Creating appropriate rating scales

06

Introduction

Rating scales are a common research tool for investigating a respondent's opinion or attitude. A simple dichotomous question may sometimes be sufficient ('Do you like or dislike this?', 'Do you agree or disagree that…?', 'Is this important or unimportant to you?'). However, frequently this approach might be over simplistic. There are often likely to be degrees of strength of feeling as attitudes and opinions can be complex. Rating scales, with scale points designed to reflect these shades of feeling, can give greater sensitivity to differences between respondents or between items that are being assessed. Rating scales are widely used by questionnaire writers. They provide a straightforward way of asking attitudinal information that is easy and versatile to analyze, and that provides comparability across time. However, there are many different types of rating scales, and there is skill in choosing which is most appropriate for a given task. In this chapter we look at the types of scales and their applications. The measurement of attitudes more generally is discussed in Chapter 8.

Itemized rating scales

The most commonly used approach is the itemized ratings scale. The researcher first develops a number of dimensions (eg attitude statements, product or service attributes, image dimensions, etc). Respondents are then asked to position how they feel about each one using a defined rating scale, usually an interval scale (see Chapter 5) with a range of evenly spaced points.

Figure 6.1 shows two typical examples: the wording on each scale is tailored to be appropriate to the question, and all have five points representing

a gradation from positive to negative. They are balanced around a neutral mid-point with equal numbers of positive and negative statements for the respondent to choose from.

Figure 6.1 Some examples of itemized rating scales

How likely are you to use the train for this journey in the near future?

Very likely	O
Quite likely	O
Neither likely nor unlikely	O
Quite unlikely	O
Very unlikely	O
Don't know	O

How effective are the management in this organization?

Highly effective	O
Effective	O
Neither effective nor ineffective	O
Not very effective	O
Not at all effective	O

Being interval data, scores can be allocated to each of the responses to assist in the analysis of responses. The allocated scores are most likely to be from 1 to 5, from the least to the most positive; or from –2 to +2, from the most negative to the most positive with the neutral point as zero.

Think ahead to whether you need to make comparisons with data from elsewhere. Consistency is often the most important factor in rating scale decisions.

Balanced scales

It is usual to balance scales by including equal numbers of positive and negative attitudes. Consider this balanced scale when asking respondents to describe the taste of a product:

Very good

Good

Average

Poor

Very poor

With two positive and two negative statements the respondents are not led in either direction. However, if the scale were as follows, the three positive dimensions would tend to result in a higher number of total positive responses:

Excellent

Very good

Good

Average

Poor

In most circumstances it is important to balance the scale to avoid this bias. However, there are occasions when an unbalanced scale can be justified. Where it is known that the response will be overwhelmingly in one direction, more categories may be given in that direction to achieve better discrimination.

This is often the case when measuring the importance of various aspects of service in customer satisfaction research. Few customers will say that any are unimportant – the customers will be looking for the best service that they can get – and the dimensions about which we ask are the ones that we believe are important anyway. The objective is mainly to distinguish between the most important aspects of service and the less important ones. An unbalanced scale might therefore be used, offering just one unimportant option, but several degrees of importance:

Extremely important

Very important

Important

Neither important nor unimportant

Not important

Here the questionnaire writer is trying to obtain a degree of discrimination between the levels of importance. The visual mid-point is 'important', and the scale implicitly assumes that this will be where the largest number of responses will be placed. The scale could have seven points extending from 'extremely unimportant' to 'extremely important' to preserve the balance,

but if we are confident they are unlikely to be used these balancing points simply add visual clutter. They may also provoke a tendency to avoid the extremes when scales become long, thus counteracting the increased sensitivity we are trying to achieve at the top end of the scale.

Unbalanced scales should only be used for a good reason and by researchers who know what the impact is likely to be.

Number of points on the scale

The illustrations in Figure 6.1 show five-point scales, which are probably the most commonly used. A five-point scale gives sufficient discrimination for most purposes and is easily understood by respondents. The size of the scale can be expanded to seven points if greater discrimination is to be attempted. Then the scale points can be written as:

Extremely likely

Very likely

Quite likely

Neither likely nor unlikely

Quite unlikely

Very unlikely

Extremely unlikely

Or:

Excellent

Very good

Good

Neither good nor poor

Poor

Very poor

Extremely poor

There is little agreement as to the optimum number of points on a scale. The only agreement is that it is between 5 and 10 (or 11). Seven is considered the optimal number by many researchers for an item-specific scale (Krosnick and Fabrigar, 1997) but there is a range of opinions on this issue and whether extending the number to 10 or more increases the validity of the data. Numeric alternatives to itemized scales provide more flexibility for more

points as there is no need to create appropriate labels for each point. Coelho and Esteves (2007) have demonstrated that a 10-point numeric scale is better than a five-point scale in that it transmits more of the available information, without encouraging response error – the characteristic given by Cox (1980) for assessing the optimum number of points. They hypothesize that, among other things, consumers may be more used these days to giving things scores out of 10 and are able to cope with them better than was the case 20 years ago. However, Revilla, Saris and Krosnick (2014) conclude that five points are the optimum for fully labelled agree-disagree scales.

The questionnaire writer's decision as to the number of points on the scale has to be taken with regard to the degree of discrimination that is sought, the feasibility of creating meaningfully distinct labels for those points, and the ability of respondents to discriminate in that much detail. With telephone interviewing, scales with more than five itemized points are difficult for respondents to remember and therefore numeric alternatives are often preferred. With multi-country surveys the feasibility of creating equally spaced itemized scales in different languages also points towards greater use of numeric scales instead (as discussed later in this Chapter).

'Don't knows' and mid-points

In Figure 6.1, each of the scales is balanced around a neutral mid-point; this is included to allow a response for people who have no strong view either way. However, this point is also frequently used by respondents who want to give a 'don't know' response but are not offered 'don't know' as a response category and do not want, or are unable, to leave the response blank.

The reluctance of respondents to leave a scale blank where they genuinely cannot give an answer has always been an issue with self-completion interviews. Unpublished work from TNS BMRB shows that up to three-quarters of those who choose the mid-point may be using it as a substitute for 'don't know', although this varies by the attribute or attitude asked about. However, 'don't know' codes or boxes are frequently not provided as the questionnaire writer is wary of prompting this as a response – instead wanting to encourage the respondent to commit to a response that, in all likelihood, may reflect an attitude unrecognized at a conscious level. In studies where it would be expected that most people would have a view, for example about crime, it can be argued that they hold a view even if they do not recognize that they do. It is therefore legitimate, it is argued, to force a response in one direction or the other. When the subject is breakfast cereals however, it must be recognized

that many people may really have no opinion one way or the other. The response points for a scale without a mid-point might look like this:

Extremely likely
Very likely
Quite likely
Quite unlikely
Very unlikely
Extremely unlikely

Or:

Excellent
Very good
Good
Poor
Very poor
Extremely poor

In an interviewer-administered study it is possible to accept a neutral response that is offered spontaneously by the respondent. However, studies have shown that including a neutral scale position significantly increases the number of neutral responses compared to accepting them spontaneously (Kalton et al, 1980; Presser and Schuman, 1980). This indicates that eliminating the neutral mid-point does increase the commitment of respondents to be either positive or negative. This is supported by Coelho and Esteves (2007), who found that the mid-point was used by respondents who are trying to reduce the effort, and so exaggerated the true mid-point score, and by Saris and Gallhofer (2007) who showed that not providing a neutral mid-point improves both the reliability and the validity of the data.

Further complications to the debate include that non-response to one scale among a battery of scales can raise issues of how to treat the data when using certain data analysis techniques. And a practical consideration is that digital scripting software often does not allow respondents to pass to the next question unless an answer of some kind is provided – reinforcing the need for a 'don't know' code if no mid-point is provided.

Figure 6.2 shows an alternative order to typical scales that places the mid-scale neutral element at the end of the options. In this case the question writer took this decision because of the subject matter, ie advertising. There

is a tendency to deny being influenced by advertising. By offering the four statements that acknowledge advertising influence together as a block, the visual impact will be such that respondents will be more prepared to consider that they may indeed be influenced. The questionnaire writer has attempted to offset one bias with another. While this could be taking a risk, in this instance the question writer felt there was good reason for doing so based on their previous experiences.

Figure 6.2 An alternative order for responses

Based on this ad, how likely will you be to purchase this product in the future?

Please select one.

Much more likely to buy it	O
Somewhat more likely to buy it	O
Somewhat less likely to buy it	O
Much less likely to buy it	O
The ad had no effect on my likelihood to buy it	O

In conclusion, since the purpose of using ratings scales (as an alternative to a simple dichotomous 'either/or') is usually to create greater sensitivity to differences, some feel it is at odds with this aim to offer a mid-point that might be used as an opt-out answer. However, mid-points continue to be widely used and the questionnaire writer must decide whether or not including one is appropriate for the particular question and subject matter. Comparability with other data will often have greater import.

Anchor strength

With all semantic scales, the wording of the anchor statement is crucial to the distribution of data that is likely to be achieved. A five-point bi-polar scale that goes from 'extremely satisfied' to 'extremely dissatisfied' is likely to discourage respondents from using the end-points and to concentrate the distribution on the middle three points. If the end-points were 'very satisfied' and 'very dissatisfied', they would be used by more respondents and the data would be more widely distributed across the scale. This can make the data more discriminatory between items. As a general rule, the stronger the anchors, the more points are required on the scale to obtain discrimination.

Likert scale

A form of itemized rating scale developed specifically to measure attitudes is the Likert scale (frequently known as an 'agree/disagree' scale). This was first published by psychologist Rensis Likert in 1932. The technique presents respondents with a series of attitude dimensions (an 'attitude battery'), for each of which they are asked whether (and how strongly) they agree or disagree, using one of a number of positions on a five-point scale (see Figure 6.3). It is increasingly common to find any type of attitudinal rating scale – regardless of the number of points – referred to as a Likert scale. Many DIY online survey providers tend to do this – probably for simplicity. Technically, however, it refers only to this specific scale.

Figure 6.3 Use of the Likert scale

Do you agree or disagree with these attitudes about shopping?

	Disagree strongly	Disagree	Neither agree nor disagree	Agree	Agree strongly
Being a smart shopper is worth the extra time it takes.	❒	❒	❒	❒	❒
Which brands I buy makes little difference to me.	❒	❒	❒	❒	❒
I take advantage of special offers.	❒	❒	❒	❒	❒
I like to try new brands.	❒	❒	❒	❒	❒
I like to shop around and look at displays.	❒	❒	❒	❒	❒

The technique is easy to administer online. It can be presented in a number of ways including radio buttons, slider scales, stars or with a range of other graphical techniques.

With face-to-face interviewer-administered scale batteries, the responses may be shown on a card while the interviewer reads out each of the statements in turn. With telephone interviewing, the respondent may sometimes

be asked to remember what the response categories are, but preferably would be asked to write them down.

Responses using the Likert scale can be given scores for each statement, usually from 1 to 5, negative to positive, or –2 to +2. As this is interval data, means and standard deviations can be calculated for each statement.

The full application of the Likert scale is to sum the scores for each respondent to provide an overall attitudinal score for each individual. Likert's intention was that the statements would represent different aspects of the same attitude. The overall score, though, is rarely calculated in commercial research (Albaum, 1997), where the statements usually cover a range of attitudes. The responses to individual statements are of more interest in determining the specific aspects of attitude that drive behaviour and choice in a market, or summations made over small groups of items. The data will tend to be used in principal component or factor analysis, to identify groups of attitudinal statements that have similar response patterns and that could therefore represent underlying attitudinal dimensions. Factor analysis can be used to create a factor score for each respondent on each of the underlying attitudinal dimensions, thereby reducing the data to a small number of individual scores.

There are four interrelated issues that questionnaire writers must be aware of when using Likert scales:

1. order effect;
2. acquiescence;
3. central tendency;
4. pattern answering.

The **order effect** arises from the order in which the response codes are presented. It has been shown (Artingstall, 1978) that there is a bias to the left on a self-completion scale presented horizontally. (Order effects are returned to in Chapter 9.)

Acquiescence is the tendency for respondents to say 'yes' to questions or to agree rather than disagree with statements (Kalton and Schuman, 1982). In Figure 6.3, the negative end of the scale is placed to the left, to be read first. With the 'agree' response to the left, the order effect and acquiescence would compound each other. With the 'disagree' response to the left, there is a possibility of the biases going some way to cancelling out each other. Importantly, it has been shown that acquiescence bias tends to be consistent for individual respondents. If measures can be found to assess the bias for each respondent, then corrections can be made. This, though, can be a complex and time-consuming exercise (Weijters et al, 2010).

Central tendency or extreme response bias is the reluctance of respondents to use extreme positions. Greenleaf (1992) showed that, like acquiescence bias, the extreme response bias is consistent within a respondent's answers. He also showed that it is related to age, income and education, but not to gender. It has been shown (Albaum, 1997) that a two-stage question elicits a higher proportion of extreme responses. This investigation used the question:

For each of the statements listed below, indicate first the extent of your agreement and second how strongly you feel about your agreement.

- A product's price will usually reflect its level of quality.

  Agree – Neither Agree nor Disagree – Disagree
- How strongly do you feel about your response?

  Very Strong – Not Very Strong

The question arises, of course, as to whether the two-stage approach is a better measure of the attitude or whether it creates its own bias towards the extreme points. Albaum et al (2007) explored this issue by correlating reported attitude to actual behaviour in charity giving. The results were not conclusive but suggested that the two-stage approach provides the truer reflection of attitudes.

With a large number of dimensions to be evaluated, this may be too time-consuming for most studies, but the questionnaire writer should be aware of this approach and of the different response patterns it is likely to give. This approach is particularly appropriate for telephone interviewing, where the complete scale cannot be shown.

Pattern answering occurs when a respondent falls into a routine of ticking boxes in a pattern, which might be straight down the page or diagonally across it. It is often a symptom of fatigue or boredom. Some online providers look at the time taken to complete such a page. Speeding through is taken as evidence of pattern answering. The best way to avoid it is to keep the interview interesting and reduce the number of items. Some advocate using both positive and negative statements so the respondent then has to read them or listen to them carefully to understand the polarity and to give consistent answers. However, additional analysis is likely to be needed to identify conflicting answers, and decisions will need to be made about how to deal with that respondent. It is also not always possible to be sure that answers really conflict. Therefore, others favour keeping consistent polarity and accepting the risk of some pattern-answering rather than subjective judgment about whether the respondent is likely to have spotted the reversal or not.

Saris et al (2005) argue that agree/disagree scales are flawed not just because of these issues but because the cognitive process involved for the respondent is more complex and burdensome than with a simpler question asked directly about the specific issue. Such construct-specific questions (Figure 6.4) are also believed to suffer less from acquiescence and order bias. This is supported by unpublished work by TNS BMRB, which looked at a number of constructs where the agree/disagree scale could be replaced by a construct-specific scale. Here it was found that while there were significant differences between the responses to end points on the agree/disagree scale when rotated between respondents, demonstrating order bias, the construct-specific scale showed far more consistency, indicating less bias.

Figure 6.4 Labelled construct-specific scale

Did you find this orange juice:

Much too sweet	O
A little too sweet	O
About right	O
Not quite sweet enough	O
Not nearly sweet enough	O

It should be noted that the European Social Survey no longer uses a Likert scale for new questions. Nevertheless, it continues to be widely used because it is simple to create.

Semantic differential scale

The semantic differential scale is a bi-polar rating scale. It differs from the Likert scale in that opposite statements of the dimension are placed at the two ends of the scale and respondents are asked to indicate which they most agree with by placing a mark along the scale. This has the advantage that there is then no need for the scale points to be individually identified. Any bias towards agreeing with a statement is avoided, as both ends of the scale have to be considered. The original development of this scale by Osgood (Osgood et al, 1957) recommended the use of seven points on the response scale, and this number continues to be the favourite of researchers (McDaniel and Gates, 1993), although both five-point and three-point scales are used for particular purposes (Oppenheim, 1992).

With semantic differential scales the statements should be kept as short and precise as possible because of the need for the respondent to read and understand fully both ends of the scale. Attitudes can be difficult to express concisely, and it is sometimes hard to find an opposite to ensure that the scale represents a linear progression from one end to the other. For these reasons semantic differential scales are usually better suited to descriptive dimensions.

Care must be taken to ensure that the two statements determine the dimension that the researcher requires. The opposite of 'modern' might be 'old-fashioned' or it might be 'traditional'. The opposite of 'sweet' might be 'savoury' or 'sour' or 'bitter'. This forces the questionnaire writer to consider exactly what the dimension is that is to be measured. This gives the semantic differential scale an advantage over the Likert scale where disagreeing with 'the brand is modern' could mean that the brand is seen as either old-fashioned or traditional, and the researcher does not know which.

Figure 6.5 comes from an advertising study, taken from a face-to-face questionnaire where the interviewer would read out much of the text. Online, this would be much simpler (as shown in Figure 6.6). The format is so simple and familiar to respondents that it may not be necessary to explain or label the scale points. Note the difficulty that the questionnaire writer has

Figure 6.5 Example of a semantic differential scale (Interviewer-administered)

Below are pairs of statements. Each one may or may not apply to the advertisement that you have just seen. Please read each pair and indicate which of the statements you agree applies to the ad by ticking one box for each pair of statements.

For example, if you agree strongly that the advertisement was 'mundane', you would tick the box closest to that statement, but if you only agreed slightly, then you should tick a box further away from the statement.

Example

Fascinating	❒	❒	❒	❒	❒	☑	❒	Mundane

Please complete the remaining items according to how you feel about the ad:

Boring	❒	❒	❒	❒	❒	❒	❒	Interesting
Important	❒	❒	❒	❒	❒	❒	❒	Unimportant
Relevant	❒	❒	❒	❒	❒	❒	❒	Irrelevant
Exciting	❒	❒	❒	❒	❒	❒	❒	Unexciting
Unappealing	❒	❒	❒	❒	❒	❒	❒	Appealing
Involving	❒	❒	❒	❒	❒	❒	❒	Uninvolving
Means nothing	❒	❒	❒	❒	❒	❒	❒	Means a lot to me

Scale items taken from Zaichkowsky (1999).

in achieving exact opposites in the first pair of statements. The ad may be worth remembering because it contains useful information, but that does not necessarily mean that it is not also easily forgettable. The questionnaire writer could have included both of the pairs: 'worth remembering – not worth remembering' and 'easy to forget – difficult to forget' but has chosen to force a decision between two statements that are not strictly opposites in order not to have to extend the number of pairs asked about.

Figure 6.6 Example of a semantic differential scale (online self-completion).

How did you feel about this ad?

For each pair of statements click closest to the one that best describes how you felt about it

Worth remembering	O	O	O	O	O	O	Easy to forget
Difficult to relate to	O	O	O	O	O	O	Easy to relate to
Lively, exciting or fun	O	O	O	O	O	O	Dull
Ordinary or boring	O	O	O	O	O	O	Clever or imaginative
Helps to make the brand different to others	O	O	O	O	O	O	Does not make the brand any different to others
Makes me less interested in the brand	O	O	O	O	O	O	Makes me more interested in the brand

Note that the questionnaire writer alternated positive and negative ends of the scale between statements to help catch the flatliners. But dimensions three and four contain potential ambiguities.

Numeric scales

A simple form of scaling is to ask respondents to award a score (eg 'out of 5', 'out of 10' or even 'out of 100'). The end points of the scale should be semantically anchored to avoid misunderstanding. It should also be made clear whether the bottom point is 0 or 1 (Figure 6.7).

- Please give us a score out of ten for how well we performed today – where 10 is good and 1 is poor.

In practice, whether a 10-point scale starts at 0 or 1 makes little difference to the distribution of the responses. To have 0 as the lowest point on the scale as is generally preferred in case there is any ambiguity as to the direction of the scale as it gives a more explicit mid-point (5). The recommended scale for the widely used Net Promoter Score (NPS) is 0 to 10 (Reicheld, 2003).

Numeric scales (Figure 6.7) are simpler to design than itemized scales where the exact language used for each scale point needs to be considered. Therefore, they are attractive for multi-country studies to avoid challenges with consistent translations. When a telephone interviewer is administering the questions,

the scale can easily be understood by the respondent without the need to remember or write down the scale point options. They take up little space which can be important for modes where this is limited (eg on a mobile phone screen).

Figure 6.7 A numeric scale question

Give *The Gingerbread Store* marks out of 100 on each of the following:
100 would mean 'perfect' and 0 'dreadful'.

The layout ☐

Helpfulness of staff ☐

Attractiveness of store ☐

However, interpretation is not always straightforward (eg in determining how people feel in absolute terms: how good is a 7 out of 10?) but where comparisons are made with previous scores or benchmarks it works well. The researcher must also remember that this is an interval scale and not a ratio scale. A score of 8 out of 10 does not mean that something is twice as good or twice as important as a score of 4. Numeric scales are not appropriate for indicating choice between two brands, because the more positive associations implicit in the higher score would bias response towards that option. Finally, a questionnaire with a large number of numeric scales can start to feel quite clinical or abstract with the risk of the respondent becoming disengaged.

Figure 6.8 Advantages and disadvantages of main types

Itemized Rating Scale	
When to use:	When absolute knowledge is required.
Advantage:	Precision of response, for both respondent and analyst.
Disadvantages:	Scale point wordings often differ between items, requiring separate questions (except Likert scale).
Semantic Differential	
When to use:	When making comparisons between items.
Advantages:	End points understood. No need to find gradations of meaning for the scale.
Disadvantages:	Requires precision in finding opposites. We cannot know what the points on the scale actually mean.
Numeric	
When to use:	When comparing with a database or over time
Advantages:	Simple to administer. Simple to understand.
Disadvantage:	Lack of consistency of interpretation by respondents.

Stapel scale

Named after Jan Stapel, in the Stapel scale the dimension or descriptor is placed at the centre of a scale that ranges from –5 to +5. Respondents indicate whether they agree positively or negatively with the statement, and how strongly, by selecting one of the points on the scale (see Figure 6.9). Thus, it is a form of numeric scale with both positive and negative scores.

Figure 6.9 A Stapel scale

Please indicate how accurately you feel each of the following words and phrases describes the Gingerbread Store. Select a positive number for the phrases you think describe the store accurately. The more accurately you think it describes it, the larger the number you should choose. Select a minus number for the phrases you think do not describe it accurately. The less accurately you think the phrase describes the store, the larger the negative number you should choose.

The Gingerbread Store

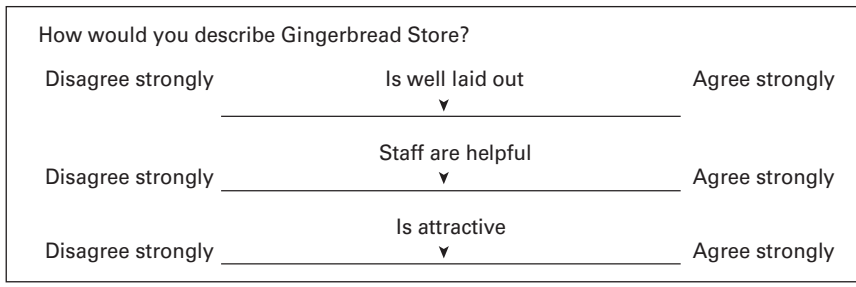
+5	+5	+5
+4	+4	+4
+3	+3	+3
+2	+2	+2
+1	+1	+1
is well laid out	has helpful staff	is attractive
–1	–1	–1
–2	–2	–2
–3	–3	–3
–4	–4	–4
–5	–5	–5

The advantage of this type of scale and other numeric scales over semantic differential scales is that it is not necessary to find an accurate opposite to each dimension to ensure bi-polarity. The data can, however, be analyzed in the same way as semantic differentials, and the scale, with 10 points, has the potential to provide greater discrimination than a five-point scale. By having no centre point, these scales also avoid the issue of whether or not there should be an odd or even number of points on the scale.

Online, this is relatively simple to administer, (Figure 6.10). A slider scale replaces the numbered points and a semantic label indicates the end points. The use of it is very intuitive, and a large amount of text is done away with.

With face-to-face or telephone interviewing, however, they are not widely used as they are thought to be confusing for respondents.

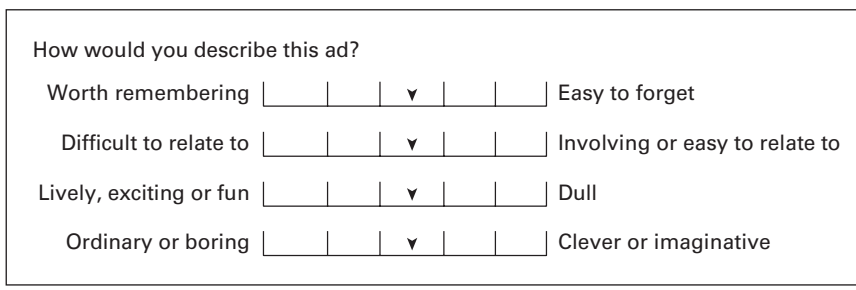
Figure 6.10 An online Stapel scale



Graphic scales

A graphic scale is one presented to the respondents visually so that they can select a position on it that best represents their desired response. In its most basic form it looks like a slider bi-polar scale with fixed points verbally anchored at either end. Here, in Figure 6.11, it is used to replace the radio buttons in a semantic differential scale.

Figure 6.11 Semantic differential slider scale



The distance from the end points of the respondent's marks is measured to provide the score for each attitudinal dimension. Essentially this is a continuously rated semantic differential scale, which provides a greater degree of precision and avoids the issue of numbers of points on the scale. It is a simple way of measuring attitudes and image perceptions but it is usually only practical online.

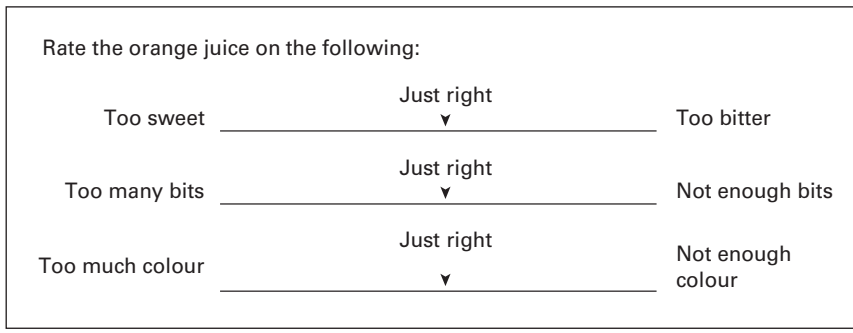
Although the data collected is continuous, the measurements will be assigned to categories and treated as interval data for analysis purposes. It is possible to have a large number of very small intervals. Some online DIY survey providers offer a choice of whether it is treated as 0 to 5, 0 to 10, 0 to 100 or whatever length scale in between that you wish. These points can often be displayed to the respondent if desired, which effectively then turn this into a numeric scale. The researcher must decide at what level the apparent accuracy of the data becomes spurious. That will depend on the length of the line used, the

accuracy with which respondents are able to place the cursor, and the degree of accuracy with which respondents are likely to have tried to place the cursor.

With some software, it is possible to place several cursors or brand logos, on the same scale on screen (see Chapter 11) so that the respondent can position them relative to each other.

We have already seen the slider scale in use as a Stapel scale (Figure 6.10). In a specific application, it can be used for new product development to rate products on specific constructs or attributes (Figure 6.12). Here a consistent centre point descriptor has been added, and the scoring will go from –50 to +50.

Figure 6.12 Semantic slider with mid-point



Compare this to the same question shown in Figure 6.4. While the slider scale is better at allowing product developers to see how much they need to adjust their product to meet expectations than would be case with a numeric scale, the labelling of the points in Figure 6.4 may provide a better indication of what the scores actually mean.

Use fully labelled construct-specific scales for key questions (as this type of scale is easier to interpret) and slider scales (which just have the end-points labelled) for quick reads on lower priority measures.

Visual analogue scales (VAS), require the respondent to place a mark or indicator at a point on the line joining two end points. They thus appear similar to slider scales, but are less frequently found in online surveys than slider scales. They are rarely offered by the online DIY survey providers. This is despite the fact that they require fewer actions by the respondent (point and click, as opposed to grab, move and release) and so should reduce the load on respondents, particularly where there are a number of scales to be answered.

It has been shown (Thomas et al, 2007) that in online surveys, respondents found visual analogue scales as easy to complete as scales using fixed points

denoted by radio buttons, and that they felt that VAS scales conveyed their responses with greater accuracy than with a numeric box entry. This view was supported by Cape (2009) with regard to slider scales. Cape also showed that respondents found the slider scale approach more interesting than the radio buttons, a finding supported by others (Roster, Luciano and Albaum, 2015).

Slider scales are popular in online surveys because of their simplicity, but care needs to be taken with them. There may be issues with software compatibility which means that they do not always display properly. There is evidence (Funke, 2016) that they are less easy to cope with on mobile phones and negatively affect completion rates.

Pictorial scales

In many instances, it is desirable to avoid using semantic scales in favour of pictorial representations:

- where the target population is children who are unable to relate their responses to verbal descriptors;
- where there are cultural differences between sub-groups of the target population that may mean that they interpret descriptors differently;
- with multi-country studies where translation of descriptors may alter shades of meaning;
- where there is a low level of literacy in the target population.

A common solution to this is the use of smiley or smiling face scales. A range of smiles and down-turned mouths is used to indicate that the respondent agrees (or is happy) with the statement or disagrees (or is unhappy) with the statement (see Figure 6.13).

Figure 6.13 Smiley scale

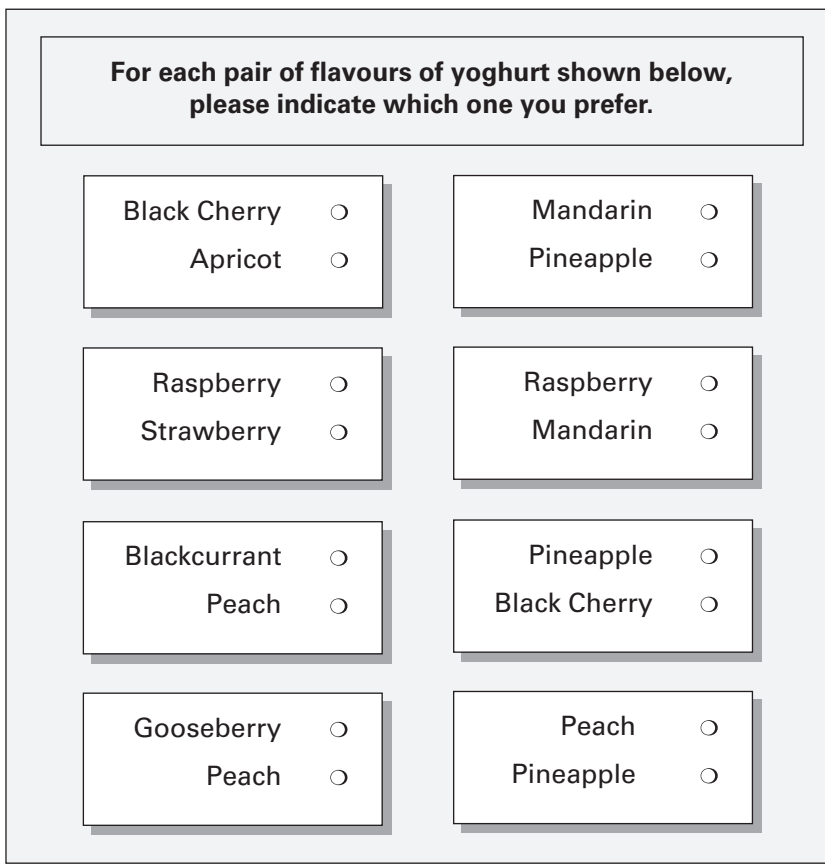
Comparative scaling techniques

Paired comparisons

With paired comparisons, respondents are asked to choose between two items based on the appropriate criterion (eg that one is more important than the

other, or preferred to the other). This can be repeated with a number of pairs chosen from a set of items, such that every item is compared against every other item (see Figure 6.14). Summing the choices made provides an evaluation of importance or preference across all of the items. The task is often easier and quicker for respondents than being asked to rank-order a list of items, because the individual judgements to be made are simpler. By careful rotation of the pairs, some of the order bias inherent in showing lists can be avoided.

Figure 6.14 Paired comparison



The disadvantage of this technique is that it is limited to a relatively small number of items. With just six items, 15 pairs are required if each is to be assessed against every other, and the number of pairs required increases geometrically. With 190 possible pairs from a list of 20 items, clearly no respondent can be shown all of them. A balanced design of the pairs shown to each respondent can provide sufficient information for the rank order of each item to be inferred.

Constant sum

With a constant sum technique, respondents are asked to allocate a fixed number of points between a set of options to indicate relative importance or relative preference. The number of points given to each option reflects the magnitude of the importance, from which we can also deduce the rank order of the options for each respondent (see Figure 6.15). Some respondents are likely to have problems with a constant sum question, as it requires some effort and mental agility on their part, both to think simultaneously across all of the items and to do the mental arithmetic.

Figure 6.15 Constant sum technique

Following is a list of items that might or might not be important to you when choosing a new car. Allocate 100 points across these five items according to how important they are to you when choosing a new car.

The engine size	☐
The colour	☐
Manual or automatic gearbox	☐
Quality of the radio/CD player	☐
Country of manufacture	☐
	100

It is easier online, where the scores allocated can be automatically summed and the respondent not allowed to move on until exactly 100 points have been allocated. However, the need to make simultaneous comparisons between a number of different items still remains. As the number of items increases, it becomes more difficult to think through and to mentally keep a running total of the scores, so this works best where a running total can be displayed.

Another way of asking this is to use a constant sum approach combined with paired comparisons. In another example, the task for respondents had been reduced to making comparisons between 10 pairs of items. Dealing with pairs is usually easier for respondents to manage. Respondents are asked to allocate 11 points between each pair. An odd number has been chosen so that the two items in any pair cannot be given the same number of points; this forces a distinction between them. Had the respondents been asked to allot 10 points per pair, this would have allowed items in a pair to be given equal weight of five points each. This technique can be used equally well for comparing preferences for products, when forcing even small distinctions can be important to the researcher.

Item sorting

When the number of objects is large, say more than 30, then a prior sorting approach can help make a ranking task manageable. Online, the respondent is asked to sort the items into a number of categories. These might be levelled by importance from 'very important' to 'not at all important'. This can be done using a drag-and-drop technique. The following screens show the items that have been put into a category, and the respondent is asked to rank order them. This is repeated for each category. In face-to face interviews a similar process is followed with each item presented on a card.

In this way, the combination of rating and ranking can produce an item scoring system that provides good discrimination across a large number of items.

Q sort

A similar approach designed for larger numbers of attributes (eg 100) is Q sorting.

The objects are sorted by respondents into a number of categories, usually 11 or 12, representing the degrees on the scale, such as appeal or interest in purchase. Respondents may be instructed to place a specific number of objects on each point of the scale so that they are distributed approximately according to a normal distribution. They are asked to put a few objects at the extremes of the scale, with increasing numbers towards the middle of the scale. Objects placed in the two extreme positions can then be rank-ordered by the respondent for increased discrimination.

Using just five scale points and 10 attributes, Chrzan and Golovashkina (2006) showed that the Q sort technique produced results that were better than several other techniques in terms of discrimination and prediction, and was quicker to administer than most. This technique is primarily suited to face-to-face interviewing.

CASE STUDY Whisky usage and attitude

Rating scales

At Q23, we need to ask the relative importance of whisky attributes when considering which brand to buy. The attributes we have are:

- depth of colour;
- smoothness of taste;

- familiarity with brand;
- distinctiveness from other brands;
- tradition associated with brand.

There are a number of ways in which we might consider asking this:

- Rating of attribute for importance. This, however, is likely to give poor discrimination because most things will be rated as important.
- Ranking of attributes. This will tell us how important each is relative to another, but not how much more important. We will know the order of importance, but not the distance between them.
- Item sort or Q sort are not appropriate because of the relatively fewer number of attributes.

We settle on using paired comparison of attributes, rotating the attributes to cover all pairs. With five attributes, this gives 10 pairs. By obtaining points allocated to each pair, the total number of points achieved by an attribute will indicate its overall importance to the respondent.

The next decision is to how to make the comparisons. We could ask respondents:

- to allocate points between each pair, eg 'Please allocate 11 points between the two attributes.' This requires quite a lot of cognitive effort from the respondents;
- to use a bi-polar slider scale to indicate the relative importance of each of the two attributes. This is simple for respondents and can be translated into a points allocation.

We decide to use the bi-polar scale. There are ten pairs which is manageable. The order of showing the pairs is randomized (Figure 6.16).

Figure 6.16 Q23 Comparative importance rating

How important are the following to you when choosing a whisky to buy?
For each pair of statements move the cursor to indicate how much one is more important than the other.

Depth of colour	▼	Smoothness of the taste
Smoothness of taste	▼	Distinct from other brands
How familiar you are with it	▼	Has lots of tradition
Distinct from other brands	▼	How familiar you are with it
Has lots of tradition	▼	Depth of colour
Depth of colour	▼	How familiar you are with it
Distinct from other brands	▼	Depth of colour
Has lots of tradition	▼	Smoothness of taste
Smoothness of taste	▼	How familiar you are with it
Has lots of tradition	▼	Distinct from other brands

Key take aways: creating appropriate rating scales

- Ratings scales allow degrees of sentiment to be expressed and therefore offer greater sensitivity when measuring opinion or attitudes than simple either/or questions.
- The question designer will have to make a number of decisions:
  - Word scales? Numbers? Pictures? A mix?
  - How many scale points are required?
  - Is a mid-point needed?
  - Is a 'don't know' response needed?
  - Can the scale be unbalanced, or should it have equal positive and negative points?
- There are very few clear cut 'rules' when it comes to making these decisions, as the most appropriate choice is likely to depend on many factors including the subject matter, objectives, data collection mode and exactly who we are interviewing.
- The most practical advice is for the question writer to think ahead to how they will interpret the results. Having a point of comparison is often important to put the results into context. Therefore, consistency with scales used elsewhere can often be the driving factor outweighing decisions that would tailor a scale more specifically to a situation.