Notes on Measurement and Scale Design

FL2012.B55.MKT.473.01 - Marketing Research
Washington University in St. Louis

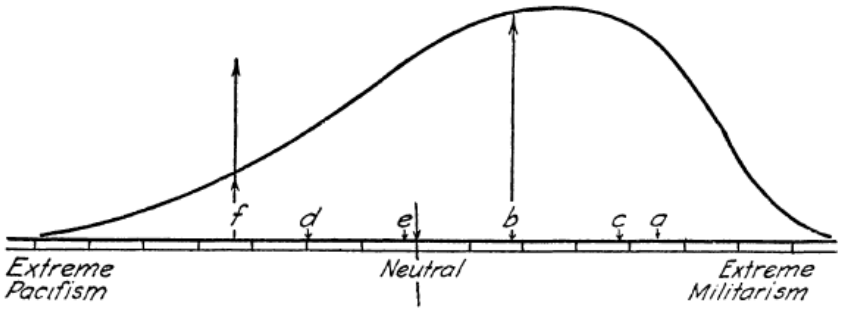
Professor César Zamudio

Contents.

1. Motivation: History and measurement challenges. 1
2. What are we measuring? Constructs. 2
3. What are we not measuring? Error, reliability, validity. 3
4. Sources of error: Biases. 6
5. Designing scales. 9
6. Measurement levels. 10
7. Attitude scales. 13
8. Relation to research proposal. 19
9. Examples of scales. 21
10. References. 26

1. Motivation: History and measurement challenges. It is quite straightforward to obtain anthropometric measures of people, such as their height or their weight. One can readily collect these measures using items such as a measurement tape or a weight scale. But can we obtain *psychometric* measures of people? Is it possible to measure their opinions, attitudes, and subjective sensations? This foundational question was raised by the psychologist Louis Leon Thurstone in a series of articles at the dawn of the 20th. century. In his seminal article, "Attitudes can be Measured" (Thurstone 1928), he established that one can, indeed, produce such a measurement, one which would yield a frequency distribution of attitudes.

Figure 1. Thurstone's (1928) Fig. 1: A hypothetical distribution for the attitude towards militarism-pacifism



How could we design an instrument to collect data that would resemble the hypothetical frequency distribution presented in Fig. 1? Thurstone spelled out these requirements. First, a *unit*

of measurement is required. For example, we use pounds to measure weight, and feet to measure height. In the same way, an underlying continuum of agreement or strength towards an object or idea (in this case, towards militarism-pacifism), is required. Presumably, such a unit of measurement could help us gauge people's attitudes and to assess the distance. Second, a *scale* need be constructed, analogously to a weight scale, measurement tape, or other physical measurement instrument. Third, the scale must be *valid,* i.e., it should indeed measure what we're interested in, in this case militarism-pacifism attitude. The construction of the scale itself was laborious and required that a number of pre-test subjects sort through attitude statements of varying intensity and then compare them (Thurstone 1927). As it turns out, these ideas and procedures, along with other early studies, laid the foundations for what is now known as Classical Test Theory (Traub 1997), after researchers recognized the crucial importance of errors in measurement. The first three sections of this handout deal with important measurement concepts relating to our scales, and to our instruments (a set of scales, for example, our questionnaire). The remaining sections deal with practical issues on measurement and scale construction.

2. What are we measuring? Constructs. Thurstone's work made clear that both physical and psychological characteristics of people could be measured. The motivation for measurement one such characteristic is clear: we assume that people are *different* along the psychological characteristic of interest, and we want to measure these differences as best we can. Furthermore, as Marketing researchers, we may also be interested in then understand how these differences relate to other measurements. For example, a research problem posed by a Marketing researcher could be: "Investigate the target markets' attitude towards the following ethnic groups:...". The researcher, then, may develop a scale on his or her own to measure attitudes, or use an existing scale, such as Bogardus's (1926) social distance scale. With this attitudinal information in hand, the Marketing researcher could now solve a potential second research problem: "Determine whether the target markets' attitude towards different ethnic groups influences store choice". In this example, the target markets' attitude towards different ethnic groups is known as a *construct:* it is a concept that has been deliberately invented by the researcher for a special scientific purpose (Kerlinger 1973). The measurements of constructs such as the above can then be used to attempt to explain observed behavior (store choice, in this case).

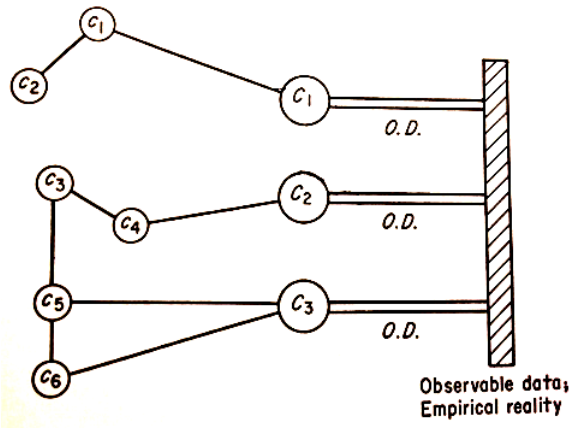
One important characteristic of constructs is that they can be defined in two ways:

- *Constitutive/conceptual definition:* When we define a construct constitutively, we are defining it in terms of other constructs. Example: We can define "Service quality" as "An attitude towards the service that results from comparisons between expectations and perceptions of service performance".

- *Operational definition*: When we define a construct operationally, we spell out how to measure the construct. Example: "Measure perception of service quality using the SERVQUAL Scale (Zeithaml, Parasuraman and Berry 1988)[1]".

The constitutive definition, as can be inferred from the above, helps us define what the construct is; additionally, the operational definition helps us define how to measure the construct. For example, Kerlinger (1973) shows a stylized model of the above:

Figure 2. Kerlinger's (1973) Fig. 3.1: Constructs defined operationally and constitutively



In Fig. 2, the smaller circles are constructs defined constitutively. Note how some of these relate to each other. The larger circles represent the constructs defined operationally - how to connect the constructs to actual empirical data, such as the data you'll collect after you design your questionnaire. Many constructs useful for Marketing can be defined in this way - social influence, popularity, attitudes, purchase intentions, and so forth. For the purposes of our class, however, we won't spell out the constitutive definition of our constructs in great detail; rather, we'll rely on quite well-known operational definitions of, say, attitude, and use these to measure constructs of interest.

3. What are we not measuring? Error, reliability, validity. Note that understanding construct definition is useful to determine what we are measuring. However, there are other constructs and random variables that we are *not* measuring, and that may impact the measurement of our construct of interest. In other words, construct measurement is most likely not precise, and would therefore be subject to errors of different nature. Essentially, an observed response has three components (Churchill 2009):

$$\text{Observed response} = \text{Truth} + \text{Systematic Error} + \text{Random Error}$$

[1] Consult this paper's Appendix to see how the instrument looks like!

The above equation suggests that the recorded responses from the consumer do not only contain information about the construct (Truth), but also error that arises from two reasons. Let us discuss these two reasons separately:

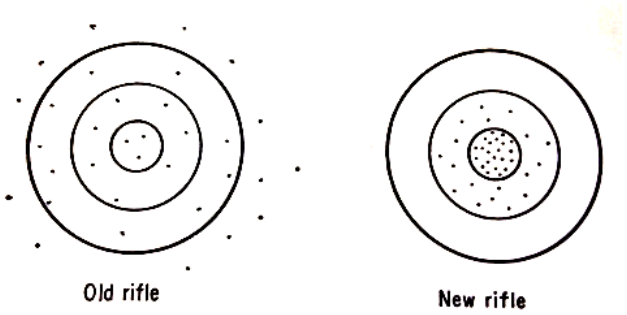
- *Systematic error* represents errors in measurement that are constant. For example, suppose that you ask the question: "How bad is Applebee's service quality?". The results of this type of "leading question" is that consumers will undervalue Applebee's service quality. A further example is if one were interested in assessing WUSTL's students attitudes towards finance jobs, and one were to sample only from the population of Finance majors. It is highly likely that the recorded attitudes would tend to be positive, and thus an overstatement of the attitudes of the whole WUSTL student body.
- *Random error* represents errors in measurement that are not systematic. These can have numerous sources. First, it may be due to changing consumer tastes or whims. For instance, if one interviews a consumer regarding attitudes towards snack items when he/she is hungry (or not), variations in attitude could be observed. Second, it may be because of situational factors. For instance, if consumers are interviewed in a crowded cafeteria, their ability to cognitively process information may diminish, leading to errors in responses. Third, elements of the instrument being administered (i.e. the questionnaire) can be the culprit. For example, the question "Check this box if you do eat snacks at night" is worded in such a way that consumers may mistakenly read "do" instead of "don't".

The above suggests that error can never be entirely removed. One could never hope to interview, for example, consumers with the exact same level of hunger because this is impossible to observe. However, error can be minimized: one could word questions and select questionnaire items which are unambiguous, ensure that the sample is as random as possible, and so forth, which will reduce, but not eliminate, error.

The consequences of measurement error are made manifest in the observed variation in the responses we record. Realize, first, that as researchers we *want* variation, as in Fig.1, where the frequency distribution of militarism-pacifism attitudes implies that not everyone shares the same attitude. A set of responses where most of these attitudes didn't vary would probably reflect that our sampling wasn't properly conducted, that we didn't condition on another variable (e.g. militarism-pacifism among young versus old people, or among people of different political inclinations). What we *do not* want as researchers is variation due to the two types of errors described above.

What are the consequences of a larger amount of systematic or random error? A less accurate measurement. This measurement accuracy is known as reliability: the larger the error, the less reliable our measurement is.

Figure 3. The rifle example and reliability (Kerlinger 1973)



As Fig. 3 depicts, an old rifle has much more "spread" than a new rifle. Consequently, the old rifle is less reliable than the new rifle. As researchers, we must pay careful attention to ensuring an adequate level of reliability, which is a minimum requirement for good measurement.

The preceding discussion assumes that the bulls-eye we're shooting at is the right bulls-eye. But could it be that we're not shooting at the right one? In terms of research, this would imply that we may be measuring a construct other than the one we're interested. If so, an incredibly reliable instrument will still yield poor measures because we're not measuring what we're supposed to. A *valid* instrument is one that measures what it's supposed to be measuring, and several types of validity exist:

- *Content validity:* An instrument or scale has content validity if the various items that compose them are substantially related to the construct they were designed to measure. For instance, suppose we propose the following scale to measure attitudes towards obtaining a college degree:

SCALE 3.1	**IN YOUR OPINION, OBTAINING A COLLEGE DEGREE IS...**	
Useful	_____ _____ _____ _____ _____ _____ _____	Useless
Risky	_____ _____ _____ _____ _____ _____ _____	Not risky
Wise	_____ _____ _____ _____ _____ _____ _____	Foolish
Fine	_____ _____ _____ _____ _____ _____ _____	Coarse
Difficult	_____ _____ _____ _____ _____ _____ _____	Easy
Prestigious	_____ _____ _____ _____ _____ _____ _____	Not prestigious

  In this example, first, notice that the "Risky/Not risky" items and "Wise/Foolish" item will probably be highly related among themselves. Furthermore, one could argue that these are also related to attitudes towards a college degree. On the contrary, the "Fine/Coarse" item is irrelevant and therefore does not contribute to a scale's content validity. To increase content validity, therefore, we must argue and make sure that the items in a scale are related to the others in some way, and that these do measure the construct of interest. Note that we say "argue" because content validation is essentially judgmental (Kerlinger 1973), although some quantitative evidence may be provided.

- *Predictive validity:* An instrument or scale has predictive validity if the resulting scores help us predict some further behavior of the target market. For example, suppose we obtain scores on attitudes toward a new flavor of Doritos. According to the Theory of Reasoned Action (Fishbein & Ajzen 1975), attitudes must contribute to intention to purchase and behavior, and thus predict these in some way. However, if the attitude scores we obtain are unrelated to intention or behavior, then our scale would be said to have low predictive validity.
- *Construct validity:* An instrument or scale has construct validity when one can show that the construct is correctly measured according to other extant theories. Establishing construct validity is quite complicated because it has many requirements which we will not discuss.

Reliability and validity are thus quite important. As Marketing researchers we must ensure that our measures are reliable so that we can consistently use our instruments without gross error, and we must ensure that our measures are valid so that we can be reasonably sure that we're measuring what we're intending to.

The remainder of these notes will deal with practical issues in scale design and measurement.

4. Sources of error. Biases. Previously we discussed some reasons why one may measure with more, or less, error. Situational, personal, and methodological reasons were given. We also reflected on the fact that error cannot be eliminated, but only minimized. A specific typology of these sources, however, would be worthwhile to study. Podsakoff and his colleagues (2003) developed such a typology that concerns *method biases,* where bias simply means a deviation from the true measurement. They are called "method" biases because these are said to depend on the researchers' chosen methodology and controls. So, for instance, if a consumer answers a survey while angry, sad, or hungry, this source of error (or bias) cannot be controlled by minimizing method bias. However, appropriate wording of an item would fall within the domain of method bias minimization. Podsakoff et al's (2003) Table 2 illustrates several of these sources of method bias. We will discuss some of these shortly along with some specific remedies. More general remedies will be discussed at the end of this section.

4.1 Common rater effects

Common rater effects are those that relate to the subjects that are responding our questionnaire or other instrument.

- *Social desirability.* A consumer answers in a socially desirable way, not in a truthful manner. Controversial topics in questionnaires, such as birth control and race relations are prone to social desirability bias. However, less controversial subjects may also be prone to this bias. For example, in a survey of employees regarding the friendliness of their workplace, an employee may feel that it is socially desirable to report a higher level

of perceived friendliness, even if he or she feels this is not the case. Survey anonymity can help remedy social desirability bias, or gathering a person's intrinsic social desirability score using a scale such as the Crowne and Marlowe Social Desirability scale (1960). We can then report our measures conditional on someone being more prone to answer in socially desirable ways.

- *Acquiesence bias.* A consumer has a tendency to agree with the questions asked *regardless of the content of the questions*. This bias is likely to be accentuated if the interviewer is close by or looking at the responder when he or she is answering a questionnaire. Furthermore, this bias can be also accentuated if the verbal labels of the questionnaire's scales are arranged in such a way that all the positive answers are in one side versus the other. For example, compare Scale 2 to Scale 1 (ignoring the Fine/Coarse item):

SCALE 4.1	**IN YOUR OPINION, OBTAINING A COLLEGE DEGREE IS...**	
Useful	_____ _____ _____ _____ _____ _____ _____	Useless
Not risky	_____ _____ _____ _____ _____ _____ _____	Risky
Wise	_____ _____ _____ _____ _____ _____ _____	Foolish
Easy	_____ _____ _____ _____ _____ _____ _____	Difficult
Prestigious	_____ _____ _____ _____ _____ _____ _____	Not prestigious

In Scale 2, all the positive elements are on the left side of this semantic differential scale. Consequently a consumer may feel enticed to simply answer these positively or negatively. Finally, time pressure may also entice the consumer to answer in an acquiescent way. Remedies include letting the subject answer the questionnaire privately (or under disguised observation), and shifting labels ("counterbalancing") so that the positive and negative verbal elements are not consistently located down a scale.

4.2 Item characteristics

Item characteristic effects are those that relate to the content of the items that the subjects are presented with.

- *Item complexity/ambiguity.* One or some of the items of a scale are hard to understand or their meaning is ambiguous. For example, if we ask the question "How much have you spent in each of your past 8 visits to Dierberg's?", this question is clearly highly difficult, and indeed impossible to answer, and thus the consumer is likely to respond with a random number or a quite inaccurate estimate. Also, the wording of a question can be ambiguous, such as in the following double-barreled question: "How aware and excited are you regarding the following Specialized Masters Programs?". The consumer may not know whether to respond something about his or her excitement, or about his or her awareness. Appropriate wording and pre-testing of items can help us mitigate this type of bias.

- *Negatively worded items:* When an item is negatively worded, e.g. "Do you not support the reelection of public officials?", consumers may fail to recognize this and thus answer incorrectly. This problem becomes more acute in the later stages of the questionnaire one the consumer has established a response pattern (Podsakoff et al. 2003). Avoid negatively worded items in general.

4.3 Item context

Item context effects are those that relate to the context that other items generate around which the item of interest is presented.

- *Context-induced mood.* When an item induces a certain consumer mood which then impacts the answer of other items. For example, a consumer may be somewhat dissatisfied with the state of the economy. If, before asking to give an overall evaluation of the current administration's economic efforts, the survey asks about attitudes and feelings towards economic depressions in general, the consumer may be temporarily put in a mood in which his feeling about the administration may worsen. This would be even more accentuated if the questions are loaded, i.e., convey negative meaning. One can minimize this bias by examining a questionnaire's flow and trying to determine if this type of mood can occur. Note that, first, mood may be induced not by items but by the context in which the consumer is interviewed (e.g. in the middle of a busy, noisy room vs. in a quiet room) as we will discuss next; second, some very guileful pollsters sometimes may use context-induced mood on purpose.

4.4 Measurement context

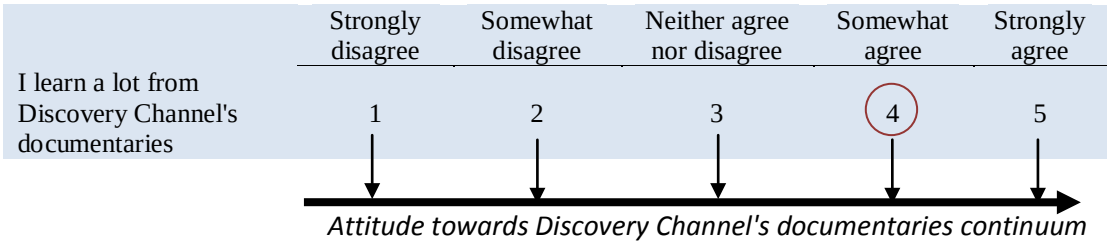
Measurement context effects are those that relate to the context around which the items are presented.

- *Time and location.* As we just discussed, the location where a survey is being administered matters, as well as the time of day and the time allotted for answering the survey. For example, suppose you are administering a personal interview questionnaire at the airport. Consumers may be in a very hurried mode, and therefore answer questions with haste, not paying particular attention, thus increasing error; in addition, consumers may be tired (if they are changing planes, for instance) or in a bad mood, which may also increase error. This source of bias can be mitigated by controlling the time and location at best, or at least paying attention to the environment in which the consumer is currently in.

There are statistical and procedural techniques to deal with bias minimization (Podsakoff et al. 2003). Statistical techniques are out of the scope of this class. Procedural techniques, in which one pays closer attention to questionnaire wording, ambiance, etc. have been already discussed to some extent for each source of bias.

5. Designing scales. We have discussed foundational concepts in scale design, such as a historical outlook on the motivation for scale development, constructs, error. Sources of error that arise in practice have also been addressed. But how do we develop a scale? A rigorous answer to this question is outside of the scope of our class, and Churchill (2009) provides a summary. We will only discuss the underlying dimensions that underlie a scale.
A scale is a measuring instrument consisting of symbols or numerals that allows a respondent to assign a value that reflects an individuals' possession of what the scale measures (Kerlinger 1973). Consider the following figure depicting a Likert item answered by two consumers:

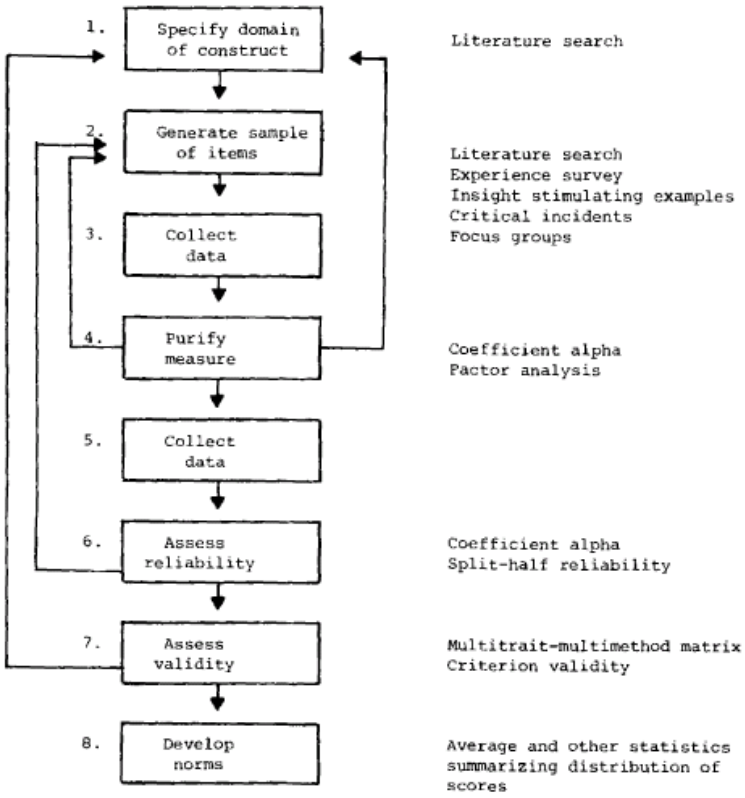
Figure 4. The underlying continuum on a Likert scale.



In Fig. 4, a Likert item is shown. Notice that underlying this item is a continuum, in this case, that of attitude towards an object. Presumably, learning from Discovery's documentaries should be related to a positive attitude towards these shows, and responses to the item shown will help us discover consumers' attitudes. Consequently, the item allows us to uncover unobserved consumers' attitudes, which they possess, just as a measuring tape or weight scale allows us to determine weight or height, which consumers possess also. The various gradations or *scale points*, thus, are assumed by the researcher to correspond to the underlying attitude continuum. Here the consumer answered "Somewhat agree" (4), which corresponds to the upper part of the continuum. As we will see shortly, however, this assumption need not hold entirely when we compare multiple consumers.

The detailed procedure to design a scale is outside of the scope of this class, but typically follows the steps shown in Figure 5 (Churchill 1979). A simplified version of this procedure is shown in the Churchill (2009) textbook. Students interested in knowing more about scale development are encouraged to start with Churchill's (1979) paper.

Figure 5. Churchill's (1979) full procedure for scale development



There are various types of scales that we discuss in Marketing research. The broadest categorization, *measurement levels*, refers to scales that allow different levels of gradation in the mapping between the item and the underlying continuum. For example, Fig. 4 shows an interval scale, where we assume the intervals between 1-2, 3-4, etc. to be equal. Other types of scales do not allow for such an assumption. A finer category is that of *attitude scales* that have been specifically designed to measure attitudes towards an object. We can also divide scales by the number of items they require: some scales require one item, whereas others are known as *multiple item scales.* We will be specific about these details as we discuss different scale types.

6. Measurement levels. There are four types of levels that compose any scales. These levels originated in the late 1940s, after mathematicians, physicists and psychologists discussed at length whether subjective sensations could be measured, and to what extent (Stevens 1946). Each level essentially assigns a number to objects, but the numbers have different meaning for each level:

- An *nominal level* scale is used to categorize objects. Therefore, one could assign the following numbers to the following brands: 1=Nike, 2=Adidas, 3=Reebok, 4=Converse. The numbers in this case have no relation to any continuum. They are simply labels. Frequency distributions and percentages can be used to analyze nominal data.

- An *ordinal level* scale is used to rank objects. Consider the following question:

Please rank these brands according to how much you prefer them, using numbers from 1 to 4, where 1=Most preferred and 4=Least preferred.

Nike	1
Adidas	2
Reebok	3
Converse	4

This response lets us know that Nike>Adidas>Reebok>Converse. Although an ordinal scale can be related to a continuum of measurement (in this case brand preference), the distance between each of these objects, therefore, cannot be assessed, as Fig. 5 shows:

Figure 5. Preference location cannot be identified with an ordinal scale

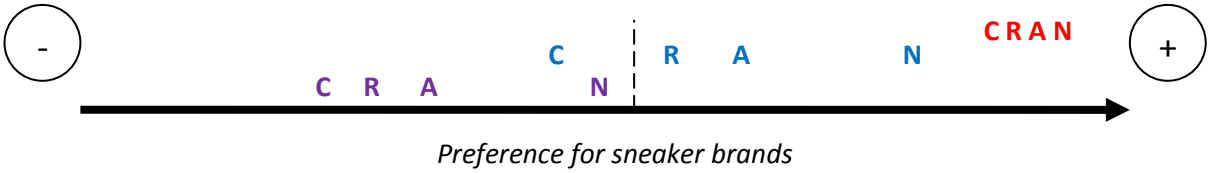


Fig. 5 depicts answers by three consumers, Red, Blue and Purple. They all ranked N>A>R>C. However, Red seems to not care about differentiation: he simply loves sneakers. Blue seems to care about differentiation and, also, he might prefer athletic to casual sneakers. Finally, Purple probably doesn't like sneakers very much and, for him, Nike is the least unliked, perhaps because of their corporate branding. These preference structures cannot be recovered using an ordinal scale. The mode, median and percentiles can be calculated to assess results from an ordinal scale.

- An *interval level scale* solves the problem above by addressing the distance between the scale points. The assumption is that the intervals between the multiple points are equally spaced in the consumers' mind. Note that as long as we have *an* interval, we are dealing with an interval scale. Consider Fig. 6, which is a modified version of Fig. 4:

Figure 6. Multiple answers and intervals on Discovery's Likert item

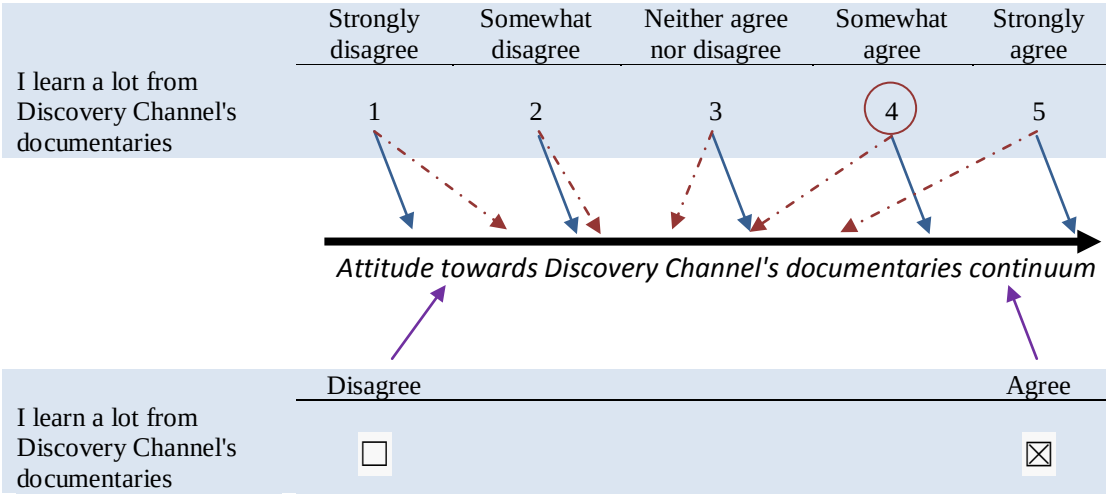


Fig. 6 shows the same Likert item as Fig. 4, but now presents two consumers answering that question, Blue and Red. They both answered "Somewhat agree". Although the distances between the answers are identical, Red's answers are less disperse and concentrate on the middle of the continuum, whereas Blue's are more disperse. As researchers, however, we would assume that they both exhibit the same level of attitude.
A third consumer, Purple, answered the same question but using an Agreement or Disagreement item, not a Likert item. This item is interval-scaled, but there is only one interval to speak of. This Agreement/Disagreement scale, however, can still be related to the attitude continuum we're interested in.
A disadvantage of the ordinal scale, as is probably evident by now, is that because intervals are assumed to be equal but are still subjective, the magnitude differences between responses cannot be assessed. So, for example, Blue's "4" response is not equivalent to Red's "2" multiplied by 2. We can employ the mean, standard deviations and correlations to understand interval data.

- An *interval level scale* solves the problem above. The interval scale has a true, absolute zero that is empirically meaningful (Kerlinger 1973). In other words, an answer on a ratio scale is a number in the fullest sense of the word, be it discrete or continuous. Consider the following constant-sum scale:

Please distribute 100 points in a way that best reflects your preference for the following brands.			
Nike	90	34	5
Adidas	5	46	10
Reebok	3	10	10
Converse	2	10	75

Figure 7. Comparing consumers' preference profiles (Radar chart)

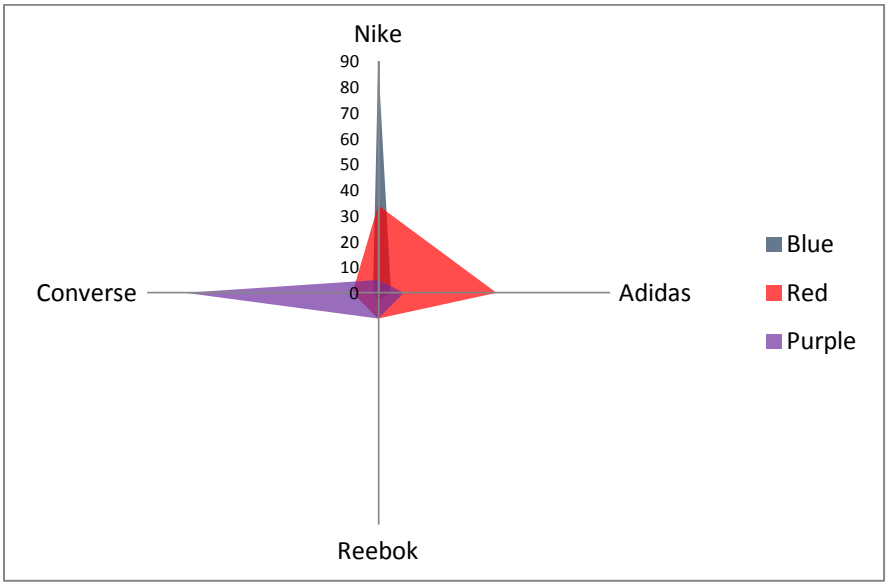
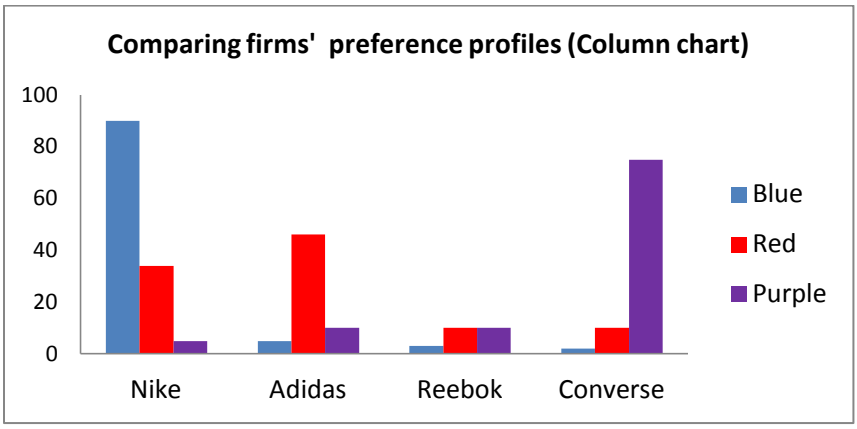


Figure 8. Comparing firms' preference profiles (Radar chart)



Here the radar chart focuses on comparing three consumers, Blue, Red and Purple (these are not the same consumers as in our previous example). As can be seen, Red's distribution is more disperse, whereas Blue and Red's preferences are quite skewed towards one brand. Red may in this case be a switcher and Blue/Purple may be brand loyals. However, it is also possible that the researcher incorrectly added Red, someone who doesn't like sneakers particularly, to the sample. Most statistical analyses can be performed using ratio data.

Different measurement scales convey different levels of information. In particular, Ratio contains more information than Interval, Interval contains more than Ordinal, and Ordinal contains more than nominal. By "information" we specifically mean information about the distance between the scale points. This suggests that, perhaps, one should always use a Ratio scale. This is not true. The Ratio scale requires that consumers are very specific in their assessments, which can have misleading responses as a consequence. For example, if we asked consumers to assess the probability that they consume a certain product in the future, it may be very difficult for them to suggest a specific probability from 0 to 100. Consumers, then, will be more likely to report numbers such as 100, 75, 50, 25 and 0. An interval scale may then be better suited to capture this information.

7. Attitude scales. Attitude scales are scales that are specifically designed to measure attitudes. Most of these attitude scales belong to a class of scales known as *summated ratings scale*, which in turn are multiple-item scales. These scales bear such a name because the researcher specified a set of items (not just one) that measures attitude; in order to obtain an attitude score for each consumer, the researcher must sum the reported ratings on each one of these items. The underpinnings of some of the most popular scales are discussed next.

7.1 The Likert scale

The Likert scale was developed by Rensis Likert (1932). It is a scale that consists of several items, in horizontal arrangement, anchored with integer numbers and verbal labels. The verbal labels must be ambivalent and measure attitude on an agreement continuum. These characteristics are spelled by Uebersax (2006) - we use his example of a Likert scale:

Figure 8. Uebersax's example Likert scale for measuring attitudes toward American politics

	Strongly Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
The President is doing a good job.	1	2	3	(4)	5
The Congress is doing a good job.	1	2	3	4	(5)
The Secretary of Defense is doing a good job.	1	2	3	(4)	5

Likert ITEM — *Likert SCALE*

Fig. 8 depicts a potential scale to measure the construct "attitudes toward American politics". Here this consumer's responses appear to reflect a positive attitude. Note that a Likert *scale* is a set of Likert *items.* Consequently, to measure attitudes towards an object using a Likert scale, we need to include several Likert items that all measure different components of the attitudinal construct we wish to measure. Using Factor Analysis of these responses we may later discover that there are underlying clusters of items or "dimensions" present in consumers' attitudes.

7.2 The Semantic Differential scale

The Semantic Differential (SD) scale was developed by Charles E. Osgood and his colleagues (1964). The SD scale also concerns the measurement of attitudes, but instead of agreement, the SD scale attempts to characterize attitude in terms of what the object *means* to consumers, or how consumers *judge* a concept - Osgood and colleagues aptly titled their seminal book "The Measurement of Meaning". Then, the SD scale is very useful to uncover attitude profiles that vary from object to object.

The SD consists of the following elements: A set of *subjects* (our respondents) will respond several scale items about many different concepts. The concepts that people can rate are varied: consumers can rate a set of brands (Nike, Adidas, Reebok, Converse...), graduate programs (3+2 MBA, SMP, Executive-Ed...). For Marketing researchers, what is critical is that *the concepts that people will rate include the alternatives they consider when choosing.* For example, suppose our

client is the brand manager of Sun Chips. In order to use a Semantic Differential scale for Sun Chips, we would need to include other concepts such as Doritos, Cheetos, and Lay's; but if, in our focus groups and interviews, we discover that beef jerky and even fruit substitute for Sun Chips, then we may need to include these too. Naturally, the more concepts that the consumer rates, the more demanding the task.

Concepts are rated by subjects in a set of scale items. The scale items in a SD scale are very different than those in the Likert scale, however. Consider the following example:

Please rate the following snack products by checking the appropriate space.

SUN CHIPS (PLAIN)

Expensive	_____;_____;_____;_____;_____;_____;_____	Cheap
Crunchy	_____;_____;_____;_____;_____;_____;_____	Mushy
Tasteful	_____;_____;_____;_____;_____;_____;_____	Tasteless
Environment-friendly	_____;_____;_____;_____;_____;_____;_____	Not environment-friendly
Fattening	_____;_____;_____;_____;_____;_____;_____	Not fattening

LAY'S POTATO CHIPS (PLAIN)

Expensive	_____;_____;_____;_____;_____;_____;_____	Cheap
Crunchy	_____;_____;_____;_____;_____;_____;_____	Mushy
Tasteful	_____;_____;_____;_____;_____;_____;_____	Tasteless
Environment-friendly	_____;_____;_____;_____;_____;_____;_____	Not environment-friendly
Fattening	_____;_____;_____;_____;_____;_____;_____	Not fattening

DORITOS (NACHO CHEESE)

Expensive	_____;_____;_____;_____;_____;_____;_____	Cheap
Crunchy	_____;_____;_____;_____;_____;_____;_____	Mushy
Tasteful	_____;_____;_____;_____;_____;_____;_____	Tasteless
Environment-friendly	_____;_____;_____;_____;_____;_____;_____	Not environment-friendly
Fattening	_____;_____;_____;_____;_____;_____;_____	Not fattening

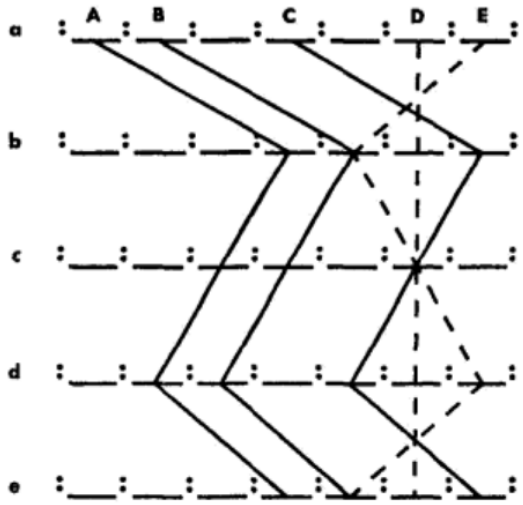
JACK LINK'S BEEF JERKY (ORIGINAL)

Expensive	_____;_____;_____;_____;_____;_____;_____	Cheap
Crunchy	_____;_____;_____;_____;_____;_____;_____	Mushy
Tasteful	_____;_____;_____;_____;_____;_____;_____	Tasteless
Environment-friendly	_____;_____;_____;_____;_____;_____;_____	Not environment-friendly
Fattening	_____;_____;_____;_____;_____;_____;_____	Not fattening

As can be seen, the SD scale consists of numerous concepts (brands, in this case: Sun Chips, Lay's, Doritos, and Jack Link's jerky) that consumers then rate on various scale items which

consist of 7 positions. The scale items are the same for each concept; consumers have to rate all of the concepts on all of the scales. Consequently, one must keep the number of concepts and scale items to a manageable number.
When a SD scale is used, the researcher assumes that attitude is a multidimensional construct. assumes that attitude is a multidimensional construct. Concepts can be profiled by plotting these dimensions one at a time using a snake diagram, such as the one below due to Osgood et al. (1964):

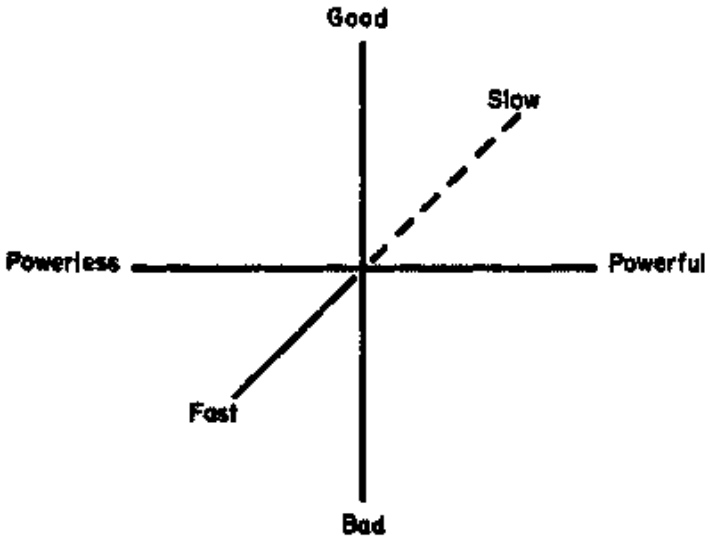
Figure 10. Osgood's hypothetical snake diagram



Here A, B, C, D, E would represent our brands, and a, b, c, d, e would represent our scale items. To make these diagrams, we simply calculate a measure of central tendency (mean or median) for each concept, for each scale, and then plot it. Importantly, one may find that several concepts share the same attitude profile. For example, Sun Chips and Lay's may a reasonably similar attitude profile as compared to Doritos or Jack Link's.

Because the attitudes that the SD scale measures are assumed to be multidimensional, one may attempt to plot them in such a multidimensional space if one knew it. Factor Analysis is a technique that is used in this regard: we try to find several underlying "dimensions" that the attitudes exhibit. An important empirical finding in many studies that deal with the SD scale is that the scale items often follow the so-called EPA structure (Heise 1970) - "Evaluation-Potency-Activity". This means that, in most cases, the attitude towards a concept will exhibit the following structure:

Figure 11. EPA structure in SD scaling (Heise 1970)



Here the Y-axis represents the "Evaluation" dimension (is the concept good or bad?); the X-axis represent the "Potency" dimensions (is the concept powerless or powerful?); the Z-axis represents the "Activity" dimension (is the concept slow or fast?). In practical terms, what this means for researchers is that we can select and use predefined adjectives for our scales, being confident in that these will be related to our concept in meaningful ways. Table 3 lists some examples of adjectives in each of these dimensions:

Table 3. Adjective pairs associated with the EPA structure (Osgood et al. 1964)

Evaluation	Evaluation (contd.)	Potency	Activity
Good-bad	Progressive-regressive	Hard-soft	Active-passive
Optimistic-pessimistic	True-false	Strong-weak	Excitable-calm
Complete-incomplete	Positive-negative	Severe-lenient	Hot-cold
Timely-untimely	Reputable-disreputable	Tenacious-yielding	Intentional-unintentional
Altruistic-Egotistic	Believing-skeptical	Constrained-free	Fast-slow
Sociable-unsociable	Wise-foolish	Constricted-spacious	Complex-simple
Kind-cruel	Healthy-sick	Heavy-light	
Grateful-ungrateful		Serious-humorous	
Harmonious-dissonant		Opaque-transparent	
Clean-dirty		Large-small	
Light-dark		Masculine-femenine	
Graceful-Awkward			
Pleasurable-painful			
Beautiful-ugly			
Successful-unsuccessful			
High-low			
Meaningful-meaningless			
Important-unimportant			

It can be readily seen that the "Evaluation" adjectives compose most of the list in Table 3. The reason is that this dimension is typically found to be the largest or most important, in terms of the variation in attitude that it helps explain. Researchers are then encouraged to include scales that reflect the EPA structure in their SD scales - typically at least 3 per dimension (Osgood 1964). If, in addition to this, the researcher believes there are other underlying dimensions (for example, if dealing with car brands, prestige may be important and not entirely related to evaluation), these can be included also.

7.3 The Stapel scale

The Stapel scale was developed by Jan Stapel and presented in the Marketing literature by Irving Crespi (1961). The Stapel scale is somewhat similar to the SD scale, in that subjects also rate concepts along scales, which typically are adjectives. The main differences are four: (1) the adjectives are no longer bipolar, and thus evaluated one by one; (2) unlike the SD scale, the Stapel scale has number anchoring, as in the Likert scale; (3) there are 10 scale positions, and (4) there is no neutral option. An example of a Stapel scale is displayed below (only one concept is rated):

Please rate the following snack products by checking the appropriate space.

SUN CHIPS (PLAIN)

	-5	**-4**	**-3**	**-2**	**-1**	**+1**	**+2**	**+3**	**+4**	**+5**
Expensive	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Crunchy	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Tasteful	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Environment-friendly	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Fattening	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

The Stapel scale may be easier to construct when the bipolarity of the adjectives the consumer is rating is difficult to produce. Both the Stapel and SD scales yield almost identical results (Hawkins et al. 1974). However, the lack of neutrality in the scale may be a concern for some researchers.

7.4 Other itemized rating scales for measuring attitudes

There are many other scales that researchers can use to measure attitudes. Thurstone's Equally Appearing Intervals (Thurstone 1927) and the Guttman cumulative rating scale (Kerlinger 1973) can be used (the Bogardus (1926) social scale cited earlier is an example of a Guttman cumulative scale). We will not discuss these two more specialized scales. However, it can be pointed out that researchers can develop their own scales according to their needs. For example,

instead of agreement or adjectives, researchers can use an importance continuum, feelings, satisfaction with quality, satisfaction with needs, fit with the subject, positivity/negativity, and so forth. Furthermore, scales can be designed with verbal and numerical anchors for the scale points, (as we discussed earlier) but also graphical anchors. For example:

Figure 12. Wong-Baker FACES Pain Rating Scale

The above rating scale is known as the Wong-Baker FACES scale, and is used to understand how children assess physical pain. Naturally, one can use this same example to develop scales that measure satisfaction, preferred beverage sizes, how one perceives some concept, himself, others, and so forth.

8. Connection with research proposal. The above knowledge is not disconnected with the research proposal the Marketer puts forth. Quite the opposite: the research problems put forth by the researcher will guide our selection of scales that will be included in the research proposal, as in the following two examples.

Decision problem	**Research problem**	**Exploratory research findings**	**Consequence for scale choice in descriptive research**
Are we perceived as a trustworthy company?	Determine consumers' level of trust in Company X	Consumers' trust seems to be multidimensional: consumers want to feel the company is dependable, reliable, and honest.	Use a semantic differential scale containing: * 2 items only from the EPA adjectives (due to space limitations) * Dependable-not dependable item * Reliable-unreliable item * Honest-not honest item * Consider including 3 more items for the three dimensions stated by consumers.
Do the number of salesperson awards impact likelihood of purchase?	Determine the impact of visible salesperson awards on purchase probability	Consumers have a strong opinion on awards, and it seems to affect not only purchase probability, but also likelihood of recommendation and willingness to pay. It is suggested to add these to the list of research problems.	*Purchase probability:* Use a single-item rating scale for likelihood of purchase. *Willingness to pay:* Use a ratio-scaled 0-100 question. *Likelihood of recommendation:* Use a single-item rating scale measuring likelihood of recomendation. *Control for attitude towards awards:* Use a Likert scale to measure attitude towards awards. Minimum 5 items.

The above examples show that our decision problems, research problems, exploratory and descriptive research have to be linked to one another. Of course, also, the above examples assume that we consistently interview subjects from the sample specified in the sampling plan. Delivering a consistent, well-planned project goes a long way towards establishing good habits as a Marketing researcher.

9. Examples of scales. In the next pages we present examples of the application of different types of scales, along with examples of proper instruction writing. Use these examples as reference for your own work. Also, we will use examples drawn from this compendium in class.

NOMINAL SCALE

Where did you purchase groceries during your last grocery trip? (You can mark more than one option)

☐ Schnucks ☐ Whole Foods ☐ Dierbergs
☐ Target ☐ Albertsons ☐ Other

Please specify your race:

_____ Asian
_____ Black or African American
_____ Hispanic or Latino
_____ White
_____ Mixed/Other

How did you hear from us? (Check all that apply)

_____ TV advertisement
_____ Radio advertisement
_____ From a friend
_____ WUSTL faculty -----> Faculty name: _____
_____ Website -----> Which website? _____

Which of the following shows do you watch the most? (Select *at most* three)

CNN

☐ The Situation Room with Wolf Blitzer ☐ Erin Burnett Out Front ☐ Anderson Cooper 360º

FOX NEWS

☐ Special Report ☐ FOX Report ☐ The O'Reilly Factor ☐ Hannity

MSNBC

☐ Hardball with Chris Matthews ☐ The ED Show ☐ The Rachel Maddow Show
☐ The Last Word with Lawrence O'Donnell

(Top of mind/Share of mind) Please list the first three detergent brands that you remember:

1. _____
2. _____
3. _____

Although a ToM/SoM question is open ended, we later code consumers' responses and categorize them.

ORDINAL SCALE

Please specify your age:

_____ Under 18
_____ 19-25 years old
_____ 26-40 years old
_____ 40-65 years old
_____ Over 65 years old

Please rank the following headphone brands in order of preference, where 1 is most preferred and 5 is least preferred.

_____ Beats by Dr. Dre
_____ Bose
_____ Koss
_____ Monster Cable
_____ Sennheiser

Please rank the following headphone brands in order of preference, where 1 is most preferred and 5 is least preferred.

_____ Beats by Dr. Dre
_____ Bose
_____ Koss
_____ Monster Cable
_____ Sennheiser

Note that attitudes can also be addressed using the ordinal scale, as in the following example, although the interval difference among the responses of course cannot be assesed.

Please rank the following characteristics of messenger bags according to your preference, where 1 is the most preferred characteristic and 5 is the least preferred characteristic.

_____ That the messenger bag looks well-made
_____ That the messenger bag design is colorful
_____ That the messenger bag is spacious
_____ That the messenger bag has a dedicated laptop compartment
_____ That the messenger bag has a dedicated tablet compartment

---

It is also noteworthy to mention that (1) there are quite large scales which use ordinal scaling, such as the Rokeach Value Survey (Johnston 1995); also, (2) some methods require consumers to order objects instead of writing a ranking, such as the procedures involved in Q-Methodology studies (see Kerlinger 1973, Ch. 34).

INTERVAL SCALE (SINGLE-ITEM RATING TASKS)

Recommendation. Would you recommend this insurance agent to a friend?

I'm sure I would NOT recommend ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ I'm sure I WOULD recommend

Likelihood of adoption. How likely are you to adopt the new Gruntmaster 6000?

Extremely likely _____ _____ _____ _____ _____ _____ _____ Extremely unlikely

Favorability. Taking everything into consideration, how favorable are you towards pursuing graduate education in general? Circle the option that best reflects this favorability.

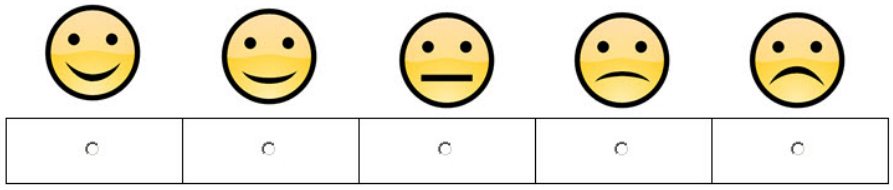
Very unfavorable	Somewhat unfavorable	Neither unfavorable nor favorable	Somewhat favorable	Very favorable
1	2	3	4	5

Note that this particular scale has numerical anchors.

Graphical rating scale. Please evaluate how satisfied you are with the taste of Doritos Nacho Cheese by marking a 'X' at the position in the line that best reflects your satisfaction

Very satisfied _____ Very unsatisfied

Graphical rating scale (2). Please evaluate how satisfied you are with the taste of Doritos Nacho Cheese by marking a 'X' at the position in the line that best reflects your satisfaction



This scale was obtained from SIMS Sensory Software, all rights reserved. Accessed 9/23/2012.

Multiple concepts, one item using self-report ratings. Please evaluate the service quality of the following hotels on a scale from 0 to 10, where 0 is the worst evaluation and 10 is the best evaluation.

Doubletree _____ Holiday Inn _____ Holiday Inn Express _____ Studio 6 _____ Motel 6 _____
Sheraton _____ Courtyard by Marriott _____ The Westin _____

INTERVAL SCALE (MULTIPLE-ITEM RATING TASKS)

Likert scale - Attitude towards Casio watches (one brand only). Please evaluate the following statements regarding Casio watches by circling the number that best reflects your agreement with the statements.

I think that Casio watches...	Strongly Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
Are modern.	1	2	3	4	5
Are for trendy people.	1	2	3	4	5
Are colorful.	1	2	3	4	5

Semantic differential scale (1) Attitudes towards TV brands (one brand only). Please describe your feelings towards **Sony HDTVs** by marking the spaces that best reflect these feelings. *Place the marks in the middle of the spaces, not on the boundaries.*

I feel Sony HDTVs are...

Expensive	_____;_____;_____;_____;_____;_____;_____	Cheap
Clear	_____;_____;_____;_____;_____;_____;_____	Not clear
Modern	_____;_____;_____;_____;_____;_____;_____	Old-fashioned
Energy saving	_____;_____;_____;_____;_____;_____;_____	Not energy saving
Beautiful	_____;_____;_____;_____;_____;_____;_____	Ugly

Note that (1) a SD scale would include more than one brand and that (2) the last item comes from Osgood et al's (1964) inventory of "Evaluation" adjectives. Remember to use this inventory!!!

Stapel scale. Attitudes towards TV brands (one brand only). Please describe your feelings towards **HDTV brands** by circling the numbers that best reflect these feelings.

I feel Sony HDTVs are...

Attractive	**Modern**	**Clear**	**Energy saving**	**Easy to install**	**Safe to handle**
+5	+5	+5	+5	+5	+5
+4	+4	+4	+4	+4	+4
+3	+3	+3	+3	+3	+3
+2	+2	+2	+2	+2	+2
+1	+1	+1	+1	+1	+1
-1	-1	-1	-1	-1	-1
-2	-2	-2	-2	-2	-2
-3	-3	-3	-3	-3	-3
-4	-4	-4	-4	-4	-4
-5	-5	-5	-5	-5	-5

RATIO SCALE

State your age in years: _____ years

Thinking about whether an insurance agent gives you a good, fair, or bad quote, what would be the chance that you buy car insurance from the agent*?* **For each type of offer,** express this chance as a number between 0 and 100, where 0 means "absolutely not" and 100 means "absolutely yes".

_____ /100	_____ /100	_____ /100
Good offer was given	**Fair offer was given**	**Bad offer was given**

Note that this scale features a visual aid to ensure subjects do not distribute 100 points among the three items, but rather give a number from 0 to 100 for each one of the items. In this particular case, the original version of the scale did not include such visual aids. Including them increased consumer understanding.

Approximately how much have you spent in your last shopping trip to the following grocery stores? If you have not visited one of these stores in the past two weeks, write "0".

Dierbergs $_____
Schnucks $_____
Target $_____
Whole Foods $_____
Albertsons $_____

Notice that this task can be quite difficult to recall. Consumers may answer with very inaccurate/rounded estimates.

Constant-sum scale. Distribute 100 points among the following five varieties of Coke soft drinks, where a larger number of points reflects a stronger preference for the variety.

_____ Classic
_____ Caffeine free
_____ Vanilla
_____ Lime
_____ Cherry

What is the farthest distance you'd be willing to travel to visit a Bennigan's restaurant?

I would be willing to travel **up to** _____ miles.

References

Bogardus, Emory S. 1926. Social Distance in the City. *Proceedings and Publications of the American Sociological Society* **20**, 40-46.

Churchill, Gilbert A. 1979. A Paradigm for Developing Better Measures of Marketing Constructs. *Journal of Marketing Research* **16**(1), 64-73.

Churchill, Gilbert A., Tom J. Brown. 2009. Basic Marketing Research. 7th. edition.

Crespi, Irving. 1961. Use of a Scaling Technique in Surveys. *Journal of Marketing* **25**(5), 69-72.

Crowne, Douglas P., David Marlowe. 1960. A New Scale of Social Desirability Independent of Psychopathology. *Journal of Consulting Psychology* **24**(4), 349-354.

Fishbein, Martin, Icek Ajzen. 1975. *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research.* 1st. ed.

Hawkins, Del I., Gerald Albaum and Roger Best. 1974. Stapel Scale or Semantic Differential in Marketing Research? *Journal of Marketing Research* **11**(3), 318-322.

Johnston, Charles S. 1995. The Rokeach Value Survey: Underlying Structure and Multidimensional Scaling. *Journal of Psychology* **129**(5), 583-597.

Heise, David R. 1970. The Semantic Differential and Attitude Research. In *Attitude Measurement*, Summers, Gene F., ed., 235-253.

Kerlinger, Fred N. 1973. *Foundations of Behavioral Research.* 2nd. ed.

Likert, Rensis. 1932. A Technique for the Measurement of Attitudes. *Archives of Psychology* **140**, 1-55.

Osgood, Charles E., George J. Suci, Percy H. Tannenbaum. 1964. *The Measurement of Meaning.* 4th. ed.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee, Nathan P. Podsakoff. 2003. Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology* **88**(5), 879-903.

Stevens, S.S. 1946. On the Theory of Scales of Measurement. *Science* **103**(2684), 677-680.

Thurstone, Louis L. 1927. A Law of Comparative Judgment. *Psychological Review* **34**, 273-286.

----------. 1928. Attitudes can be Measured. *American Journal of Sociology* **33**(4), 529-554.

Traub, Ross E. 1997. Classical Test Theory in Historical Perspective. *Educational Measurement: Issues and Practice* **16**(4), 8-14.

Uebersax, John S. 2006. Likert Scales: Dispelling the Confusion. Statistical Methods for Rater Agreement (website). Available at http://john-uebersax.com/stat/likert.htm. Accessed: 9/22/2012.

Zeithaml, Valarie, A. Parasuraman and Leonard L. Berry. SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. *Journal of Retailing* **64**(1), 12-40.