort scaling, are discussed and illustrated with examples. The non-comparative techniques are composed of continuous and itemised rating scales. We discuss and illustrate the popular itemised rating scales – the Likert, semantic differential and Stapel scales – as well as the construction of multi-item rating scales. We show how scaling techniques should be evaluated in terms of reliability and validity and consider how the researcher selects a particular scaling technique. Mathematically derived scales are also presented. The considerations involved in implementing scaling techniques when researching international markets are discussed. The chapter concludes with a discussion of several ethical issues that arise in scale construction. We begin with an example of how the use of different types of scale can give quite different powers of analysis and interpretation.

Example

Numbers, rankings and ratings: Brazil is on top

According to the international football federation (FIFA) (**www.fifa.com**) post-2006 world cup rankings, the 2002 world champions Brazil maintained their supremacy at the top of the rankings with 1,630 points and the champions Italy took second spot with 1,550 points. The top 10 countries were as follows:

Number	Country	July 2006 ranking	Points
1	Argentina	3	1,472
2	Brazil	1	1.630
3	Czech Republic	10	1,223
4	England	5	1,434
5	France	4	1,462
6	Germany	9	1,229
7	Italy	2	1,550
8	Netherlands	6	1,332
9	Portugal	8	1,301
10	Spain	7	1,309

Note that the countries have been placed in alphabetical order and that at first glance this gives the impression that South American countries have performed better than European countries. An alphabetical order is used to illustrate the first column 'Number'. The 'number' assigned to denote countries is not in any way related to their football-playing capabilities but simply serves the purpose of identification, e.g. drawing numbered balls to decide which teams may play each other in a competition. This identification number constitutes a nominal scale, which says nothing about the respective performances of the

→

countries. So whilst Germany is numbered 6 and the Netherlands is numbered 8, this does not reflect the superior performance of the Netherlands.

A much clearer way to present the list would be to place the countries in the order of their ranking, with Brazil at the top and the Czech Republic at the bottom of the table. The ranking would represent an ordinal scale, where it would be clear to see that the lower the number, the better the performance. But what is still missing from the ranking is the magnitude of differences between the countries.

The only way to really understand how much one country is better than another is to examine the points awarded to each country. The points awarded represent an interval scale. Based on the points awarded, note that only ten points separates the closely ranked Argentina (1,472) and France (1,462), or eight points between Spain (1,309) and Portugal (1,301), but that the difference between Brazil (1,630) ranked at number 1 and Italy (1,550) ranked at number 2 is 80 points.

Measurement and scaling

Measurement
The assignment of numbers or other symbols to characteristics of objects according to certain pre-specified rules.

Measurement means assigning numbers or other symbols to characteristics of objects according to certain pre-specified rules.[1] We measure not the object but some characteristic of it. Thus, we do not measure consumers, only their perceptions, attitudes, preferences or other relevant characteristics. In marketing research, numbers are usually assigned for one of two reasons. First, numbers permit statistical analysis of the resulting data. Second, numbers facilitate universal communication of measurement rules and results.

The most important aspect of measurement is the specification of rules for assigning numbers to the characteristics. The assignment process must be isomorphic, i.e. there must be one-to-one correspondence between the numbers and the characteristics being measured. For example, the same euro (€) figures can be assigned to households with identical annual incomes. Only then can the numbers be associated with specific characteristics of the measured object, and vice versa. In addition, the rules for assigning numbers should be standardised and applied uniformly. They must not change over objects or time.

Scaling
The generation of a continuum upon which measured objects are located.

Scaling may be considered an extension of measurement. Scaling involves creating a continuum upon which measured objects are located. To illustrate, consider a scale for locating consumers according to the characteristic 'attitude towards Formula One racing'. Each respondent is assigned a number indicating an unfavourable attitude (measured as 1), a neutral attitude (measured as 2) or a favourable attitude (measured as 3). Measurement is the actual assignment of 1, 2 or 3 to each respondent. Scaling is the process of placing the respondents on a continuum with respect to their attitude towards Formula One. In our example, scaling is the process by which respondents would be classified as having an unfavourable, neutral or positive attitude.

Primary scales of measurement

There are four primary scales of measurement: nominal, ordinal, interval and ratio.[2] These scales are illustrated in Figure 12.1, and their properties are summarised in Table 12.1 and discussed in the following sections.

Nominal scale
A scale whose numbers serve only as labels or tags for identifying and classifying objects with a strict one-to-one correspondence between the numbers and the objects.

Nominal scale

A **nominal scale** is a figurative labelling scheme in which the numbers serve only as labels or tags for identifying and classifying objects. For example, the numbers assigned to the respondents in a study constitute a nominal scale; thus a female respondent may be

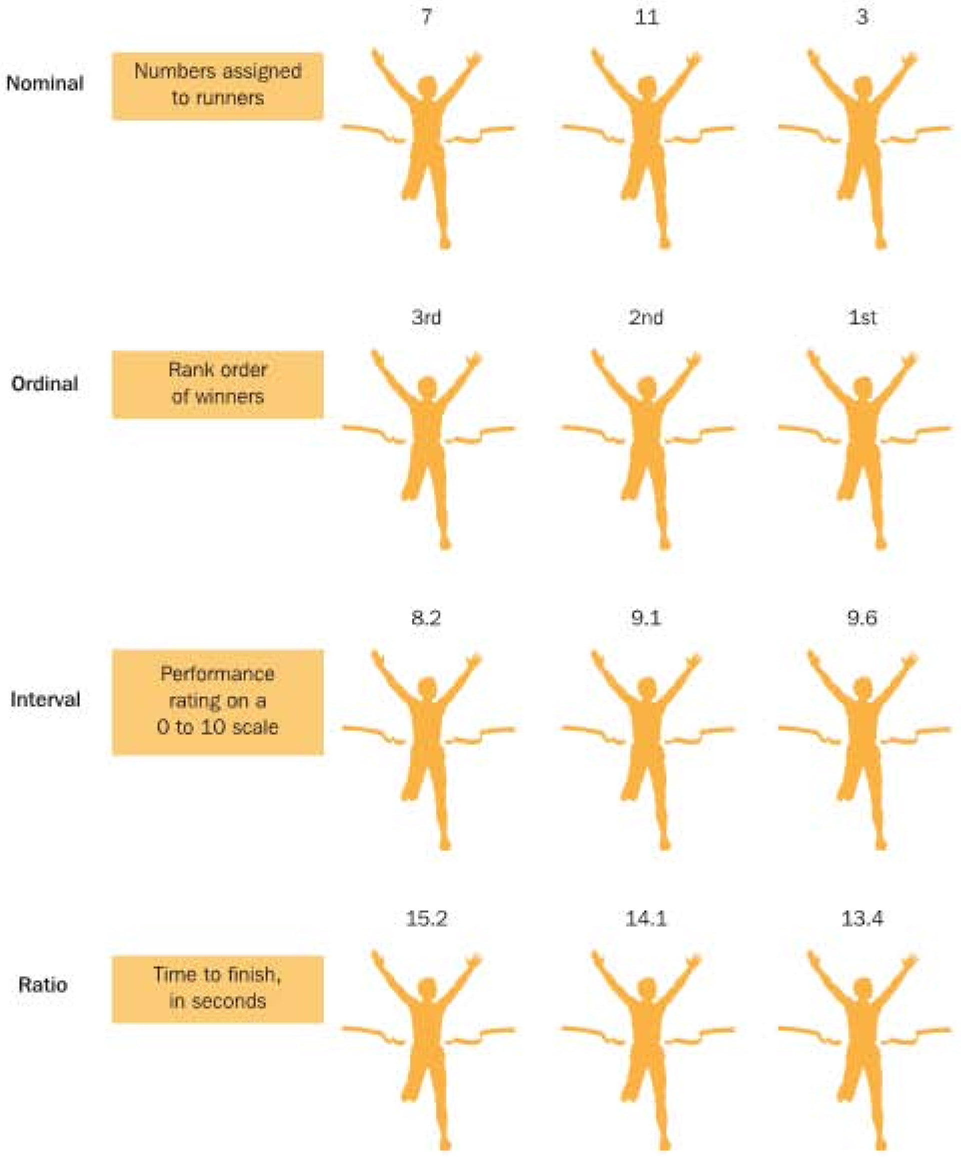

7
11
3
Nominal
Numbers assigned to runners
3rd
2nd
1st
Ordinal
Rank order of winners
8.2
9.1
9.6
Interval
Performance rating on a 0 to 10 scale
15.2
14.1
13.4
Ratio
Time to finish, in seconds

Figure 12.1 An illustration of primary scales of measurement

assigned a number 1 and a male respondent 2. When a nominal scale is used for the purpose of identification, there is a strict one-to-one correspondence between the numbers and the objects. Each number is assigned to only one object, and each object has only one number assigned to it.

Common examples include student registration numbers at their college or university and numbers assigned to football players or jockeys in a horse race. In marketing research, nominal scales are used for identifying respondents, brands, attributes, banks and other objects.

When used for classification purposes, the nominally scaled numbers serve as labels for classes or categories. For example, you might classify the control group as group 1 and the experimental group as group 2. The classes are mutually exclusive and collectively exhaustive. The objects in each class are viewed as equivalent with respect to the characteristic represented by the nominal number. All objects in the same class have the same number, and no two classes have the same number.

The numbers in a nominal scale do not reflect the amount of the characteristic possessed by the objects. For example, a high number on a football player's shirt does not

Table 12.1 Primary scales of measurement

Scale	Basic characteristics	Common examples	Marketing example		Permissible statistics
				Descriptive	Inferential
Nominal	Numbers identify and classify objects	Student registration numbers, numbers on football players' shirts	Gender classification, bank 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 banks. Rank order of favourite TV programmes	Percentile, median	Rank-order correlation, Friedman ANOVA
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

imply that the footballer is a better player than one with a low number or vice versa. The same applies to numbers assigned to classes. 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. These include percentages, mode, chi-square and binomial tests (see Chapter 18). It is not meaningful to compute an average student registration number, the average gender of respondents in a survey, or the number assigned to an average Formula One team, as in the example opposite.

Ordinal scale

Ordinal scale
A ranking scale in which numbers are assigned to objects to indicate the relative extent to which some characteristic is possessed. Thus, it is possible to determine whether an object has more or less of a characteristic than some other object.

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. An ordinal scale allows you to determine whether an object has more or less of a characteristic than some other object, but not how much more or less. Thus, an ordinal scale indicates relative position, not the magnitude of the differences between the objects. The object ranked first has more of the characteristic as compared with the object ranked second, but whether the object ranked second is a close second or a poor second is not known. Common examples of ordinal scales include quality rankings, rankings of teams in a tournament and occupational status. In marketing research, ordinal scales are used to measure relative attitudes, opinions, perceptions and preferences. Measurements of this type include 'greater than' or 'less than' judgements from respondents.

In an ordinal scale, as in a nominal scale, equivalent objects receive the same rank. Any series of numbers can be assigned that preserves the ordered relationships between the objects. Ordinal scales can be transformed in any way as long as the basic ordering of the objects is maintained.[3] In other words, any monotonic positive (order-preserving) transformation of the scale is permissible, since the differences in numbers are void of any meaning other than order (see the example opposite). For these reasons, in addition

Sports Marketing Surveys

Nominal scale

In the Racetrack study of the Formula One, the numbers 1 through to 10 were assigned to the racing teams (see extracts from the list in Table 12.2). Thus, team 2 referred to Jaguar. It did not imply that Jaguar was in any way superior or inferior to Williams, which was assigned the number 10. Any reassignment of the numbers, such as transposing the numbers assigned to Jaguar and Williams, would have no effect on the numbering system, because the numerals did not reflect any characteristics of the teams. It is meaningful to make statements such as '40% of French respondents named Ferrari as their favourite team'. Although the average of the assigned numbers is 5.5, it is not meaningful to state that the number of the average Formula One team is 5.5.

Table 12.2 Illustration of primary scales of measurement

No.	Nominal scale	Ordinal scale		Interval scale		Ratio scale
	Sponsor	Preference rankings		Preference ratings		Amount (€) spent on merchandise on this team in the past three months
				1–7	11–17	
1	BAR	5	53	5	15	35
2	Ferrari	1	10	7	17	200
3	Jaguar	6	61	5	15	100
4	Jordan	8	82	4	14	0
5	McLaren	2	25	7	17	200
6	Minardi	9	95	4	14	0
7	Renault	3	30	6	16	100
8	Sauber	10	115	2	12	10
9	Toyota	7	79	5	15	0
10	Williams	4	45	6	16	0

Sports Marketing Surveys

Ordinal scale

Table 12.2 gives a particular respondent's preference rankings. Respondents ranked the teams in order of who they preferred, by assigning a rank 1 to the first, rank 2 to the second, and so on. Note that Ferrari (ranked 1) is preferred to McLaren (ranked 2), but how much it is preferred we do not know. Also, it is not necessary that we assign numbers from 1 to 10 to obtain a preference ranking. The second ordinal scale, which assigns a number 10 to Ferrari, 25 to McLaren and 30 to Renault, is an equivalent scale, as it was obtained by a monotonic positive transformation of the first scale. The two scales result in the same ordering of the teams according to preference.

Source: © Alamy

to the counting operation allowable for nominal scale data, ordinal scales permit the use of statistics based on centiles. It is meaningful to calculate percentile, quartile, median (Chapter 18), rank order correlation (Chapter 20) or other summary statistics from ordinal data.

Interval scale

Interval scale
A scale in which the numbers are used to rank objects such that numerically equal distances on the scale represent equal distances in the characteristic being measured.

In an **interval scale**, numerically equal distances on the scale represent equal values in the characteristic being measured. An interval scale contains all the information of an ordinal scale, but it also allows you to compare the differences between objects. The difference between any two scale values is identical to the difference between any other two adjacent values of an interval scale. There is a constant or equal interval between scale values. The difference between 1 and 2 is the same as the difference between 2 and 3, which is the same as the difference between 5 and 6. A common example in everyday life is a temperature scale. In marketing research, attitudinal data obtained from rating scales are often treated as interval data.[4]

In an interval scale, the location of the zero point is not fixed. Both the zero point and the units of measurement are arbitrary. Hence, any positive linear transformation of the form $y = a + bx$ will preserve the properties of the scale. Here, x is the original scale value, y is the transformed scale value, b is a positive constant, and a is any constant. Therefore, two interval scales that rate objects A, B, C and D as 1, 2, 3 and 4 or as 22, 24, 26 and 28 are equivalent. Note that the latter scale can be derived from the former by using a = 20 and $b = 2$ in the transforming equation.

Because the zero point is not fixed, it is not meaningful to take ratios of scale values. As can be seen, the ratio of D to B values changes from 2:1 to 7:6 when the scale is transformed. Yet, ratios of differences between scale values are permissible. In this process, the constants a and b in the transforming equation drop out in the computations. The ratio of the difference between D and B values to the difference between C and B values is 2:1 in both the scales.

Statistical techniques that may be used on interval scale data include all those that can be applied to nominal and ordinal data in addition to the arithmetic mean, standard deviation (Chapter 18), product moment correlations (Chapter 20), and other statistics commonly used in marketing research. Certain specialised statistics such as geometric mean, harmonic mean and coefficient of variation, however, are not meaningful on interval scale data. The Sports Marketing Surveys example gives a further illustration of an interval scale.

Focus on

Sports Marketing Surveys

Interval scale

In Table 12.2, a respondent's preferences for the 10 teams are expressed on a seven-point rating scale (where a higher number represents a greater preference for a team). We can see that although Williams received a preference rating of 6 and Sauber a rating of 2, this does not mean that Williams is preferred three times as much as Sauber. When the ratings are transformed to an equivalent 11-to-17 scale (next column), the ratings for those teams become 16 and 12, and the ratio is no longer 3 to 1. In contrast, the ratios of preference differences are identical on the two scales. The ratio of preference difference between Ferrari and Sauber to the preference difference between BAR and Sauber is 5 to 3 on both the scales.

Ratio scale

Ratio scale
The highest scale. This scale allows the researcher to identify or classify objects, rank order the objects, and compare intervals or differences. It is also meaningful to compute ratios of scale values.

A **ratio scale** possesses all the properties of the nominal, ordinal and interval scales, and, in addition, an absolute zero point. Thus, in ratio scales we can identify or classify objects, rank the objects, and compare intervals or differences. It is also meaningful to compute ratios of scale values. Not only is the difference between 2 and 5 the same as the difference between 14 and 17, but also 14 is seven times as large as 2 in an absolute sense. Common examples of ratio scales include height, weight, age and money. In marketing, sales, costs, market share and number of customers are variables measured on a ratio scale.

Ratio scales allow only proportionate transformations of the form $y = bx$, where b is a positive constant. One cannot add an arbitrary constant, as in the case of an interval scale. An example of this transformation is provided by the conversion of metres to yards ($b = 1.094$). The comparisons between the objects are identical whether made in metres or yards.

All statistical techniques can be applied to ratio data. These include specialised statistics such as geometric mean, harmonic mean and coefficient of variation. The ratio scale is further illustrated in the context of a Sports Marketing Surveys example.

Focus on

Sports Marketing Surveys

Ratio scale

In the ratio scale illustrated in Table 12.2, a respondent is asked to indicate how much had been spent on team merchandise in the last three months. Note that this respondent spent €200 on Ferrari merchandise and only €10 on Sauber. The respondent spent 20 times more euros on Ferrari compared with Sauber. Also, the zero point is fixed because 0 means that the respondent did not spend any money on teams such as Jordan and Minardi. Multiplying these numbers by 100 to convert euros to cents results in an equivalent scale.

Metric scale
A scale that is either interval or ratio in nature.

The four primary scales discussed above do not exhaust the measurement-level categories. It is possible to construct a nominal scale that provides partial information on order (the partially ordered scale). Likewise, an ordinal scale can convey partial information on distance, as in the case of an ordered **metric scale.** A discussion of these scales is beyond the scope of this text.[5]

A comparison of scaling techniques

The scaling techniques commonly employed in marketing research can be classified into comparative and non-comparative scales (see Figure 12.2).

Comparative scales
One of two types of scaling techniques in which there is direct comparison of stimulus objects with one another.

Non-metric scale
A scale that is either nominal or ordinal in nature.

Comparative scales involve the direct comparison of stimulus objects. For example, respondents may be asked whether they prefer Coke or Pepsi. Comparative scale data must be interpreted in relative terms and have only ordinal or rank order properties. For this reason, comparative scaling is also referred to as **non-metric scaling**. As shown in Figure 12.2, comparative scales include paired comparisons, rank order, constant sum scales, Q-sort and other procedures.

Carryover effects
Where the evaluation of a particular scaled item significantly affects the respondent's judgement of subsequent scaled items.

The major benefit of comparative scaling is that small differences between stimulus objects can be detected. As they compare the stimulus objects, respondents are forced to choose between them. In addition, respondents approach the rating task from the same known reference points. Consequently, comparative scales are easily understood and can be applied easily. Other advantages of these scales are that they involve fewer theoretical assumptions, and they also tend to reduce halo or **carryover effects** from one judgement to

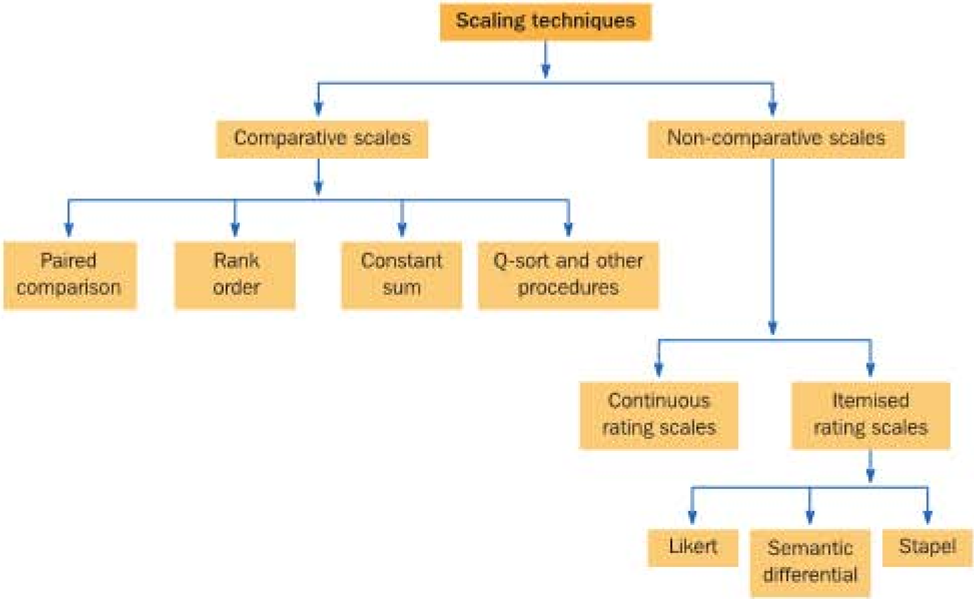


Figure 12.2
A classification of scaling techniques

another.[6] The major disadvantages of comparative scales include the ordinal nature of the data and the inability to generalise beyond the stimulus objects scaled. For instance, to compare Virgin Cola with Coke and Pepsi the researcher would have to do a new study. These disadvantages are substantially overcome by the non-comparative scaling techniques.

Non-comparative scale
One of two types of scaling techniques in which each stimulus object is scaled independently of the other objects in the stimulus set. Also called monadic scale.

In **non-comparative scales**, also referred to as monadic or metric 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.[7] For example, respondents may be asked to evaluate Coke on a 1 to 6 preference scale (1 = not at all preferred, 6 = greatly preferred). Similar evaluations would be obtained for Pepsi and Virgin Cola. As can be seen in Figure 12.2, non-comparative scales can be continuous rating or itemised rating scales. The itemised rating scales can be further classified as Likert, semantic differential or Stapel scales. Non-comparative scaling is the most widely used scaling technique in marketing research.

Comparative scaling techniques

Paired comparison scaling

Paired comparison scaling
A comparative scaling technique in which a respondent is presented with two objects at a time and asked to select one object in the pair according to some criterion. The data obtained are ordinal in nature.

As its name implies, in **paired comparison scaling** a respondent is presented with two objects and asked to select one according to some criterion.[8] The data obtained are ordinal in nature. A respondent may state that he or she prefers Belgian chocolate to Swiss, likes Kellogg's cereals better than supermarket home brands, or likes Adidas more than Nike. Paired comparison scales are frequently used when the stimulus objects are physical products. Coca-Cola is reported to have conducted more than 190,000 paired comparisons before introducing New Coke.[9] Paired comparison scaling is the most widely used comparative scaling technique.

Figure 12.3 shows paired comparison data obtained to assess a respondent's bottled beer preferences. As can be seen, this respondent made 10 comparisons to evaluate five brands. In general, with n brands, $[n(n-1)/2]$ paired comparisons include all possible pairings of objects.[10]

Paired comparison data can be analysed in several ways.[11] The researcher can calculate the percentage of respondents who prefer one stimulus over another by summing the

Instructions

We are going to present you with ten pairs of bottled beer brands. For each pair, please indicate which of the two brands of beer in the pair you prefer.

Recording form

	Holsten	Stella Artois	Grolsch	Carlsberg	Budvar
Holsten		0	0	1	0
Stella Artois	1[a]		0	1	0
Grolsch	1	1		1	1
Carlsberg	0	0	0		0
Budvar	1	1	0	1	
Number of times preferred[b]	3	2	0	4	1

[a] 1 in a particular box means that the brand in that column was preferred over the brand in the corresponding row. 0 means that the row brand was preferred over the column brand.

[b] The number of times a brand was preferred is obtained by summing the 1s in each column.

Figure 12.3
Obtaining bottled beer preferences using paired comparisons

matrices of Figure 12.3 for all the respondents, dividing the sum by the number of respondents, and multiplying by 100. Simultaneous evaluation of all the stimulus objects is also possible. Under the assumption of transitivity, it is possible to convert paired comparison data to a rank order.

Transitivity of preference
An assumption made to convert paired comparison data to rank order data. It implies that if Brand A is preferred to Brand B, and Brand B is preferred to Brand C, then Brand A is preferred to Brand C.

Transitivity of preference implies that if Brand A is preferred to B, and Brand B is preferred to C, then Brand A is preferred to C. To arrive at a rank order, the researcher determines the number of times each brand is preferred by summing the column entries in Figure 12.3. Therefore, this respondent's order of preference, from most to least preferred, is Carlsberg, Holsten, Stella Artois, Budvar and Grolsch. It is also possible to derive an interval scale from paired comparison data using the Thurstone case V procedure. Refer to the appropriate literature for a discussion of this procedure.[12]

Several modifications of the paired comparison technique have been suggested. One involves the inclusion of a neutral/no difference/no opinion response. Another extension is graded paired comparisons. In this method, respondents are asked which brand in the pair is preferred and how much it is preferred. The degree of preference may be expressed by how much more the respondent is willing to pay for the preferred brand. The resulting scale is a euro metric scale. Another modification of paired comparison scaling is widely used in obtaining similarity judgements in multidimensional scaling (see Chapter 24).

Paired comparison scaling is useful when the number of brands is limited, since it requires direct comparison and overt choice. With a large number of brands, however, the number of comparisons becomes unwieldy. Other disadvantages are that violations of the assumption of transitivity may occur, and the order in which the objects are presented may bias the results.[13] Paired comparisons bear little resemblance to the marketplace situation, which involves selection from multiple alternatives. Also respondents may prefer one object over certain others, but they may not like it in an absolute sense.

Example

Paired comparison scaling[14]

The most common method of taste testing is paired comparison. Respondents are asked to taste two different products and say which one they prefer. The test is done in private, either in respondents' homes or some other location such as a hotel suite near to a shopping centre. In these tests a minimum of 1,000 responses is considered an adequate sample.

Ocean Spray (**www.oceanspray.com**), the producer of bottled and canned juices/juice drinks, makes extensive use of taste tests in developing new products. Respondents are asked to sample its new drinks which are presented in pairs. They are evaluated on taste and aspects of flavour and then respondents choose the one they like more than the other. Taste tests showed that a segment of consumers preferred white cranberries to the strong tart taste of red cranberries. Therefore in early 2002, Ocean Spray added White Cranberry drinks, made with natural white cranberries harvested a few weeks earlier than the red variety, and Juice Spritzers, lightly carbonated juice drinks, to its product line.

Rank order scaling

Rank order scaling
A comparative scaling technique in which respondents are presented with several objects simultaneously and asked to order or rank them according to some criterion.

After paired comparisons, the most popular comparative scaling technique is **rank order scaling**. In rank order scaling respondents are presented with several objects simultaneously and asked to order or rank them according to some criterion. For example, respondents may be asked to rank according to overall preference. As shown in Figure 12.4, these rankings are typically obtained by asking the respondents to assign a rank of 1 to the most preferred Fomula One teams, 2 to the second most preferred, and so on, until a rank of *n* is assigned to the least preferred team. Like paired comparison, this approach is also comparative in nature, and it is possible that the respondents may dislike the team ranked 1 in an absolute sense. Furthermore, rank order scaling also results in ordinal data. See Table 12.2, which uses rank order scaling to derive an ordinal scale.

Rank order scaling is commonly used to measure attributes of products and services as well as preferences for brands. Rank order data are frequently obtained from respondents in conjoint analysis (see Chapter 24), since rank order scaling forces the respondents to discriminate among the stimulus objects. Moreover, compared with paired comparisons, this type of scaling process more closely resembles the shopping environment. It also takes less time and eliminates intransitive responses. If there are *n* stimulus objects, only $(n - 1)$ scaling decisions need be made in rank order scaling. However, in paired comparison scaling, $[n(n - 1)/2]$ decisions would be required. Another advantage is that most respondents

Instructions

Rank the listed Formula One teams in order of preference. Begin by picking out the team that you like most and assign it a number 1. Then find the second most preferred team and assign it a number 2. Continue this procedure until you have ranked all the teams in order of preference. The least preferred team should be assigned a rank of 10.

No two teams should receive the same rank number.

The criterion of preference is entirely up to you. There is no right or wrong answer. Just try to be consistent.

	Brand	*Rank order*
1	BAR	
2	Ferrari	
3	Jaguar	
4	Jordan	
5	McLaren	
6	Minardi	
7	Renault	
8	Sauber	
9	Toyota	
10	Williams	

Figure 12.4
Preference for Formula One using rank order scaling

easily understand the instructions for ranking. The major disadvantage is that this technique produces only ordinal data.

Finally, under the assumption of transitivity, rank order data can be converted to equivalent paired comparison data, and vice versa. This point was illustrated by examining the 'Number of times preferred' in Figure 12.3. Hence, it is possible to derive an interval scale from rankings using the Thurstone case V procedure. Other approaches for deriving interval scales from rankings have also been suggested.[15]

Constant sum scaling

Constant sum scaling
A comparative scaling technique in which respondents are required to allocate a constant sum of units such as points, euros, chits, stickers or chips among a set of stimulus objects with respect to some criterion.

In **constant sum scaling**, respondents allocate a constant sum of units, such as points or euros, among a set of stimulus objects with respect to some criterion. As shown in Figure 12.5, respondents may be asked to allocate 100 points to attributes of bottled beers in a way that reflects the importance they attach to each attribute. If an attribute is unimportant, the respondent 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.

The attributes are scaled by counting the points assigned to each one by all the respondents and dividing by the number of respondents. These results are presented for three groups, or segments, of respondents in Figure 12.5. Segment I attaches overwhelming importance to price. Segment II considers a high alcoholic level to be of prime importance. Segment III values bitterness, hop flavours, fragrance and the aftertaste. Such information cannot be obtained from rank order data unless they are transformed into interval data. Note that the constant sum also has an absolute zero; 10 points are twice as many as 5 points, and the difference between 5 and 2 points is the same as the difference between 57 and 54 points. For this reason, constant sum scale data are sometimes treated as metric. Although this may be appropriate in the limited context of the stimuli scaled, these results are not generalisable to other stimuli not included in the

Instructions

Below are eight attributes of bottled beers. Please allocate 100 points among the attributes so that your allocation reflects the relative importance you attach to each attribute. The more points an attribute receives, the more important an attribute is. If an attribute is not at all important, assign it no points. If an attribute is twice as important as some other attribute, it should receive twice as many points.

Note: the figures below represent the mean points allocated to bottled beers by three segments of a target market.

Form

		MEAN POINTS ALLOCATED		
	Attribute	Segment I	Segment II	Segment III
1	Bitterness	8	2	17
2	Hop flavours	2	4	20
3	Fragrance	3	9	19
4	Country where brewed	9	17	4
5	Price	53	5	7
6	High alcohol level	7	60	9
7	Aftertaste	5	0	15
8	Package design	13	3	9
	Sum	100	100	100

Figure 12.5
Importance of bottled beer attributes using a constant sum scale

study. Hence, strictly speaking, the constant sum should be considered an ordinal scale because of its comparative nature and the resulting lack of generalisability. It can be seen that the allocation of points in Figure 12.5 is influenced by the specific attributes included in the evaluation task.

The main advantage of the constant sum scale is that it allows for fine discrimination among stimulus objects without requiring too much time. It has two primary disadvantages, however. Respondents may allocate more or fewer units than those specified. For example, a respondent may allocate 108 or 94 points. The researcher must modify such data in some way or eliminate this respondent from analysis. Another potential problem is rounding error if too few units are used. On the other hand, the use of a large number of units may be too taxing on the respondent and cause confusion and fatigue.

Q-sort and other procedures

Q-sort scaling
A comparative scaling technique that uses a rank order procedure to sort objects based on similarity with respect to some criterion.

Q-sort scaling was developed to discriminate among a relatively large number of objects quickly. This technique uses a rank order procedure in which objects are sorted into piles based on similarity with respect to some criterion. For example, respondents are given 100 attitude statements on individual cards and asked to place them into 11 piles, ranging from 'most highly agreed with' to 'least highly agreed with'. The number of objects to be sorted should not be less than 60 nor more than 140; a reasonable range is 60 to 90 objects.[16] The number of objects to be placed in each pile is pre-specified, often to result in a roughly normal distribution of objects over the whole set.

Another comparative scaling technique is magnitude estimation.[17] In this technique, numbers are assigned to objects such that ratios between the assigned numbers reflect ratios on the specified criterion. For example, respondents may be asked to indicate whether they agree or disagree with each of a series of statements measuring attitude towards different sports. Then they assign a number between 0 and to 100 to each statement to indicate the intensity of their agreement or disagreement. Providing this type of number imposes a cognitive burden on the respondents.

Verbal protocol
A technique used to understand respondents' cognitive responses or thought processes by having them think aloud while completing a task or making a decision.

Another particularly useful procedure (that could be viewed as a very structured combination of observation and depth interviewing) for measuring cognitive responses or thought processes consists of **verbal protocols**. Respondents are asked to 'think out loud' and verbalise anything going through their heads while making a decision or performing a task.[18] The researcher says, 'If you think anything, say it aloud, no matter how trivial the thought may be.' Even with such an explicit instruction, the respondents may be silent. At these times, the researcher will say, 'Remember to say aloud everything you are thinking.' Everything that the respondents say is tape recorded. This record of the respondents' verbalised thought processeses is referred to as a protocol.[19]

Protocols have been used to measure consumers' cognitive responses in actual shopping trips as well as in simulated shopping environments. An interviewer accompanies the respondent and holds a microphone into which the respondent talks. Protocols, thus collected, have been used to determine the attributes and cues used in making purchase decisions, product usage behaviour and the impact of the shopping environment on consumer decisions. Protocol analysis has also been employed to measure consumer response to advertising. Immediately after seeing an ad, the respondent is asked to list all the thoughts that came to mind while watching the ad. The respondent is given a limited amount of time to list the thoughts so as to minimise the probability of collecting thoughts generated after, rather than during, the message. After the protocol has been collected, the individual's thoughts or cognitive responses can be coded into three categories as illustrated in Table 12.3.[20]

Table 12.3 Coded verbal protocols

Category	Definition	Example
Support argument	Support the claim made by the message	'Diet Coke tastes great'
Counter-argument	Refute the claim made by the message	'Diet Coke has an aftertaste'
Source derogation	Negative opinion about the source of the message	'Coca-Cola is not an honest company'

Protocols are, typically, incomplete. The respondent has many thoughts that he or she cannot or will not verbalise. The researcher must take the incomplete record and infer from it a measure of the underlying cognitive response.

Non-comparative scaling techniques

Respondents using a non-comparative scale employ whatever rating standard seems appropriate to them. They do not compare the object being rated either with another object or with some specified standard, such as 'your ideal brand'. They evaluate only one object at a time; thus, non-comparative scales are often referred to as monadic scales. Non-comparative techniques consist of continuous and itemised rating scales, which are described in Table 12.4 and discussed in the following sections.

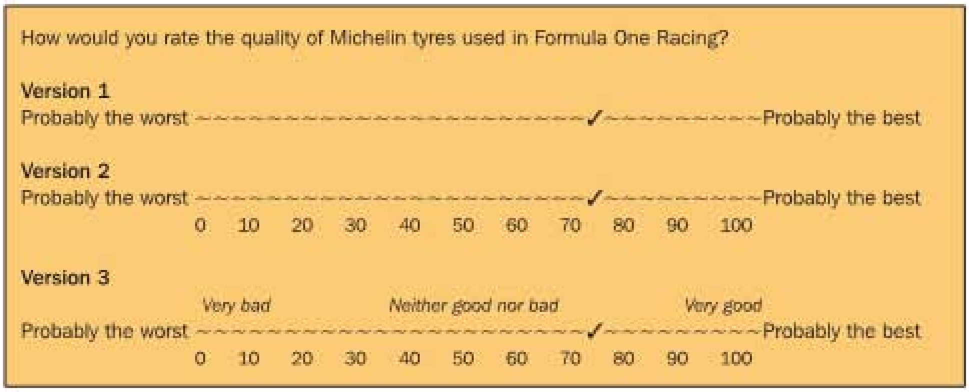
Table 12.4 Basic non-comparative scales

Scale	Basic characteristics	Examples	Advantages	Disadvantages
Continuous rating scale	Place a mark on a continuous line	Reaction to TV commercials	Easy to construct	Scoring can be cumbersome unless computerised
Itemised rating scales				
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 phone	Confusing and difficult to apply

Continuous rating scale

Continuous rating scale
A measurement scale that has 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 the other. The form may vary considerably. Also called graphic rating scale.

In a **continuous rating scale**, also referred to as a graphic rating scale, 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 the other. Thus, the respondents are not restricted to selecting from marks previously set by the researcher. The form of the continuous scale may vary considerably. For example, the line may be vertical or horizontal: scale points, in the form of numbers or brief descriptions, may be provided; and if provided, the scale points may be few or many. Three versions of a continuous rating scale are illustrated in Figure 12.6.


How would you rate the quality of Michelin tyres used in Formula One Racing?

Version 1
Probably the worst ~~~~~~~~~~~~~~~~~~~~~~~~~~✓~~~~~~~~~~Probably the best

Version 2
Probably the worst ~~~~~~~~~~~~~~~~~~~~~~~~~~✓~~~~~~~~~~Probably the best
0 10 20 30 40 50 60 70 80 90 100

Version 3
Very bad *Neither good nor bad* *Very good*
Probably the worst ~~~~~~~~~~~~~~~~~~~~~~~~~~✓~~~~~~~~~~Probably the best
0 10 20 30 40 50 60 70 80 90 100

Figure 12.6
Continuous rating scale

Once the respondent has provided the ratings, the researcher divides the line into as many categories as desired and assigns scores based on the categories into which the ratings fall. In Figure 12.6, the respondent exhibits a favourable attitude towards Michelin tyres. These scores are typically treated as interval data. The advantage of continuous scales is that they are easy to construct; however, scoring is cumbersome and unreliable.[21] Moreover, continuous scales provide little new information. Hence, their use in marketing research has been limited. Recently, however, with the increased popularity of computer-assisted personal interviewing and other technologies, their use has become more frequent.

Itemised rating scales

Itemised rating scale
A measurement scale having numbers or brief descriptions associated with each category. The categories are ordered in terms of scale position.

In an **itemised rating scale**, respondents are provided with a scale that has a number or brief description associated with each category. The categories are ordered in terms of scale position, and the respondents are required to select the specified category that best describes the object being rated. Itemised rating scales are widely used in marketing research and form the basic components of more complex scales, such as multi-item rating scales. We first describe the commonly used itemised rating scales – the Likert, semantic differential and Stapel scales – and then examine the major issues surrounding the use of itemised rating scales.

Likert scale
A measurement scale with five response categories ranging from 'strongly disagree' to 'strongly agree' that requires respondents to indicate a degree of agreement or disagreement with each of a series of statements related to the stimulus objects.

Likert scale

Named after its developer, Rensis Likert, the **Likert scale** is a widely used rating scale that requires the respondents to indicate a degree of agreement or disagreement with each of a series of statements about the stimulus objects.[22] Typically, each scale item has five response categories, ranging from 'strongly disagree' to 'strongly agree'. We illustrate with a Likert scale for evaluating attitudes towards Renault cars.

To conduct the analysis, each statement is assigned a numerical score, ranging either from –2 to +2 or from 1 to 5. The analysis can be conducted on an item-by-item basis (profile analysis), or a total (summated) score can be calculated for each respondent by summing across items. Suppose that the Likert scale in Figure 12.7 was used to measure attitudes towards Renault as well as Ford. Profile analysis would involve comparing the two car manufacturers in terms of the average respondent ratings for each item. The summated approach is most frequently used, and, as a result, the Likert scale is also referred to as a summated scale.[23] When using this approach to determine the total score for each respondent on each car manufacturer, it is important to use a consistent scoring procedure so that a high (or low) score consistently reflects a favourable response. This requires that the categories assigned to the negative statements by the respondents be scored by reversing the scale. Note that for a negative statement, an agreement reflects an unfavourable response, whereas for a positive statement, agreement represents a favourable response. Accordingly, a 'strongly agree' response to a favourable statement and a 'strongly disagree' response to an unfavourable statement would both receive scores of 5.[24] In the example in Figure 12.7, if a higher score is to denote a more favourable attitude, the scoring of items 2, 4, 5 and 7 will be reversed. The respondent to this set of statements has an attitude score of 26. Each respondent's total score for each car manufacturer is calculated. A respondent will have the most favourable attitude towards a car manufacturer with the highest score. The procedure for developing summated Likert scales is described later in the section on the development and evaluation of scales.

The Likert scale has several advantages. It is easy to construct and administer, and respondents readily understand how to use the scale, making it suitable for Internet

Instructions

Listed below are different beliefs about Renault cars. Please indicate how strongly you agree or disagree with each by putting a tick next to your choice on the following scale:

1 = Strongly disagree, 2 = Disagree, 3 = Neither agree nor disagree, 4 = Agree, 5 = Strongly agree

	Strongly disagree	*Disagree*	*Neither agree nor disagree*	*Agree*	*Strongly agree*
1 Renault produces high-quality cars	1	2✓	3	4	5
2 Renault has poor after-sales service	1	2✓	3	4	5
3 I like to visit Renault dealerships	1	2	3✓	4	5
4 Renault does not offer a good range of optional extras for its cars	1	2	3	4✓	5
5 The credit terms at Renault dealerships are terrible	1	2	3	4✓	5
6 Renault is the embodiment of European excellence in car manufacturing	1✓	2	3	4	5
7 I do not like Renault advertising	1	2	3	4✓	5
8 Renault has an excellent selection of car types	1	2	3	4✓	5
9 The price of Renault cars is fair	1	2✓	3	4	5

Figure 12.7
The Likert scale

surveys, mail, telephone or personal interviews. The major disadvantage of the Likert scale is that it takes longer to complete than other itemised rating scales because respondents have to read and fully reflect upon each statement.

Semantic differential scale

Semantic differential
A seven-point rating scale with end points associated with bipolar labels.

The **semantic differential** is a seven-point rating scale with end points associated with bipolar labels that have semantic meaning. In a typical application, respondents rate objects on a number of itemised, seven-point rating scales bounded at each end by one of two bipolar adjectives, such as 'boring' and 'exciting'.[25] We illustrate this scale in Figure 12.8 by presenting a respondent's evaluation of Formula One racing on five attributes.

The respondents mark the blank that best indicates how he or she would describe the object being rated.[26] Thus, in our example, Formula One is evaluated as exciting, innovative, safe, dynamic, though uninspiring. The negative adjective or phrase sometimes appears at the left side of the scale and sometimes at the right. This controls the tendency of some respondents, particularly those with very positive or very negative attitudes, to mark the right- or left-hand sides without reading the labels.

Individual items on a semantic differential scale may be scored either on a –3 to +3 or on a 1 to 7 scale. The resulting data are commonly analysed through profile analysis. In profile analysis, means or median values on each rating scale are calculated and compared by plotting or statistical analysis. This helps determine the overall differences and similarities among the objects. To assess differences across segments of respondents, the researcher can compare mean responses of different segments. Although the mean is most often used as a summary statistic, there is some controversy as to whether the data obtained should be treated as an interval scale.[27] On the other hand, in cases when the researcher requires an overall comparison of objects, such as to determine car manufacturer preference, the individual item scores are summed to arrive at a total score.

Its versatility makes the semantic differential a popular rating scale in marketing research. It has been widely used in comparing brand, product and company images. It has also been used to develop advertising and promotion strategies and in new product development studies.[28]

Stapel scale

Stapel scale
A scale for measuring attitudes that consists of a single adjective in the middle of an even-numbered range of values.

The **Stapel scale**, named after its developer, Jan Stapel, is a unipolar rating scale with 10 categories numbered from –5 to +5, without a neutral point (zero).[29] This scale is usually presented vertically. Respondents are asked to indicate, by selecting an appropriate numerical response category, how accurately or inaccurately each term describes the object. The higher the number, the more accurately the term describes the object, as shown in Figure 12.9. In this example, Formula One is perceived as being prestigious but not elitist.

Instructions
What does Formula One racing mean to you? The following descriptive scales, bounded at each end by bipolar adjectives, summarise characteristics of the sport. Please mark ✗ the blank that best indicates what Formula One means to you.

Form
Formula One is:

Boring	:__::__::__::__::__::✗::__:	Exciting
Conservative	:__::__::__::__::__::✗::__:	Innovative
Dangerous	:__::__::__::__::__::__::✗:	Safe
Staid	:__::__::__::__::__::✗::__:	Dynamic
Uninspiring	:__::✗::__::__::__::__::__:	Inspirational

Figure 12.8
Semantic differential scale

Instructions

Please evaluate how accurately each word or phrase describes Formula One racing. Select a positive number for the phrases you think describe the sport accurately. The more accurately you think the phrase describes the sport, the larger the plus number you should choose. You should select a minus number for the phrases you think do not describe the sport accurately. The less accurately you think the phrase describes the sport, the larger the negative number you should choose. You can select any number from +5 for phrases you think are very accurate, to –5 for phrases you think are very inaccurate.

Form

Formula One

+5	+5
+4✗	+4
+3	+3
+2	+2
+1	+1
Prestigious	*Elitist*
–1	–1
–2	–2✗
–3	–3
–4	–4
–5	–5

Figure 12.9
The Stapel scale

The data obtained by using a Stapel scale can be analysed in the same way as semantic differential data. The Stapel scale produces results similar to the semantic differential.[30] The Stapel scale's advantages are that it does not require a pretest of the adjectives or phrases to ensure true bipolarity and that it can be administered over the telephone. Some researchers, however, believe the Stapel scale is confusing and difficult to apply. Of the three itemised rating scales considered, the Stapel scale is used least.[31] Nonetheless, this scale merits more attention than it has received.

Itemised rating scale decisions

As is evident from the discussion so far, non-comparative itemised rating scales can take many different forms. The researcher must make six major decisions when constructing any of these scales:

1 The number of scale categories to use
2 Balanced versus unbalanced scale
3 Odd or even number of categories
4 Forced versus non-forced choice
5 The nature and degree of the verbal description
6 The physical form of the scale.

Number of scale categories

Two conflicting considerations are involved in deciding the number of scale categories or response options. The greater the number of scale categories, the finer the discrimination among stimulus objects that is possible. On the other hand, most respondents cannot handle more than a few categories. Traditional guidelines suggest that the appropriate number of categories should be between five and nine.[32] Yet there is no single optimal number of categories. Several factors should be taken into account in deciding on the number of categories.

If the respondents are interested in the scaling task and are knowledgeable about the objects, many categories may be employed. On the other hand, if the respondents are not very knowledgeable or involved with the task, fewer categories should be used. Likewise, the nature of the objects is also relevant. Some objects do not lend themselves to fine discrimination, so a small number of categories are sufficient. Another important factor is the mode of data collection. If telephone interviews are involved, many categories may confuse the respondents. Likewise, space limitations may restrict the number of categories in mail questionnaires.

How the data are to be analysed and used should also influence the number of categories. In situations where several scale items are added together to produce a single score for each respondent, five categories are sufficient. The same is true if the researcher wishes to make broad generalisations or group comparisons. If, however, individual responses are of interest or if the data will be analysed by sophisticated statistical techniques, seven or more categories may be required. The size of the correlation coefficient, a common measure of relationship between variables (Chapter 20), is influenced by the number of scale categories. The correlation coefficient decreases with a reduction in the number of categories. This, in turn, has an impact on all statistical analysis based on the correlation coefficient.[33]

Balanced versus unbalanced scale

Balanced scale
A scale with an equal number of favourable and unfavourable categories.

In a **balanced scale**, the number of favourable and unfavourable categories is equal; in an unbalanced scale, the categories are unequal.[34] Examples of balanced and unbalanced scales are given in Figure 12.10.

In general, in order to obtain objective data, the scale should be balanced. If the distribution of responses is likely to be skewed, however, either positively or negatively, an unbalanced scale with more categories in the direction of skewness may be appropriate. If an unbalanced scale is used, the nature and degree of imbalance in the scale should be taken into account in data analysis.

Odd or even number of categories

With an odd number of categories, the middle scale position is generally designated as neutral or impartial. The presence, position and labelling of a neutral category can have a significant influence on the response. The Likert scale is a balanced rating scale with an odd number of categories and a neutral point.[35]

The decision to use an odd or even number of categories depends on whether some of the respondents may be neutral on the response being measured. If a neutral or indifferent response is possible from at least some of the respondents, an odd number of categories should be used. If, on the other hand, the researcher wants to force a response or believes that no neutral or indifferent response exists, a rating scale with an even number of categories should be used. A related issue is whether the choice should be forced or non-forced.

Balanced scale

Clinique moisturiser for men is:

Extremely good	
Very good	✓
Good	
Bad	
Very bad	
Extremely bad	

Unbalanced scale

Clinique moisturiser for men is:

Extremely good	
Very good	✓
Good	
Somewhat good	
Bad	
Very bad	

Figure 12.10
Balanced and unbalanced scales

Forced versus non-forced choice

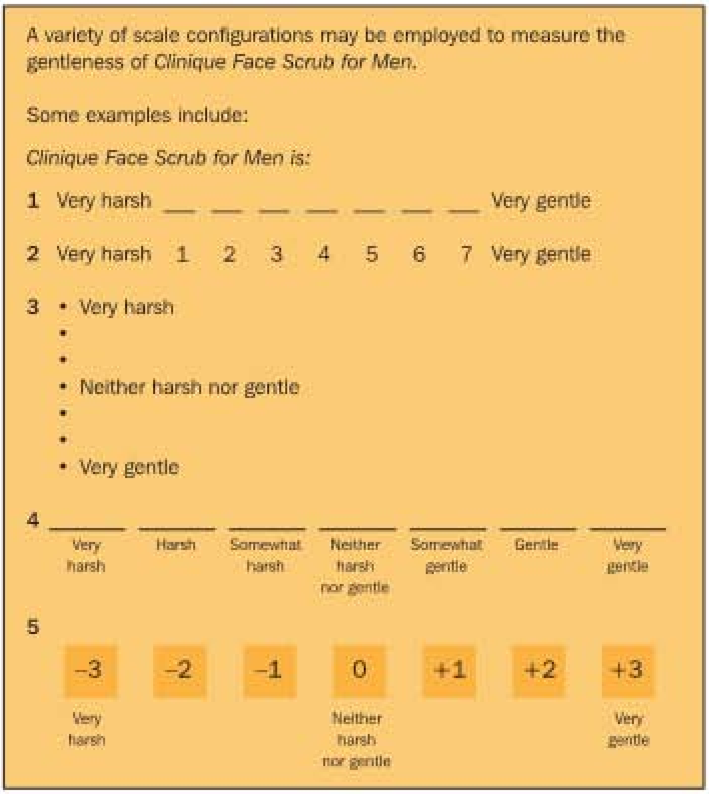
Forced rating scale
A rating scale that forces respondents to express an opinion because a 'no opinion' or 'no knowledge' option is not provided.

On **forced rating scales** the respondents are forced to express an opinion because a 'no opinion' option is not provided. In such a case, respondents without an opinion may mark the middle scale position. If a sufficient proportion of the respondents do not have opinions on the topic, marking the middle position will distort measures of central tendency and variance. In situations where the respondents are expected to have no opinion, as opposed to simply being reluctant to disclose it, the accuracy of data may be improved by a non-forced scale that includes a 'no opinion' category.[36]

Nature and degree of verbal description

The nature and degree of verbal description associated with scale categories varies considerably and can have an effect on the responses. Scale categories may have verbal, numerical or even pictorial descriptions. Furthermore, the researcher must decide whether to label every scale category, label only some scale categories, or label only extreme scale categories. Surprisingly, providing a verbal description for each category may not improve the accuracy or reliability of the data. Yet, an argument can be made for labelling all or many scale categories to reduce scale ambiguity. The category descriptions should be located as close to the response categories as possible.

The strength of the adjectives used to anchor the scale may influence the distribution of the responses. With strong anchors (1 = completely disagree, 7 = completely agree), respondents are less likely to use the extreme scale categories. This results in less variable and more peaked response distributions. Weak anchors (1 = generally disagree, 7 = generally agree), in contrast, produce uniform or flat distributions. Procedures have been developed to assign values to category descriptors to result in balanced or equal interval scales.[37]



A variety of scale configurations may be employed to measure the gentleness of *Clinique Face Scrub for Men*.

Some examples include:

Clinique Face Scrub for Men is:

1 Very harsh __ __ __ __ __ __ __ Very gentle

2 Very harsh 1 2 3 4 5 6 7 Very gentle

3
- Very harsh
-
-
- Neither harsh nor gentle
-
-
- Very gentle

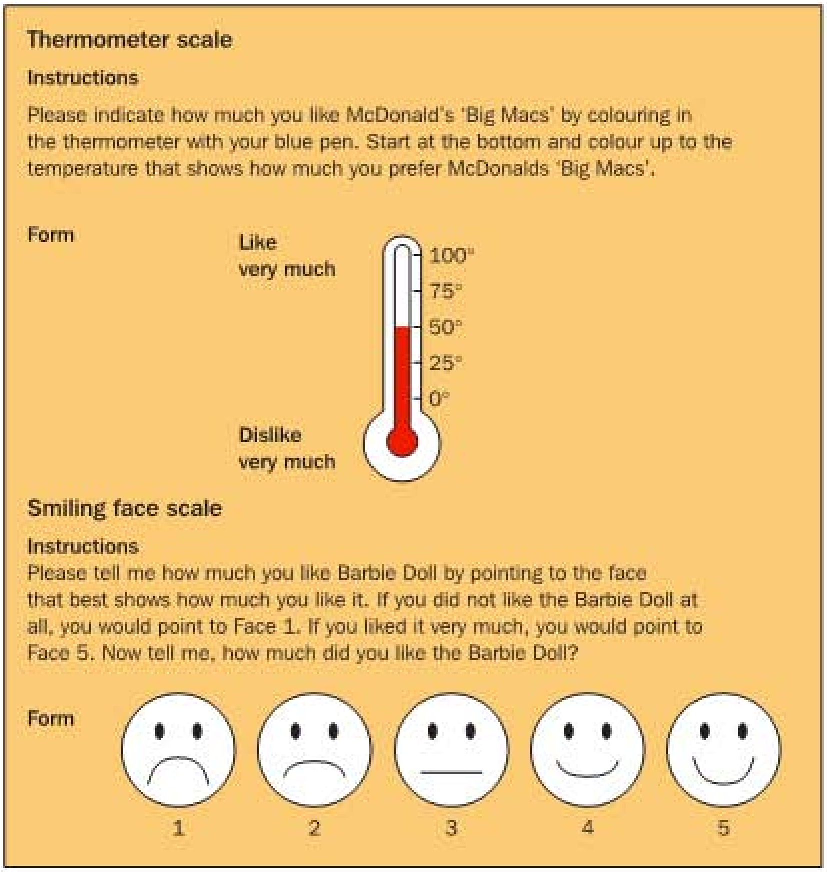
4

____	____	____	____	____	____	____
Very harsh	Harsh	Somewhat harsh	Neither harsh nor gentle	Somewhat gentle	Gentle	Very gentle

5

−3	−2	−1	0	+1	+2	+3
Very harsh			Neither harsh nor gentle			Very gentle

Figure 12.11
Rating scale configurations



Thermometer scale

Instructions

Please indicate how much you like McDonald's 'Big Macs' by colouring in the thermometer with your blue pen. Start at the bottom and colour up to the temperature that shows how much you prefer McDonalds 'Big Macs'.

Smiling face scale

Instructions

Please tell me how much you like Barbie Doll by pointing to the face that best shows how much you like it. If you did not like the Barbie Doll at all, you would point to Face 1. If you liked it very much, you would point to Face 5. Now tell me, how much did you like the Barbie Doll?

Figure 12.12 Some unique rating scale configurations

Physical form of the scale

A number of options are available with respect to scale form or configuration. Scales can be presented vertically or horizontally. Categories can be expressed by boxes, discrete lines or units on a continuum and may or may not have numbers assigned to them. If numerical values are used, they may be positive, negative or both. Several possible configurations are presented in Figure 12.11.

Two unique rating scale configurations used in marketing research are the thermometer scale and the smiling face scale. For the thermometer scale, the higher the temperature, the more favourable the evaluation. Likewise, happier faces indicate evaluations that are more favourable. These scales are especially useful for children.[38] Examples of these scales are shown in Figure 12.12. Table 12.5 summarises the six decisions in designing rating scales.

Table 12.5 Summary of itemised rating scale decisions

1 Number of categories	Although there is no single, optimal number, traditional guidelines suggest that there should be between five and nine categories
2 Balanced versus unbalanced	In general, the scale should be balanced to obtain objective data
3 Odd or even number of categories	If a neutral or indifferent scale response is possible from at least some of the respondents, an odd number of categories should be used
4 Forced versus unforced	In situations where the respondents are expected to have no opinion, the accuracy of the data may be improved by a non-forced scale
5 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
6 Physical form	A number of options should be tried and the best one selected

The development and evaluation of scales

The development of multi-item rating scales requires considerable technical expertise.[39] Figure 12.13 presents a sequence of operations needed to construct multi-item scales.

The characteristic to be measured is frequently called a construct. Scale development begins with an underlying theory of the construct being measured. Theory is necessary not only for constructing the scale but also for interpreting the resulting scores. The next step is to generate an initial pool of scale items. Typically, this is based on theory, analysis of secondary data and qualitative research. From this pool, a reduced set of potential scale items is generated by the judgement of the researcher and other knowledgeable individuals. Some qualitative criterion is adopted to aid their judgement. The reduced set of items may still be too large to constitute a scale. Thus, further reduction is achieved in a quantitative manner.

Data are collected on the reduced set of potential scale items from a large pretest sample of respondents. The data are analysed using techniques such as correlations, factor analysis, cluster analysis, discriminant analysis and statistical tests discussed later in this book. As a result of these statistical analyses, several more items are eliminated, resulting in a purified scale. The purified scale is evaluated for reliability and validity by collecting more data from a different sample (these concepts will be explained on pages 357–359). On the basis of these assessments, a final set of scale items is selected. As can be seen from Figure 12.13, the scale development process is an iterative one with several feedback loops.[40]

A multi-item scale should be evaluated for accuracy and applicability.[41] As shown in Figure 12.14, this involves an assessment of reliability, validity and generalisability of the scale. Approaches to assessing reliability include test–retest reliability, alternative-forms

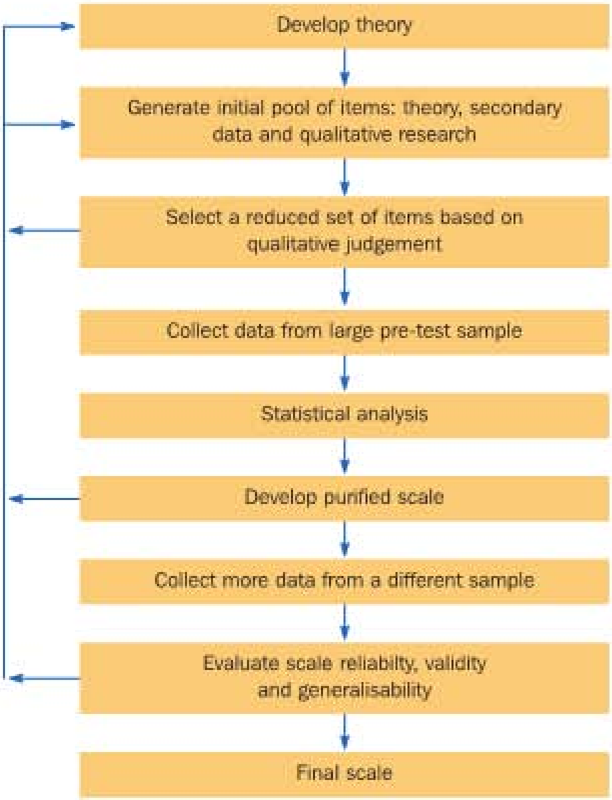

Develop theory
Generate initial pool of items: theory, secondary data and qualitative research
Select a reduced set of items based on qualitative judgement
Collect data from large pre-test sample
Statistical analysis
Develop purified scale
Collect more data from a different sample
Evaluate scale reliabilty, validity and generalisability
Final scale

Figure 12.13
Development of a multi-item scale

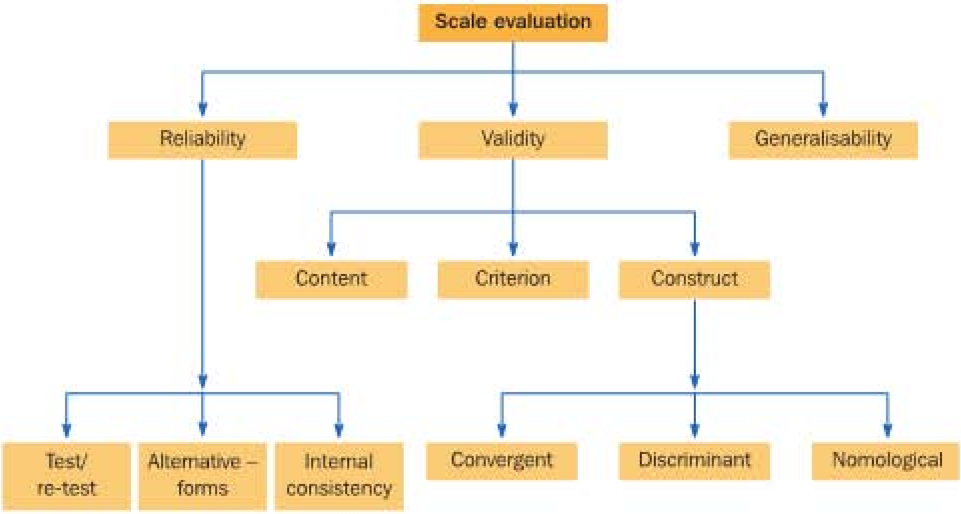


Figure 12.14 Scale evaluation

reliability and internal consistency reliability. Validity can be assessed by examining content validity, criterion validity and construct validity.

Before we can examine reliability and validity we need an understanding of measurement accuracy; it is fundamental to scale evaluation.

Measurement accuracy

A measurement is a number that reflects some characteristic of an object. A measurement is not the true value of the characteristic of interest but rather an observation of it. A variety of factors can cause **measurement error**, which results in the measurement or observed score being different from the true score of the characteristic being measured (see Table 12.6).

Measurement error
The variation in the information sought by the researcher and the information generated by the measurement process employed.

The **true score model** provides a framework for understanding the accuracy of measurement.[42] According to this model,

True score model
A mathematical model that provides a framework for understanding the accuracy of measurement.

$$X_O = X_T + X_S + X_R$$

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

Table 12.6 Potential sources of error in measurement

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

Systematic error
An error that affects the measurement in a constant way and represents stable factors that affect the observed score in the same way each time the measurement is made.

Random error
An error that arises from random changes or differences in respondents or measurement situations.

Reliability
The extent to which a scale produces consistent results if repeated measurements are made on the characteristic.

Note that the total measurement error includes the systematic error, X_S, and the random error, X_R. **Systematic error** affects the measurement in a constant way. It represents stable factors that affect the observed score in the same way each time the measurement is made, such as mechanical factors (see Table 12.6). **Random error**, on the other hand, is not constant. It represents transient factors that affect the observed score in different ways each time the measurement is made, such as short-term transient personal factors or situational factors (see Table 12.6). The distinction between systematic and random error is crucial to our understanding of reliability and validity.

Reliability

Reliability refers to the extent to which a scale produces consistent results if repeated measurements are made.[43] Systematic sources of error do not have an adverse impact on reliability, because they affect the measurement in a constant way and do not lead to inconsistency. In contrast, random error produces inconsistency, leading to lower reliability. 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.

Reliability is assessed by determining the proportion of systematic variation in a scale. This is done by determining the association between scores obtained from different administrations of the scale. If the association is high, the scale yields consistent results and is therefore reliable. Approaches for assessing reliability include the test–retest, alternative-forms and internal consistency methods.

Test–retest reliability
An approach for assessing reliability, in which respondents are administered identical sets of scale items at two different times, under as nearly equivalent conditions as possible.

In **test–retest reliability**, respondents are administered identical sets of scale items at two different times, under as nearly equivalent conditions as possible. The time interval between tests or administrations is typically two to four weeks. The degree of similarity between the two measurements is determined by computing a correlation coefficient (see Chapter 20). The higher the correlation coefficient, the greater the reliability.

Several problems are associated with the test–retest approach to determining reliability. First, it is sensitive to the time interval between testing. Other things being equal, the longer the time interval, the lower the reliability. Second, the initial measurement may alter the characteristic being measured. For example, measuring respondents' attitude towards low-alcohol beer may cause them to become more health conscious and to develop a more positive attitude towards low-alcohol beer. Third, it may be impossible to make repeated measurements (e.g. the research topic may be the respondent's initial reaction to a new product). Fourth, the first measurement may have a carryover effect to the second or subsequent measurements. Respondents may attempt to remember answers they gave the first time. Fifth, the characteristic being measured may change between measurements. For example, favourable information about an object between measurements may make a respondent's attitude more positive. Finally, the test–retest reliability coefficient can be inflated by the correlation of each item with itself. These correlations tend to be higher than correlations between different scale items across administrations. Hence, it is possible to have high test–retest correlations because of the high correlations between the same scale items measured at different times even though the correlations between different scale items are quite low. Because of these problems, a test–retest approach is best applied in conjunction with other approaches, such as alternative-forms reliability.[44]

Alternative-forms reliability
An approach for assessing reliability that requires two equivalent forms of the scale to be constructed and then the same respondents to be measured at two different times.

In **alternative-forms reliability**, two equivalent forms of the scale are constructed. The same respondents are measured at two different times, usually two to four weeks apart (e.g. by initially using Likert scaled items and then using Stapel scaled items). The scores from the administrations of the alternative scale forms are correlated to assess reliability.[45]

The two forms should be equivalent with respect to content, i.e. each scale item should attempt to measure the same items. The main problems with this approach are that it is difficult, time consuming and expensive to construct an equivalent form of the scale. In a strict sense, it is required that the alternative sets of scale items should have the same means, variances and intercorrelations. Even if these conditions are satisfied, the two forms may not be equivalent in content. Thus, a low correlation may reflect either an unreliable scale or non-equivalent forms.

Internal consistency reliability
An approach for assessing the internal consistency of a set of items, where several items are summated in order to form a total score for the scale.

Split-half reliability
A form of internal consistency reliability in which the items constituting the scale are divided into two halves and the resulting half scores are correlated.

Coefficient alpha
A measure of internal consistency reliability that is the average of all possible split-half coefficients resulting from different splittings of the scale items.

Internal consistency reliability is used to assess the reliability of a summated scale where several items are summed to form a total score. In a scale of this type, each item measures some aspect of the construct measured by the entire scale, and the items should be consistent in what they indicate about the construct. This measure of reliability focuses on the internal consistency of the set of items forming the scale.

The simplest measure of internal consistency is **split-half reliability**. The items on the scale are divided into two halves and the resulting half scores are correlated. High correlations between the halves indicate high internal consistency. The scale items can be split into halves based on odd- and even-numbered items or randomly. The problem is that the results will depend on how the scale items are split. A popular approach to overcoming this problem is to use the coefficient alpha.

The **coefficient alpha**, or Cronbach's alpha, is the average of all possible split-half coefficients resulting from different ways of splitting the scale items.[46] This coefficient varies from 0 to 1, and a value of 0.6 or less generally indicates unsatisfactory internal consistency reliability. An important property of coefficient alpha is that its value tends to increase with an increase in the number of scale items.[47] Therefore, coefficient alpha may be artificially, and inappropriately, inflated by including several redundant scale items. Another coefficient that can be employed in conjunction with coefficient alpha is coefficient beta. Coefficient beta assists in determining whether the averaging process used in calculating coefficient alpha is masking any inconsistent items.

Some multi-item scales include several sets of items designed to measure different aspects of a multidimensional construct. For example, car manufacturer image is a multidimensional construct that includes country of origin, range of cars, quality of cars, car performance, service of car dealers, credit terms, dealer location and physical layout of dealerships. Hence, a scale designed to measure car manufacturer image could contain items measuring each of these dimensions. Because these dimensions are somewhat independent, a measure of internal consistency computed across dimensions would be inappropriate. If several items are used to measure each dimension, however, internal consistency reliability can be computed for each dimension.

Validity

Validity
The extent to which a measurement represents characteristics that exist in the phenomenon under investigation.

Content validity
A type of validity, sometimes called face validity, that consists of a subjective but systematic evaluation of the representativeness of the content of a scale for the measuring task at hand.

Criterion validity
A type of validity that examines whether the measurement scale performs as expected in relation to other selected variables as meaningful criteria.

The **validity** of a scale may be considered 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 ($X_O = X_T$, $X_R = 0$, $X_S = 0$). Researchers may assess content validity, criterion validity or construct validity.[48]

Content validity, sometimes called face validity, is a subjective but systematic evaluation of how well the content of a scale represents the measurement task at hand. The researcher or someone else examines whether the scale items adequately cover the entire domain of the construct being measured. Thus, a scale designed to measure car manufacturer image would be considered inadequate if it omitted any of the major dimensions (country of origin, range of cars, quality of cars, car performance, etc.). Given its subjective nature, content validity alone is not a sufficient measure of the validity of a scale, yet it aids in a common-sense interpretation of the scale scores. A more formal evaluation can be obtained by examining criterion validity.

Criterion validity reflects whether a scale performs as expected in relation to other selected variables (criterion variables) as meaningful criteria. If, for example, a scale is designed to measure loyalty in customers, criterion validity might be determined by comparing the results generated by this scale with results generated by observing the extent of repeat purchasing. Based on the time period involved, criterion validity can take two forms, concurrent validity and predictive validity.

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

Concurrent validity is assessed when the data on the scale being evaluated (e.g. loyalty scale) and the criterion variables (e.g. repeat purchasing) are collected at the same time. The scale being developed and the alternative means of encapsulating the criterion variables would be administered simultaneously and the results compared.

Predictive validity
A type of validity that is concerned with how well a scale can forecast a future criterion.

Predictive validity is concerned with how well a scale can forecast a future criterion. To assess predictive validity, the researcher collects data on the scale at one point in time and data on the criterion variables at a future time. For example, attitudes towards how loyal customers feel to a particular brand could be used to predict future repeat purchases of that brand. The predicted and actual purchases are compared to assess the predictive validity of the attitudinal scale.

Construct validity
A type of validity that addresses the question of what construct or characteristic the scale is measuring. An attempt is made to answer theoretical questions of why a scale works and what deductions can be made concerning the theory underlying the scale.

Construct validity addresses the question of what construct or characteristic the scale is, in fact, measuring. When assessing construct validity, the researcher attempts to answer theoretical questions about why the scale works and what deductions can be made concerning the underlying theory. Thus, construct validity requires a sound theory of the nature of the construct being measured and how it relates to other constructs. Construct validity is the most sophisticated and difficult type of validity to establish. As Figure 12.14 shows, construct validity includes convergent, discriminant and nomological validity.

Convergent validity
A measure of construct validity that measures the extent to which the scale correlates positively with other measures of the same construct.

Discriminant validity
A type of construct validity that assesses the extent to which a measure does not correlate with other constructs from which it is supposed to differ.

Nomological validity
A type of validity that assesses the relationship between theoretical constructs. It seeks to confirm significant correlations between the constructs as predicted by a theory.

Convergent validity is the extent to which the scale correlates positively with other measurements of the same construct. It is not necessary that all these measurements be obtained by using conventional scaling techniques. **Discriminant validity** is the extent to which a measure does not correlate with other constructs from which it is supposed to differ. It involves demonstrating a lack of correlation among differing constructs. **Nomological validity** is the extent to which the scale correlates in theoretically predicted ways with measures of different but related constructs. A theoretical model is formulated that leads to further deductions, tests and inferences.

An instance of construct validity can be evaluated in the following example. A researcher seeks to provide evidence of construct validity in a multi-item scale, designed to measure the concept of 'self-image'. These findings would be sought:[49]

- High correlations with other scales designed to measure self-concepts and with reported classifications by friends (convergent validity).
- Low correlations with unrelated constructs of brand loyalty and variety seeking (discriminant validity).
- Brands that are congruent with the individual's self-concept are more preferred, as postulated by the theory (nomological validity).
- A high level of reliability.

Note that a high level of reliability was included as evidence of construct validity in this example. This illustrates the relationship between reliability and validity.

Relationship between reliability and validity

The relationship between reliability and validity can be understood in terms of the true score model. If a measure is perfectly valid, it is also perfectly reliable. In this case, $X_O = X_T$, $X_R = 0$ and $X_S = 0$. Thus, perfect validity implies perfect reliability. 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., $X_S \neq 0$. Thus, unreliability implies invalidity. If a measure is perfectly reliable, it may or may not be perfectly valid, because systematic error may still be present ($X_O = X_T + X_S$). In other words, a reliable scale can be constructed to measure 'customer loyalty' but it may not necessarily be a valid measurement of 'customer loyalty'. Conversely, a valid measurement of 'customer loyalty' has to be reliable. Reliability is a necessary, but not sufficient, condition for validity.

Generalisability

Generalisability refers to the extent to which one can generalise from the observations at hand to a universe of generalisations. The set of all conditions of measurement over which the investigator wishes to generalise is the universe of generalisation. These conditions may include items, interviewers and situations of observation. A researcher may wish to generalise a scale developed for use in personal interviews to other modes of data collection, such as mail and telephone interviews. Likewise, one may wish to generalise from a sample of items to the universe of items, from a sample of times of measurement to the universe of times of measurement, from a sample of observers to a universe of observers, and so on.[50]

Generalisability
The degree to which a study based on a sample applies to the population as a whole.

In generalisability studies, measurement procedures are designed to investigate each universe of interest by sampling conditions of measurement from each of them. For each universe of interest, an aspect of measurement called a facet is included in the study. Traditional reliability methods can be viewed as single-facet generalisability studies. A test–retest correlation is concerned with whether scores obtained from a measurement scale are generalisable to the universe scores across all times of possible measurement. Even if the test–retest correlation is high, nothing can be said about the generalisability of the scale to other universes. To generalise to other universes, generalisability theory procedures must be employed.

Choosing a scaling technique

In addition to theoretical considerations and evaluation of reliability and validity, certain practical factors should be considered in selecting scaling techniques for a particular marketing research problem.[51] Selecting an appropriate rating scale is a necessary first step in developing a good measurement instrument; establishing statistical reliability and validity through a multi-step testing and retesting process should be accorded the highest priority in selecting a scale. A good rating scale should have the following characteristics:[52]

- Minimal response bias
- Respondent interpretation and understanding
- Discriminating power
- Ease of administration
- Ease of use by respondents
- Credibility and usefulness of results.

As a general rule, using the scaling technique that will yield the highest level of information feasible in a given situation will permit using the greatest variety of statistical analyses. Also, regardless of the type of scale used, whenever feasible, several scale items should measure the characteristic of interest. This provides more accurate measurement than a single-item scale. In many situations, it is desirable to use more than one scaling technique or to obtain additional measures using mathematically derived scales.

Mathematically derived scales

All the scaling techniques discussed in this chapter require the respondents to evaluate directly the constructs that the researcher believes to comprise the object of study, e.g. the cognitive state of customer satisfaction. In contrast, mathematical scaling techniques allow researchers to infer respondents' evaluations of the constructs of the object of study. These evaluations are inferred from the respondents' overall judgements. Two popular mathematically derived scaling techniques are multidimensional scaling and conjoint analysis, which are discussed in detail in Chapter 24.

International marketing research

In designing the scale or response format, respondents' educational or literacy levels should be taken into account.[53] One approach is to develop scales that are pan-cultural, or free of cultural biases. Of the scaling techniques we have considered, the semantic differential scale may be said to be pan-cultural. It has been tested in a number of countries and has consistently produced similar results. The consistency of results occurred in the following example where Xerox successfully used a Russian translation of an equivalent English semantic differential scale.

Example

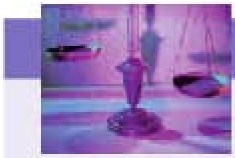
Copying the name Xerox[54]

Xerox was a name well received in the former Soviet Union since the late 1960s. In fact, the act of copying documents was called Xeroxing, a term coined after the name of the company. It was a brand name people equated with quality. With the disintegration of the Soviet Union into the Commonwealth of Independent States, however, Xerox's sales started to fall. The management initially considered this problem to be the intense competition with strong competitors such as Canon, Ricoh, Mitsubishi and Minolta. First attempts to make the product more competitive did not help. Subsequently, marketing research was undertaken to measure the image of Xerox and its competitors in Russia. Semantic differential scales were used, as examples of this type of scale translated well in other countries and were thus considered pan-cultural. The bipolar labels used were carefully tested to ensure that they had the intended semantic meaning in the Russian context.

The results of the study revealed that the real problem was a growing negative perception of Russian customers toward Xerox products. What could have gone wrong? The problem was not with Xerox, but with several independent producers of copying machines that had illegally infringed Xerox's trademark rights. With the disintegration of the Soviet Union, the protection of these trademarks was unclear and trademark infringement kept growing. As a result, customers developed a misconception that Xerox was selling low-quality products.

Although the semantic differential worked well in the Russian context, an alternative approach is to develop scales that use a self-defined cultural norm as a base referent. For example, respondents may be required to indicate their own anchor point and position relative to a culture-specific stimulus set. This approach is useful for measuring attitudes that are defined relative to cultural norms (e.g. attitude towards marital roles). In developing response formats, verbal rating scales appear to be the most suitable. Even less educated respondents can readily understand and respond to verbal scales. Special attention should be devoted to determining equivalent verbal descriptors in different languages and cultures. The end points of the scale are particularly prone to different interpretations. In some cultures, 1 may be interpreted as best, whereas in others it may be interpreted as worst, regardless of how it is scaled. It is important that the scale end points and the verbal descriptors be employed in a manner consistent with the culture.

Finally, in international marketing research, it is critical to establish the equivalence of scales and measures used to obtain data from different countries. This topic is complex and is discussed in some detail in Chapter 26.



Ethics in marketing research

Researchers should not bias scales so as to slant the findings in any particular direction. This is easy to do by biasing the wording of statements (Likert-type scales), the scale descriptors or other aspects of the scales. Consider, for example, the use of scale descriptors. The descriptors used to frame a scale can be manipulated to bias results in any direction. They can be manipulated to generate a positive view of the client's brand or a negative view of a competitor's brand. A researcher who wants to project the client's brand favourably can ask respondents to indicate their opinion of the brand on several attributes using seven-point scales framed by the descriptors 'extremely poor' to 'good'. Using a strongly negative descriptor with only a mildly positive one has an interesting effect. As long as the product is not the worst, respondents will be reluctant to rate the product extremely poorly. In fact, respondents who believe the product to be only mediocre will end up responding favourably. Try this yourself. How would you rate BMW cars on the following attributes?

Reliability	Horrible	1	2	3	4	5	6	7	Good
Performance	Very poor	1	2	3	4	5	6	7	Good
Quality	One of the worst	1	2	3	4	5	6	7	Good
Prestige	Very low	1	2	3	4	5	6	7	Good

Did you find yourself rating BMW cars positively? Using this same technique, a researcher can negatively bias evaluations of competitors' products by providing a mildly negative descriptor (somewhat poor) against a strong positive descriptor (extremely good).

Thus we see how important it is to use balanced scales with comparable positive and negative descriptors. When this guide is not practised, responses are biased and should be interpreted accordingly. This concern also underscores the need to establish adequately the reliability, validity and generalisability of scales before using them in a research project. Scales that are invalid, unreliable or not generalisable to the target market provide the client with flawed results and misleading findings, thus raising serious ethical issues. The researcher has a responsibility to both the client and respondents to ensure the applicability and usefulness of the scale.

Internet and computer applications

All the primary scales of measurement that we have considered can be implemented on the Internet. The same is true for the commonly used comparative scales. Paired comparisons involving verbal, visual or auditory comparisons can be implemented with ease. However, taste, smell and touch comparisons are difficult to implement. It may also be difficult to implement specialised scales such as the Q-sort. The process of implementing comparative scales may be facilitated by searching the Internet for similar scales that have been implemented by other researchers.

Continuous rating scales may be easily implemented on the Internet. The cursor can be moved on the screen in a continuous fashion to select the exact position on the scale that best describes the respondent's evaluation. Moreover, the scale values can be automatically scored by the computer, thus increasing the speed and accuracy of processing the data.

Similarly, it is also easy to implement all of the three itemised rating scales on the Internet. Again, you can use the Internet to search for and locate cases and examples where scales have been used by other researchers. It is also possible that other researchers have reported reliability and validity assessments for multi-item scales. Before generating new scales, a researcher should first examine similar scales used by other researchers and consider using them if they meet the measurement objectives. The following example illustrates how Domino's Pizza uses the Internet to conduct customer surveys and uses the full range of scale types.

Example

Primary scales help Domino's to become a primary competitor[55]

Domino's Pizza builds websites to communicate its image and give information on its products. It also sees its website as a medium to collect information on customers and therefore conduct marketing research. Although no pizza is sold online, the company has a main website (**www.dominos.com**) in addition to websites for its local subsidiaries. For local subsidiaries, the customer is asked to fill in a comment form on the website. This survey helps the local team better to understand its customers' needs and better service them. Different scales are utilised to obtain the following information:

- Name, phone number, email address (nominal scale).
- Preference for pizza restaurants in the local area (ordinal scale).
- Impressions on the service offered by Domino's Pizza as a whole (interval scale).
- Assessments on the products and price (interval scale).
- Customer satisfaction (interval scale).
- Amount spent on pizza and fast foods (ratio scale).

This enables the company to measure customer satisfaction and to use that information for a variety of purposes, including linking it to employee salaries.

Summary

Measurement is the assignment of numbers or other symbols to characteristics of objects according to set rules. Scaling involves the generation of a continuum upon which measured objects are located. The four primary scales of measurement are nominal, ordinal, interval and ratio. Of these, the nominal scale is the most basic in that the numbers are used only for identifying or classifying objects. In the ordinal scale, the numbers indicate the relative position of the objects but not the magnitude of difference between them. The interval scale permits a comparison of the differences between the objects. Because it has an arbitrary zero point, however, it is not meaningful to calculate ratios of scale values on an interval scale. The highest level of measurement is represented by the ratio scale in which the zero point is fixed. The researcher can compute ratios of scale values using this scale. The ratio scale incorporates all the properties of the lower level scales.

Scaling techniques can be classified as comparative or non-comparative. Comparative scaling involves a direct comparison of stimulus objects. Comparative scales include paired comparisons, rank order, constant sum and the Q-sort. The data obtained by these procedures have only ordinal properties. Verbal protocols, where the respondent is instructed to think out loud, can be used for measuring cognitive responses.

In non-comparative scaling, each object is scaled independently of the other objects in the stimulus set. The resulting data are generally assumed to be interval or ratio scaled. Non-comparative rating scales can be either continuous or itemised. The itemised rating scales are further classified as Likert, semantic differential or Stapel scales. When using non-comparative itemised rating scales, the researcher must decide on the number of scale categories, balanced versus unbalanced scales, an odd or even number of categories, forced versus non-forced choices, the nature and degree of verbal description, and the physical form or configuration.

Multi-item scales consist of a number of rating scale items. These scales should be evaluated in terms of reliability and validity. Reliability refers to the extent to which a scale produces consistent results if repeated measurements are made. Approaches to assessing reliability include test–retest, alternative forms and internal consistency. The validity of a measurement may be assessed by evaluating content validity, criterion validity and construct validity.

The choice of particular scaling techniques in a given situation should be based on theoretical and practical considerations. Generally, the scaling technique used should be the one that will yield the highest level of information feasible. Also, multiple measures should be obtained.

In international marketing research, special attention should be devoted to determining equivalent verbal descriptors in different languages and cultures. The misuse of scale descriptors also raises serious ethical concerns. The researcher has a responsibility to both the client and respondents to ensure the applicability and usefulness of scales.

Questions

1 What is measurement?

2 Highlight any marketing phenomena that you feel may be problematic in terms of assigning numbers to characteristics of those phenomena.

3 Describe and illustrate, with examples, the differences between a nominal and an ordinal scale.

4 What are the advantages of a ratio scale over an interval scale? Are these advantages significant?

5 What is a comparative rating scale?

6 What is a paired comparison? What are the advantages and disadvantages of paired comparison scaling?

7 Describe the constant sum scale. How is it different from the other comparative rating scales?

8 Identify the type of scale (nominal, ordinal, interval or ratio) used in each of the following. Give reasons for your choice.

a I like to listen to the radio when I am revising for exams

Disagree Agree
1 2 3 4 5

b How old are you? _____

c Rank the following activities in terms of your preference by assigning a rank from 1 to 5 (1 = most preferred, 2 = second most preferred, etc.):

(i) Reading magazines.
(ii) Watching television.
(iii) Going to the cinema.
(iv) Shopping for clothes.
(v) Eating out.

d What is your university/college registration number? _____

e In an average weekday, how much time do you spend doing class assignments?
   (i) Less than 15 minutes
   (ii) 15 to 30 minutes
   (iii) 31 to 60 minutes
   (iv) 61 to 120 minutes
   (v) More than 120 minutes.

f How much money did you spend last week in the Student Union Bar? _____

9 Describe the semantic differential scale and the Likert scale. For what purposes are these scales used?

10 What are the major decisions involved in constructing an itemised rating scale? How many scale categories should be used in an itemised rating scale? Why?

11 Should an odd or even number of categories be used in an itemised rating scale?

12 How does the nature and degree of verbal description affect the response to itemised rating scales?

13 What is reliability? What are the differences between test–retest and alternative-forms reliability?

14 What is validity? What is criterion validity? How is it assessed?

15 How would you select a particular scaling technique?

Exercises

1 You work in the marketing 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 towards 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.

2 Develop three comparative (paired comparison, rank order and constant sum) scales to measure attitude towards five popular brands of beer (e.g. Heineken, Guinness, Carlsberg, Stella and Holsten). Administer each scale to five students. No student should be administered more than one scale. Note the time it takes each student to respond. Which scale was the easiest to administer? Which scale took the shortest time?

3 Develop a constant sum scale to determine preferences for restaurants. Administer this scale to a pilot sample of 20 students to determine their preferences for some of the popular restaurants in your town or city. Based on your pilot, evaluate the efficacy of the scale items you chose, and design new scale items that could be used for a full survey.

4 Design Likert scales to measure the usefulness of Renault's website. Visit the site at (**www.renault.com**) and rate it on the scales that you have developed. After your site visit, were there any aspects of usefulness that you had not considered in devising your scales, what were they and why were they not apparent before you made your site visit?

5 In a small group discuss the following issues: 'A brand could receive the highest median rank on a rank order scale of all the brands considered and still have poor sales' and 'It really does not matter which scaling technique you use. As long as your measure is reliable, you will get the right results.'

Notes

1 Newell, S. J., 'The development of a scale to measure perceived corporate credibility', *Journal of Business Research* (June 2001), 235; Gofton, K., 'If it moves measure it', *Marketing* (Marketing Technique Supplement) (4 September 1997), 17; Nunnally, J.C., *Psychometric Theory*, 2nd edn (New York: McGraw-Hill, 1978), 3.

2 Subabrata, B.B., 'Corporate environmentalism: the construct and its measurement', *Journal of Business Research* 55 (3) (March 2002), 177; Stevens, S., 'Mathematics, measurement and psychophysics', in Stevens, S. (ed.), *Handbook of Experimental Psychology* (New York: Wiley, 1951).

3 Moshkovich, H.M., 'Ordinal judgments in multiattribute decision analysis,' *European Journal of Operational Research* 137 (3) (16 March 2002), 625; Cook, W.D., Kress, M. and Seiford, L.M., 'On the use of ordinal data in data envelopment analysis', *Journal of the Operational Research Society* 44 (2) (February 1993), 133–140; Barnard, N.R. and Ehrenberg, A.S.C., 'Robust measures of consumer brand beliefs', *Journal of*

Marketing Research 27 (November 1990), 477–484; Perreault, W.D. Jr. and Young, F.W., 'Alternating least squares optimal scaling: analysis of nonmetric data in marketing research', *Journal of Marketing Research* 17 (February 1980), 1–13.

4 Halme, M., 'Dealing with interval scale data in data envelopment analysis', *European Journal of Operational Research* 137 (1) (February 16, 2002), 22; Lynn, M. and Harriss, J., 'The desire for unique consumer products: a new individual difference scale', *Psychology and Marketing* 14 (6) (September 1997), 601–616.

5 For a discussion of these scales, refer to Miller, D.C., and Salkind, N.J., *Handbook of Research Design and Social Measurement*, 6th edn (Thousand Oaks, CA: Sage, 2002); Taiwo, A., 'Overall evaluation rating scales: an assessment', *International Journal of Market Research* (Summer 2000), 301 –311; Coombs, C.H., 'Theory and methods of social measurement', in Festinger, L. and Katz, D. (eds), *Research Methods in the Behavioral Sciences* (New York: Holt, Rinehart & Winston, 1953).

6 Bastell, R.R. and Wind, Y., 'Product development: current methods and needed developments', *Journal of the Market Research Society* 8 (1980), 122–126.

7 There is, however, some controversy regarding this issue. See Campbell, D. T. and Russo, M.J., *Social Measurement* (Thousand Oaks, CA: Sage, 2001); Amoo, T., 'Do the numeric values influence subjects' responses to rating scales?', *Journal of International Marketing and Marketing Research* (February 2001), 41; Kang, M. and Stam, A., 'PAHAP: a pairwise aggregated hierarchical analysis of ratio-scale preferences', *Decision Sciences* 25 (4) (July/August 1994), 607–624.

8 Kellogg, D.L. and Chase, R.B., 'Constructing an empirically derived measure for customer contact', *Management Science* 41 (11) (November 1995), 1734–1749; Corfman, K.P., 'Comparability and comparison levels used in choices among consumer products', Journal of Marketing Research 28 (August 1991), 368–374.

9 Anon, 'Competition between Coca-Cola and Pepsi to start,' *Asiainfo Daily China News* (19 March 2002), 1; Rickard, L., 'Remembering New Coke', *Advertising Age* 66 (16) (17 April 1995), 6; 'Coke's flip-flop underscores risks of consumer taste tests', *Wall Street Journal* (18 July 1985), 25.

10 It is not necessary to evaluate all possible pairs of objects, however. Procedures such as cyclic designs can significantly reduce the number of pairs evaluated. A treatment of such procedures may be found in Bemmaor, A.C. and Wagner, U., 'A multiple-item model of paired comparisons: separating chance from latent performance', *Journal of Marketing Research* 37 (4) (November 2000), 514–524; Malhotra, N.K., Jain, A.K. and Pinson, C., 'The robustness of MDS configurations in the case of incomplete data', *Journal of Marketing Research* 25 (February 1988), 95–102.

11 For an advanced application involving paired comparison data, see Bemmaor, A.C. and Wagner, U., 'A multiple-item model of paired comparisons: separating chance from latent performance', *Journal of Marketing Research* 37 (4) (November 2000), 514 –524; Genest, C. and Zhang, S.S., 'A graphical analysis of ratio-scaled paired comparison data', *Management Science* 42 (3) (March 1996), 335–349.

12 Campbell, D. T. and Russo, M.J., *Social Measurement* (Thousand Oaks, CA: Sage, 2001); Likert, R., Roslow, S. and Murphy, G., 'A simple and reliable method of scoring the Thurstone Attitude Scales', *Personnel Psychology* 46 (3) (Autumn 1993), 689–690; Thurstone, L.L., *The Measurement of Values* (Chicago: University of Chicago Press, 1959). For an application of the case V procedure, see Malhotra, N.K., 'Marketing linen services to hospitals: a conceptual framework and an empirical investigation using Thurstone's case V analysis', *Journal of Health Care Marketing* 6 (March 1986), 43–50.

13 Daniles, E. and Lawford, J., 'The effect of order in the presentation of samples in paired comparison tests', *Journal of the Market Research Society* 16 (April 1974), 127–133.

14 Anon., 'Cranberry juice in a can', *Grocer* 225 (7538) (26 January 2002), 64; The Beverage Network, **www.bevnet.com**.

15 Bottomley, P.A., 'Testing the reliability of weight elicitation methods: direct rating versus point allocation,' *Journal of Marketing Research* 37 (4) (November 2000), 508–513; Herman, M.W. and Koczkodaj, W.W., 'A Monte Carlo study of pairwise comparison', *Information Processing Letters* 57 (1) (15 January 1996), 25–29.

16 Kerlinger, F., *Foundations of Behavioral Research*, 3rd edn (New York: Holt, Rinehart & Winston, 1973), 583–592.

17 Siciliano, T., 'Magnitude estimation', *Quirk's Marketing Research Review* (November 1999); Noel, N.M. and Nessim, H., 'Benchmarking consumer perceptions of product quality with price: an exploration', *Psychology & Marketing* 13 (6) (September 1996), 591–604; Steenkamp, J.-B. and Wittink, D.R., 'The metric quality of full –profile judgments and the number of attribute levels effect in conjoint analysis', *International Journal of Research in Marketing* 11 (3) (June 1994), 275–286.

18 Hayes, J.R., 'Issues in protocol analysis', in Ungson, G.R. and Braunste, D.N. (eds), *Decision Making: An Interdisciplinary Inquiry* (Boston, MA: Kent, 1982), 61–77.

19 For an application of verbal protocols, see Harrison, D.A., McLaughlin, M.E. and Coalter, T.M., 'Context, cognition and common method variance: psychometric properties and verbal protocol evidence', *Organizational Behavior and Human Decision Processes* 68 (3) (December 1996), 246–261; Gardial, S.F., Clemons, D.S., Woodruff, R.B., Schumann, D.W. and Bums, M.J., 'Comparing consumers' recall of prepurchase and postpurchase product evaluation experiences', *Journal of Consumer Research* 20 (March 1994), 548–560.

20 Mick, D.G., 'Levels of subjective comprehension in advertising processing and their relations to ad perceptions, attitudes, and memory', *Journal of Consumer Research* 18 (March 1992), 411–424; Wright, P.L., 'Cognitive processes mediating acceptance of advertising', *Journal of Marketing Research* 10 (February 1973), 53–62; Wright, P.L., 'Cognitive responses to mass media advocacy and cognitive choice processes', in Petty, R., Ostrum, T. and Brock, T. (eds), *Cognitive Responses to Persuasion* (New York: McGraw-Hill, 1978).

21 Murphy, I.P., 'RAMS helps Best Western tout worldwide positioning', *Marketing News* 31 (1) (6 January 1996), 25.

22 Amoo, T. and Friedman, H.H., 'Overall evaluation rating scales: an assessment,' *International Journal of market Research* 42 (3). (Summer 2000), 301–310; Albaum, G., 'The Likert scale revisited – an alternative version', *Journal of the Market Research Society* 39 (2) (April 1997), 331–348; Brody, C.J. and Dietz, J., 'On the dimensionality of 2-question format Likert attitude scales', *Social Science Research* 26 (2) (June 1997), 197–204; Likert, R., 'A technique for the measurement of attitudes', *Archives of Psychology* 140 (1932).

23 However, when the scale is multidimensional, each dimension should be summed separately. See Stanton, J.M., 'Issues and strategies for reducing the length of self-report scales,' *Personnel Psychology* 55 (1) (Spring 2002), 167–194; Aaker, J.L., 'Dimensions of brand personality', *Journal of Marketing Research* 34 (August 1997), 347–356.

24 Herche, J. and Engelland, B., 'Reversed-polarity items and scale unidimensionality', *Journal of the Academy of Marketing Science* 24 (4) (Fall 1996), 366–374.

25 Sethi, R., Smith, D.C. and Whan Park, C., 'Cross-functional product development teams, creativity and the innovativeness of new consumer products,' *Journal of Marketing Research* 38 (1) (February 2001) 73–85; Chandler, T.A. and Spies, C.J., 'Semantic differential comparisons of attributions and dimensions among respondents from 7 nations', *Psychological Reports* 79 (3 pt 1) (December 1996), 747–758.

26 Miller, D.C. and Salkind, N.J., *Handbook of research design and social measurement*, 6th edn. (Thousand Oaks, CA: Sage, 2002); Bearden, W.O. and Netemeyer, R.G., *Handbook of Marketing Scales: Multi-Item measures for marketing and consumer behaviour research* (Thousand Oaks, CA: Sage, 1999), 456–464; Millar, R. and Brotherton, C., 'Measuring the effects of career interviews on young people – a preliminary study', *Psychological Reports* 79 (3 pt 2) (December 1996), 1207–1215.

27 There is little difference in the results based on whether the data are ordinal or interval; however, see Nishisato, S., *Measurement and multivariate analysis* (New York: Springer-Verlag, New York, 2002); Gaiton, J., 'Measurement scales and statistics: resurgence of an old misconception', *Psychological Bulletin* 87 (1980), 567.

28 Ofir, C., 'In search of negative customer feedback: the effect of expecting to evaluate on satisfaction evaluations', *Journal of Marketing Research* (May 2001), 170–182; Reisenwitz, T.H. and Wimbush, G.J., Jr. 'Over-the-counter pharmaceuticals: exploratory research of consumer preferences toward solid oral dosage forms', *Health Marketing Quarterly* 13 (4) (1996), 47–61; Malhotra, S., Van Auken, S. and Lonial, S.C., 'Adjective profiles in television copy testing', *Journal of Advertising Research* (August 1981), 21–25.

29 Brady, M.K., 'Performance only measurement of service quality: a replication and extension', *Journal of Business Research* 55 (1) (January 2002), 17; Stapel, J., 'About 35 years of market research in the Netherlands', *Markonderzock Kwartaalschrift* 2 (1969), 3–7.

30 Hawkins, D.I., Albaum, G. and Best, R., 'Stapel scale or semantic differential in marketing research?', *Journal of Marketing Research* 11 (August 1974), 318–322; Menezes, D. and Elbert, N.E., 'Alternative semantic scaling formats for measuring store image: an evaluation', *Journal of Marketing Research* 16 (February 1979), 80–87.

31 Devellis, R.F., *Scale Development: Theories and Applications* (Thousand Oaks, CA: Sage, 1991); Etzel, M.J., Williams, T.G., Rogers, J.C. and Lincoln, D.J., 'The comparability of three Stapel scale forms in a marketing setting', in Bush, R.F. and Hunt, S.D. (eds), *Marketing Theory: Philosophy of Science Perspectives* (Chicago: American Marketing Association, 1982), 303–306.

32 Anderson, E.W., 'Foundations of the American customer satisfaction index', *Total Quality Management* 11 (7) (September 2000), 5869–5882; Coleman, A.M., Norris, C.E. and Peterson, C.C., 'Comparing rating scales of different lengths – equivalence of scores from 5-point and 7-point scales', *Psychological Reports* 80 (2) (April 1997), 355–362; Viswanathan, M., Bergen, M. and Childers, T., 'Does a single response category in a scale completely capture a response?', *Psychology and Marketing* 13 (5) (August 1996), 457–479; Cox, E.P., III, 'The optimal number of response alternatives for a scale: a review', *Journal of Marketing Research* 17 (November 1980), 407–422.

33 Dodge, Y., 'On asymmetric properties of the correlation coefficient in the regression setting', *The American Statistician* 55 (1) (February 2001), 51–54; Alwin, D.F., 'Feeling thermometers versus 7-point scales – which are better?', *Sociological Methods and Research* 25 (3) (February 1997), 318–340; Givon, M.M. and Shapira, Z., 'Response to rating scales: a theoretical model and its application to the number of categories problem', *Journal of Marketing Research* 21 (November 1984), 410–419; Stem D.E. Jr. and Noazin, S., 'The effects of number of objects and scale positions on graphic position scale reliability', in Lusch, R.E. *et al.*, *1985 AMA Educators' Proceedings* (Chicago, IL: American Marketing Association, 1985), 370–372.

34 Jones, B.S., 'Modeling direction and intensity in semantically balanced ordinal scales: an assessment of Congressional incumbent approval,' *American Journal of Political Science* 44 (1) (January 2000), 174; Watson, D., 'Correcting for acquiescent response bias in the absence of a balanced scale – an application to class-consciousness', *Sociological Methods and Research* 21 (1) (August 1992), 52–88; Schuman, H. and Presser, S., *Questions and Answers in Attitude Surveys* (New York: Academic Press, 1981), 179–201.

35 Morrel-Samuels, P., 'Getting the truth into workplace surveys,' *Harvard Business Review* 80 (2) (February 2002) 111; and Spagna, G.J., 'Questionnaires: which approach do you use?', *Journal of Advertising Research* (February–March 1984), 67–70.

36 McColl-Kennedy, J., 'Measuring customer satisfaction: why, what and how', *Total Quality Management* 11 (7) (September 2000), 5883–5896; Hasnich, K.A., 'The job descriptive index revisited: questions about the question mark', *Journal of Applied Psychology* 77 (3) (June 1992), 377–382; Schneider, K.C., 'Uninformed response rate in survey research', *Journal of Business Research* (April 1985), 153–162.

37 Amoo, T., 'Do numeric values influence subjects' responses to rating scales?', *Journal of International Marketing and Market Research* (February 2001), 41; Gannon, K.M. and Ostrom, T.M., 'How meaning is given to rating scales – the effects of response language on category activation', *Journal of Experimental Social Psychology* 32 (4) (July 1996), 337–360; Friedman, H.H. and Leefer, J.R., 'Label versus position in rating scales', *Journal of the Academy of Marketing Science* (Spring 1981), 88–92.

38 Alwin, D.F., 'Feeling thermometers versus 7-point scales – which are better?' *Sociological Methods and Research* 25 (3) (February 1997), 318–340.

39 For an example of a multi-item scale, see Brown, T., 'The customer orientation of service workers: personality trait effects on self and supervisor-performance ratings', *Journal of Marketing Research* 39 (1) (February 2002), 110–119; Mathwick, C., Malhotra, N.K. and Ridgon, E., 'Experiential value: conceptualization, measurement and application in the catalog and internet shopping environment,' *Journal of Retailing* 77 (2001), 39–56; Aaker, J.L., 'Dimensions of brand personality', *Journal of Marketing Research* 34 (August 1997), 347–356.

40 For example, see Flynn, L. R. and Pearcy, D., 'Four subtle sins in scale development: some suggestions for strengthening the current paradigm', *International Journal of Market Research* 43 (4) (Fourth Quarter 2001), 409–423; King, M.F., 'Social desirability bias: a neglected aspect of validity testing', *Psychology and Marketing* 17 (2) (February 2000), 79; Singhapakdi, A., Vitell, S.J., Rallapalli, K.C. and Kraft, K.L., 'The perceived role of ethics and social responsibility: a scale development', *Journal of Business Ethics* 15 (11) (November 1996), 1131–1140.

41 Borman, W.C., 'An examination of the comparative reliability, validity and accuracy and performance ratings made using computerised adaptive rating scales', *Journal of Applied*

Psychology 86 (5) (October 2001), 965; Kim, K. and Frazier, G.L., 'Measurement of distributor commitment in industrial channels of distribution', *Journal of Business Research* 40 (2) (October 1997), 139–154; Greenleaf, E.A., 'Improving rating scale measures by detecting and correcting bias components in some response styles', *Journal of Marketing Research* 29 (May 1992), 176–188.

42 The true score model is not the only theory of measurement. See Lord, E.M. and Novick, M.A., *Statistical Theories of Mental Test-Scores* (Reading, MA: Addison-Wesley, 1968).

43 Thompson, B., *Score reliability: Contemporary thinking on reliability issues* (Thousand Oaks, CA: Sage, 2002); Sinha, P., 'Determination of reliability of estimations obtained with survey research: a method of simulation', *International Journal of Market Research* 42 (3) (Summer 2000), 311–317; Wilson, E.J., 'Research design effects on the reliability of rating scales in marketing – an update on Churchill and Peter', *Advances in Consumer Research* 22 (1995), 360–365; Perreault, W.D. Jr and Leigh, L.E., 'Reliability of nominal data based on qualitative judgements', *Journal of Marketing Research* 25 (May 1989), 135–148; Peter, J.P., 'Reliability: a review of psychometric basics and recent marketing practices', *Journal of Marketing Research* 16 (February 1979), 6–17.

44 Campbell, D. T. and Russo, M.J., *Social Measurement* (Thousand Oaks, CA: Sage, 2001); Lam, S.S.K. and Woo, K.S., 'Measuring service quality: a test–re-test reliability investigation of SERVQUAL', *Journal of the Market Research Society* 39 (2) (April 1997), 381–396.

45 Hunt, D., Measurement and scaling in statistics, (London: Edward Arnold, 2001); Armstrong, D., Gosling, A., Weinman, J. and Marteau, T., 'The place of inter-rater reliability in qualitative research: an empirical study', *Sociology: The Journal of the British Sociological Association* 31 (3) (August 1997), 597–606; Segal, M.N., 'Alternate form conjoint reliability', *Journal of Advertising Research* 4 (1984), 31–38.

46 Cronbach, L.J., 'Coefficient alpha and the internal structure of tests', *Psychometrika* 16 (1951), 297–334.

47 Brown, T.J., Mowen, J.C., Donavan, D.T. and Licata, J.W., 'The customer orientation of service workers: personality trait effects on self – and supervisor performance ratings', *Journal of Marketing Research* 39 (1) (February 2002), 110–119; Peterson, R.A., 'A meta-analysis of Cronbach's coefficient alpha', *Journal of Consumer Research* 21 (September 1994), 381–391.

48 Chen, G., 'Validation of a new general self-efficacy scale', *Organizational Research Methods* 4 (1) (January 2001), 62–83; McTavish, D.G., 'Scale validity – a computer content analysis approach', *Social Science Computer Review* 15 (4) (Winter 1997), 379–393; Peter, J.P., 'Construct validity: a review of basic issues and marketing practices', *Journal of Marketing Research* 18 (May 1981), 133–145.

49 For further details on validity, see B. Keillor, 'A cross-cultural/cross national study of influencing factors and socially desirable response biases', *International Journal of Market Research* (1st Quarter 2001), 63–84; Sirgy, M.J., Grewal, D., Mangleburg, T.F., Park, J. *et al.*, 'Assessing the predictive ability of two methods of measuring self-image congruence', *Journal of the Academy of Marketing Science* 25(3) (Summer 1997), 229–241; Spiro, R.L. and Weitz, B.A., 'Adaptive selling: conceptualization, measurement, and nomological validity', *Journal of Marketing Research* 27 (February 1990), 61–69.

50 For a discussion of the generalisability theory and its applications in marketing research, see Middleton, K.L., 'Socially desirable response sets: the impact of country culture,' *Psychology and Marketing* (February 2000), 149; Abe, S., Bagozzi, R.P. and Sadarangani, P., 'An investigation of construct validity and generalizability in the self concept: self consciousness in Japan and the United States', *Journal of International Consumer Marketing* 8 (3,4) (1996), 97–123; Rentz, J.O., 'Generalisability theory: a comprehensive method for assessing and improving the dependability of marketing measures', *Journal of Marketing Research* 24 (February 1987), 19–28.

51 Myers, M., 'Academic insights: an application of multiple-group causal models in assessing cross-cultural measurement equivalence,' *Journal of International Marketing* 8 (4) (2000), 108–121; Hinkin, T.R., 'A review of scale development practices in the study of organisations', *Journal of Management* 21 (5) (1995), 967–988.

52 Devlin, S.J., Dong, H.K. and Brown, M., 'Selecting a scale for measuring quality', *Marketing Research* (Fall 2003), 13–16.

53 Page Fisk, A., 'Using individualism and collectivism to compare cultures – a critique of the validity and measurement of the constructs: Comment on Oyserman', *Psychological Bulletin* 128 (1) (January 2002), 78; Mullen, M.R., Milne, G.R. and Didow, N.M., 'Determining cross-cultural metric equivalence in survey research: a new statistical test', *Advances in International Marketing* 8 (1996), 145–157; Gencturk, E., Childers, T.L. and Ruekert, R.W., 'International marketing involvement – the construct, dimensionality, and measurement', *Journal of International Marketing* 3 (4) (1995), 11–37.

54 Unikel, A.L., 'Imitation might be flattering, but beware of trademark infringement,' *Marketing News* 21 (19) (11 September 1997), 200–221; Mckay, B., 'Xerox fights trademark battle', *Advertising Age International* (27 April 1992), 1–39.

55 Zuber, A., 'Pizza chains top customer satisfaction poll', *Nation's Restaurant News* 36 (9) (4 March 2002), 4–5 and **www.dominos.com**.

Visit the *Marketing Research* Companion Website at www.pearsoned.co.uk/malhotra_euro for additional learning resources including annotated weblinks, an online glossary and a suite of downloadable video cases.