Data

3

Keywords

Armstrong and Overton procedure • Case • Census • Constant • Construct • Construct validity • Content validity • Criterion validity • Dependence of observations • Discriminant validity • Equidistance • Face validity • Formative constructs • Index • Index construction • Internal consistency reliability • Inter-rater reliability • Items • Latent concept • Latent variable • Measurement scaling • Multi-item constructs • Net Promoter Score (NPS) • Nomological validity • Non-probability sampling • Observation • Operationalization • Population • Predictive validity • Primary data • Probability sampling • Qualitative data • Quantitative data • Reflective constructs • Reliability • Sample size • Sampling • Sampling error • Scale development • Secondary data • Single-item constructs • Test-retest reliability • Unit of analysis • Validity • Variable

Learning Objectives

After reading this chapter, you should understand:

- How to explain what kind of data you use.
- The differences between primary and secondary data.
- The differences between quantitative and qualitative data.
- What the unit of analysis is.
- When observations are independent and when dependent.
- The difference between dependent and independent variables.
- Different measurement scales and equidistance.
- Validity and reliability from a conceptual viewpoint.
- How to set up different sampling designs.
- How to determine acceptable sample sizes.

E. Mooi et al., *Market Research*, Springer Texts in Business and Economics,
DOI 10.1007/978-981-10-5218-7_3

3.1 Introduction

Data are at the heart of market research. By data we mean a collection of facts that can be used as a basis for analysis, reasoning, or discussions. Think, for example, of people's answers to surveys, existing company records, or observations of shoppers' behaviors. "Good" data are vital, because they form the basis of useful market research. In this chapter, we discuss different types of data. This discussion will help you explain the data you use and why you do so. Subsequently, we introduce strategies for collecting data in Chap. 4.

3.2 Types of Data

Before we start discussing data, it is a good idea to introduce the terminology we will use. In the next sections, we will discuss the following four concepts:

- variables,
- constants,
- cases, and
- constructs.

A **variable** is an attribute whose value can change. For example, the price of a product is an attribute of that product and generally varies over time. If the price does not change, it is a **constant**. A **case** (or **observation**) consists of all the variables that belong to an object such as a customer, a company, or a country.

The relationship between variables and cases is that within one case we usually find multiple variables. Table 3.1 includes six variables: *type of car bought* and the customer's *age*, as well as *brand_1*, *brand_2*, and *brand_3*, which capture statements related to brand trust. In the lower rows, you see four observations.

Another important and frequently used term in market research is **construct**, which refers to a variable that is not directly observable (i.e., a **latent variable**). More precisely, a construct is used to represent latent concepts in statistical analyses. **Latent concepts** represent broad ideas or thoughts about certain phenomena that researchers have established and want to measure in their research (e.g., Bollen 2002). However, constructs cannot be measured directly, as respondents cannot articulate a single response that will completely and perfectly provide a measure of that concept. For example, constructs such as satisfaction, loyalty, and brand trust cannot be measured directly. However, we can measure satisfaction, loyalty, and brand trust by means of several **items**. The term items (or *indicators*) is normally used to indicate posed survey questions. Measuring constructs requires combining items to form a *multi-item scale*, an example of which appears in Table 3.1 in the form of three items *Brand_1* ("This brand's product claims are believable"), *Brand_2* ("This brand delivers what it promises"), and *Brand_3* ("This brand has a name that you can trust"). Bear in mind that not all items are

Table 3.1 Quantitative data

Variable name	*Type of car bought*	*Customer's Age*	*Brand_1*	*Brand_2*	*Brand_3*
Description	Name of car bought	Age in years	This brand's product claims are believable	This brand delivers what it promises	This brand has a name that you can trust
Customer 1	BMW 328i	29	6	5	7
Customer 2	Mercedes C180K	45	6	6	6
Customer 3	VW Passat 2.0 TFSI	35	7	5	5
Customer 4	BMW 525ix	61	5	4	5

Coding for *Brand_1*, *Brand_2*, and *Brand_3*: 1 = fully disagree, 7 = fully agree

constructs, for example, the *type of car bought* and *customer's age* in Table 3.1 are not a construct, as these are observable and a single response can measure it accurately and fully.

Like constructs, an **index** also consists of sets of variables. The difference is that an index is created by the variable's "causes." For example, we can create an index of information search activities, which is the sum of the information that customers require from dealers, the promotional materials, the Internet, and other sources. This measure of information search activities is also referred to as a *composite measure*, but, unlike a construct, the items in an index define what we want to measure. For example, the *Retail Price Index* consists of a "shopping" bag of common retail products multiplied by their price. Unlike a construct, each item in a scale captures a part of the index perfectly.

The procedure of combining several items is called **scale development**, **operationalization,** or, in the case of an index, **index construction**. These procedures involve a combination of theory and statistical analysis, such as factor analysis (discussed in Chap. 8) aimed at developing an appropriate construct measure. For example, in Table 3.1, *Brand_1*, *Brand_2*, and *Brand_3* are items that belong to a construct called *brand trust* (as defined by Erdem and Swait 2004). The construct is not an individual item that you see in the list, but it is captured by calculating the average of several related items. Thus, in terms of brand trust, the score of customer 1 is $(6 + 5 + 7)/3 = 6$.

But how do we decide which and how many items to use when measuring specific constructs? To answer these questions, market researchers make use of scale development procedures. These procedures follow an iterative process with several steps and feedback loops. For example, DeVellis (2017) provides a thorough introduction to scale development. Unfortunately, scale development requires much (technical) expertise. Describing each step is therefore beyond this book's scope. However, many scales do not require this procedure, as existing scales can

be found in scale handbooks, such as the *Handbook of Marketing Scales* by Bearden et al. (2011). Furthermore, marketing and management journals frequently publish research articles that introduce new scales, such as for measuring the reputation of non-profit organizations (e.g., Sarstedt and Schloderer 2010) or for refining existing scales (e.g., Kuppelwieser and Sarstedt 2014). In Box 3.1, we introduce two distinctions that are often used to discuss constructs.

Box 3.1 Types of Constructs

In **reflective constructs**, the items are considered to be manifestations of an underlying construct (i.e., the items reflect the construct). Our brand trust example suggests a reflective construct, as the items reflect trust. Thus, if a respondent changes his assessment of brand trust (e.g., due to a negative brand experience), this is reflected in the answers to the three items. Reflective constructs typically use multiple items (3 or more) to increase the measurement stability and accuracy. If we have multiple items, we can use analysis techniques to inform us about the measurement quality, such as factor analysis and reliability analysis (discussed in Chap. 8). **Formative constructs** consist of several items that define a construct. A typical example is socioeconomic status, which is formed by a combination of education, income, occupation, and residence. If any of these measures increases, the socioeconomic status would increase (even if the other items did not change). Conversely, if a person's socioeconomic status increases, this would not necessarily go hand in hand with an increase in all four measures. The distinction between reflective and formative constructs is that they require different approaches to decide on the type and number of items. For example, reliability analyses (discussed in Chap. 8) cannot be run on formative measures. For an overview of this distinction, see Bollen and Diamantopoulos (2017), Diamantopoulos et al. (2008), or Sarstedt et al. (2016c).

Instead of using multiple items to measure constructs (i.e., **multi-item constructs**), researchers and practitioners frequently use single items (i.e., **single-item constructs**). For example, we may only use "This brand has a name that you can trust" to measure brand trust, instead of using three items. A popular single-item measure is the **Net Promoter Score** (**NPS**), which aims to measure loyalty by using the single question: *"How likely are you to recommend our company/product/service to a friend or colleague?"* (Reichheld 2003). While this is a good way of making the questionnaire shorter, it also reduces the quality of your measures (e.g., Diamantopoulos et al. 2012, Sarstedt et al. 2016a, b). You should therefore avoid using single items to measure constructs unless you only need a rough proxy measure of a latent concept.

3.2.1 Primary and Secondary Data

Generally, we can distinguish between **primary data** and **secondary data**. While primary data are data that a researcher has collected for a specific purpose, another researcher collected the secondary data for another purpose.

The US Consumer Expenditure Survey (www.bls.gov/cex), which makes data available on what people in the US buy, such as insurance, personal care items, and food, is an example of secondary data. It also includes the prices people pay for these products and services. Since these data have already been collected, they are secondary data. If a researcher sends out a survey with various questions to find an answer to a specific issue, the collected data are primary data. If primary data are re-used to answer another research question, they become secondary data.

Secondary and primary data have their own specific advantages and disadvantages, which we illustrate in Table 3.2. The most important reasons for using secondary data are that they tend to be cheaper and quick to obtain access to (although lengthy processes may be involved). For example, if you want to have access to the US Consumer Expenditure Survey, all you have to do is to use your web browser to go to www.bls.gov/cex and to download the required files. However, the authority and competence of some research organizations could be a factor. For example, the claim that Europeans spend 9% of their annual income on health may be more believable if it comes from Eurostat (the statistical office of the European Community) rather than from a single, primary research survey.

However, important secondary data drawbacks are that they may not answer your research question. If you are, for example, interested in the sales of a specific

Table 3.2 The advantages and disadvantages of secondary and primary data

	Secondary data	Primary data
Advantages	– Tends to be cheaper	– Are recent
	– Sample sizes tend to be greater	– Are specific for the purpose
	– Tend to have more authority	– Are proprietary
	– Are usually quick to access	
	– Are easier to compare to other research using the same data	
	– Are sometimes more accurate (e.g., data on competitors)	
Disadvantages	– May be outdated	– Are usually more expensive
	– May not fully fit the problem	– Take longer to collect
	– There may be hidden errors in the data – difficult to assess the data quality	
	– Usually contain only factual data	
	– No control over data collection	
	– May not be reported in the required form (e.g., different units of measurement, definitions, aggregation levels of the data)	

product (and not in a product or service category), the US Expenditure Survey may not help much. In addition, if you are interested in the reasons for people buying products, this type of data may not help answer your question. Lastly, as you did not control the data collection, there may be errors in the data.

In contrast, primary data tend to be highly specific, because the researcher (you!) can influence what the research comprises. In addition, research to gather primary data can be carried out when and where required and competitors cannot access it. However, gathering primary data often requires much time and effort and is therefore usually expensive compared to secondary data.

As a rule, start looking for secondary data first. If they are available, and of acceptable quality, use them! We will discuss ways to gather primary and secondary data in Chap. 4.

3.2.2 Quantitative and Qualitative Data

Data can be quantitative or qualitative. **Quantitative data** are presented in values, whereas qualitative data are not. **Qualitative data** can take many forms, such as words, stories, observations, pictures, and audio. The distinction between qualitative and quantitative data is not as black-and-white as it seems, because quantitative data are based on qualitative judgments. For example, the questions on brand trust in Table 3.1 take the values of 1–7. There is no reason why we could not have used other values to code these answers, such as 0–6, but it is common practice to code the answers to a construct's items on a range of 1–7.

In addition, many of the sources that market researchers use, such as Twitter feeds or Facebook posts, produce qualitative data. Researchers can code attributes of the data, which describe a particular characteristic, thereby turning it into quantitative data. Think, for example, of how people respond to a new product in an interview. We can code the data by setting neutral responses to 0, somewhat positive responses to 1, positive responses to 2, and very positive responses to 3. We have therefore turned qualitative data into quantitative data. Box 3.2 shows an example of how to code qualitative data.

Box 3.2 Coding Qualitative Data

In 2016 Toyota launched new Prius, the Prius Prime (www.toyota.com/priusprime/). Not surprisingly, Facebook reactions were divided. Here are some examples of Facebook posts:

- "Love it! But why only 2 seats at the back? Is there any technical reason for that?"
- "Wondering if leather seats are offered? The shape of the seats looks super comfy!"

(continued)

Box 3.2 (continued)

– "Here's that big black grill on yet another Toyota. Will be very glad when this 'fashion faze' is over."

One way of structuring these responses is to consider the attributes mentioned in the posts. After reading them, you may find that, for example, the seat and styling are attributes. You can then categorize in respect of each post whether the response was negative, neutral, or positive. If you add the actual response, this can later help identify the aspect the posts liked (or disliked). As you can see, we have now turned qualitative data into quantitative data!

Attribute	Negative	Neutral	Positive
Seats	1-why only two seats?	2-are leather seats offered?	2-shape looks super comfy!
Styling	3-big black grill		1-love it!

Qualitative data's biggest strength is their richness, as they have the potential to offer detailed insights into respondents' perceptions, attitudes, and intentions. However, their downside is that qualitative data can be interpreted in many ways. Thus, the process of interpreting qualitative data is subjective. To reduce subjectivity, (multiple) trained researchers should code qualitative data. The distinction between quantitative and qualitative data is closely related to that between quantitative and qualitative research, which we discuss in Box 3.3. Most people think of quantitative data as being more factual and precise than qualitative data, but this is not necessarily true. Rather, how well qualitative data have been collected and/or coded into quantitative data is important.

3.3 Unit of Analysis

The **unit of analysis** is the level at which a variable is measured. Researchers often ignore this aspect, but it is crucial because it determines what we can learn from the data. Typical measurement levels include that of the respondents, customers, stores, companies, or countries. It is best to use data at the lowest possible level, because this provides more detail. If we need these data at another level, we can aggregate them. *Aggregating data* means that we sum up a variable at a lower level to create a variable at a higher level. For example, if we know how many cars all car dealers in a country sell, we can take the sum of all the dealer sales, to create a variable measuring countrywide car sales. Aggregation is not possible if we have incomplete or missing data at the lower levels.

Box 3.3 Quantitative Research and Qualitative Research

Market researchers often label themselves as either quantitative or qualitative researchers. The two types of researchers use different methodologies, different types of data, and focus on different research questions. Most people regard the difference between qualitative and quantitative as the difference between numbers and words, with quantitative researchers focusing on numbers and qualitative researchers on words. This distinction is not accurate, as many qualitative researchers use numbers in their analyses. The distinction should instead depend on when the information is quantified. If we know the values that may occur in the data even before the research starts, we conduct quantitative research. If we only know this after the data have been collected, we conduct qualitative research. Think of it in this way: If we ask survey questions and use a few closed questions, such as "Is this product of good quality?," and the respondents can either choose "Completely disagree," "Somewhat disagree," "Neutral," "Somewhat agree," and "Completely agree," we know that the data we will obtain from this will—at most—contain five different values. Because we know all possible values beforehand, the data are quantified beforehand. If, on the other hand, we ask someone "Is this product of good quality?," he or she could give many different answers, such as "Yes," "No," "Perhaps," "Last time yes, but lately…". This means we have no idea what the possible answer values are. Therefore, these data are qualitative. We can, however, recode these qualitative data, for example, as described in Box 3.2, and assign values to each response. Thus, we quantify the data, allowing further statistical analysis.

Qualitative research accounts for 17% of money spent in the market research industry, with quantitative research making up the rest.[1] Practically, market research is often hard to categorize as qualitative or quantitative, as it may include elements of both. Research that includes both elements is sometimes called *hybrid market research*, *fused market research*, or simply *mixed methodology*.

3.4 Dependence of Observations

A key issue for any data is the degree to which observations are related, or the **dependence of observations**. If we have exactly one observation from each individual, store, company, or country, we label the observations independent.

[1] See ESOMAR Global Market Research Report 2013.

That is, the observations are unrelated. If we have multiple observations of each individual, store, company, or country, we label them dependent. For example, we could ask respondents to rate a type of Cola, then show them an advertisement, and again ask them to rate the same type of Cola. Although the advertisement may influence the respondents, it is likely that the first response and second response will be related. That is, if the respondents first rated the Cola negatively, the chance is higher that they will continue to rate the Cola negative rather than positive after the advertisement. If the observations are dependent, this often impacts the type of analysis we should use. For example, in Chap. 6, we discuss the difference between the independent samples *t*-test (for *independent observations*) and the paired samples *t*-test (for *dependent observations*).

3.5 Dependent and Independent Variables

Dependent variables represent the outcome that market researchers study, while *independent variables* are those used to explain the dependent variable(s). For example, if we use the amount of advertising to explain sales, then advertising is the independent variable and sales the dependent.

This distinction is artificial, as all variables depend on other variables. For example, the amount of advertising depends on how important the product is for a company, the company's strategy, and other factors. However, this distinction is frequently used in the application of statistical methods. While researching relationships between variables, we, on the basis of theory and practical considerations, need to distinguish between the dependent and the independent variables beforehand.

3.6 Measurement Scaling

Not all data are equal! For example, we can calculate the respondents' average age in Table 3.1. However, if we would have coded the color of the car as black = 1, blue = 2, silver = 3 it would not make any sense to calculate the average. Why is this? The values that we have assigned 1, 2, and 3 are arbitrary; we could just as well have changed these value for any other. Therefore, choosing a different coding would lead to different results, which is meaningless. **Measurement scaling** refers to two things: the variables we use for measuring a certain construct (see discussion above) and the level at which a variable is measured, which we discuss in this section. This can be highly confusing!

There are four *levels of measurement*:

- nominal scale,
- ordinal scale,
- interval scale, and
- ratio scale.

Table 3.3 Measurement Scaling

	Label	Order	Differences	Origin is 0
Nominal scale	✓			
Ordinal scale	✓	✓		
Interval scale	✓	✓	✓	
Ratio scale	✓	✓	✓	✓

Fig. 3.1 Meaningless!

These scales relate to how we quantify what we measure. It is vital to know the scale on which something is measured, because, as the gender example above illustrates, the measurement scale determines the analysis techniques we can, or cannot, use. For example, as indicated above, it makes no sense to calculate. We will return to this issue in Chap. 5 and beyond. However, even when we know the scale, be aware that, as Fig. 3.1 shows, meaningful calculations are not always possible!

The *nominal scale* is the most basic level at which we can measure something. Essentially, if we use a nominal scale, we substitute a word for a numerical value. For example, we could code the color of each Prius sold: black = 1, blue = 2, silver = 3. In this example, the numerical values represent nothing more than a label.

The *ordinal scale* provides more information. If a variable is measured on an ordinal scale, increases or decreases in values give meaningful information. For example, if we code the Prius version people bought as the first generation = 1, second generation = 2, third generation = 3, and fourth generation = 4, we know

whether the model is more recent. The ordinal scale provides information about the order of our observations. However, we do not know if the differences in the order are equally spaced. That is, we do not know if the difference between first generation and second generation is the same as between second and third generation, even though the difference in values (1–2 and 2–3) is equal.

If something is measured on an *interval scale*, we have precise information on the rank order at which something is measured and we can interpret the magnitude of the differences in values directly. For example, if the temperature in a car showroom is 23°C, we know that if it drops to 20°C, the difference is exactly 3°C. This difference of 3°C is the same as the increase from 23 to 26°C. This exact "spacing" is called **equidistance.** Equidistant scales are necessary for some analysis techniques, such as factor analysis (discussed in Chap. 8). What the interval scale does not give us, is an absolute zero point. If the temperature is 0°C it may feel cold, but the temperature can drop further. The value of 0 does not therefore mean that there is no temperature at all.

The *ratio scale* provides the most information. If something is measured on a ratio scale, we know that a value of 0 means that that the attribute of that particular variable is not present. For example, if a dealer sells zero Prius cars (value = 0) then he or she really sells none. Or, if we spend no money on advertising a Prius (value = 0), we really spend no money. Therefore, the origin of the variable is equal to 0.

While it is relatively easy to distinguish between the nominal and the interval scales, it is sometimes hard to see the difference between the interval and the ratio scales. The difference between the interval and the ratio scales can be ignored in most statistical methods. Table 3.3 shows the differences between these four scales.

3.7 Validity and Reliability

In any market research process, it is paramount to use "good" measures. Good measures are those that measure what they are supposed to measure and do so consistently. For example, if we are interested in knowing whether customers like a new TV commercial, we could show a commercial and ask the following two questions afterwards:

1. "Did you enjoy watching the commercial?," and
2. "Did the commercial provide the essential information required for a purchase decision?"

How do we know if these questions really measure whether or not the viewers liked the commercial? We can think of this as a measurement problem through which we relate what we want to measure—whether existing customers like a new TV commercial—with what we actually measure in terms of the questions we ask. If these relate perfectly, our actual measurement is equal to what we intend to

measure and we have no measurement error. If these do not relate perfectly, we have *measurement error*.

This measurement error can be divided into a *systematic error* and a *random error*. We can express this as follows, where X_O stands for the observed score (i.e., what the customers indicated), X_T for the true score (i.e., what the customers' true liking of the commercial is), E_S for the systematic error, and E_R for the random error.

$$X_O = X_T + E_S + E_R$$

Systematic error is a measurement error through which we consistently measure higher, or lower, than we want to measure. If we were to ask customers, for example, to evaluate a TV commercial and offer them remuneration in return, they may provide more favorable information than they would otherwise have. This may cause us to think that the TV commercial is systematically more enjoyable than it is in reality. There may also be random errors. Some customers may be having a good day and indicate that they like a commercial, whereas others, who are having a bad day, may do the opposite.

Systematic errors cause the actual measurement to be consistently higher, or lower, than what it should be. On the other hand, random error causes (random) variation between what we actually measure and what we want to measure.

The systematic and random error concepts are important, because they relate to a measure's validity and reliability. **Validity** refers to whether we are measuring what we want to measure and, therefore, to a situation where the systematic error E_S is small. **Reliability** is the degree to which what we measure is free from random error and therefore relates to a situation where the E_R is zero. In Fig. 3.2, we illustrate the difference between reliability and validity by means of a target comparison. In this analogy, different measurements (e.g., of a customer's satisfaction with a specific service) are compared to arrows shot at a target. To measure each score, we have five measurements (indicated by the black circles), which correspond to, for example, questions asked in a survey. The cross indicates their average. Validity describes the cross's proximity to the bull's eye at the target center. The closer the average to the true score, the higher the validity. If several arrows are fired, reliability is the degree to which the arrows are apart. If all the arrows are close together, the measure is reliable, even though it is not necessarily near the bull's eye. This corresponds to the upper left box where we have a scenario in which the measure is reliable, but not valid. In the upper right box, both reliability and validity are given. In the lower left box, though, we have a situation in which the measure is neither reliable, nor valid. This is obviously because the repeated measurements are scattered around and the average does not match the true score. However, even if the latter were the case (i.e., if the cross were in the bull's eye), we would still not consider the measure valid. An unreliable measure can never be valid. If we repeated the measurement, say, five more times, the random error would probably shift the cross to a different position. Reliability is therefore a necessary condition for validity. This is also

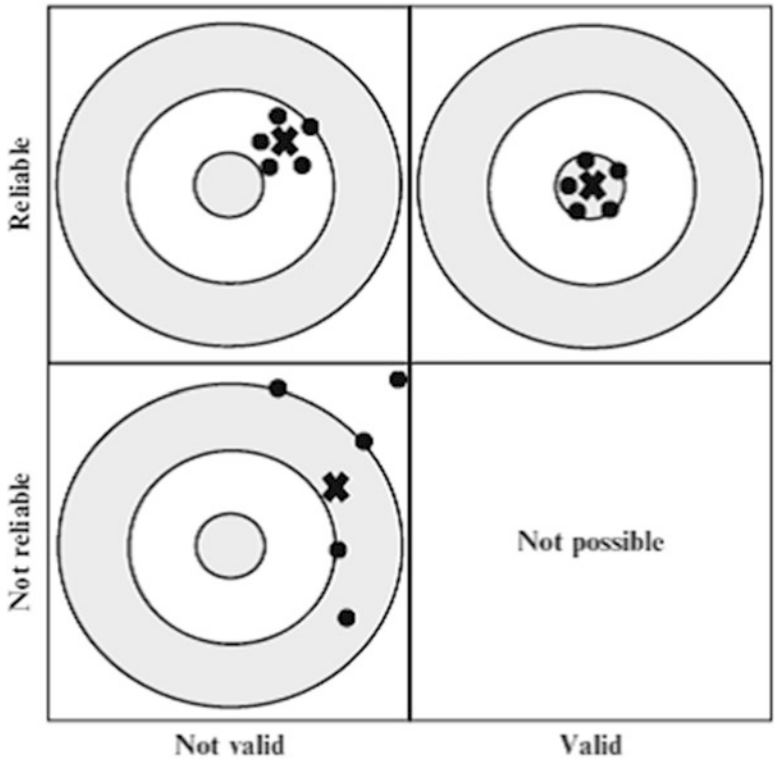

Fig. 3.2 Validity and reliability

why the scenario that is not reliable/valid (lower right box) is not included, as it is not possible for a measure to be valid, but not reliable.

3.7.1 Types of Validity

For some variables, such as length or income, we can objectively verify what the true score is. For constructs, such as satisfaction, loyalty and brand trust, this is impossible. From a philosophical point of view, one could even argue that there is no "true" score for a construct. So how do we know if a measure is valid? Because there is no objective way of verifying what we are measuring, several forms of validity have been developed, including face, content, predictive, criterion, discriminant, and nomological validity (Netemeyer et al. 2003). Researchers frequently summarize these validity types under the umbrella term **construct validity**, which relates to the correspondence between a measure at the conceptual level and a purported measure. The different types of validity help us understand the association between what we should measure and what we actually measure, thereby increasing the likelihood of adequately measuring the latent concept under consideration.

- **Face validity** is an absolute minimum requirement for a variable to be valid and refers to whether a variable reflects what you want to measure. Essentially, face validity exists if a measure seems to make sense. For example, if you want to measure trust, using items such as "this company is honest and truthful" makes a lot of sense, whereas "this company is not well known" makes little sense. Researchers should agree on the face validity before starting the actual measurement. Face validity is usually determined by using a sample of experts who discuss and agree on the degree of face validity (this is also referred to as *expert validity*).
- **Content validity** is strongly related to face validity, but is more formalized. To assess content validity, researchers need to first define what they want to measure and discuss what is included in the definition and what not. For example, trust between businesses is often defined as the extent to which a firm believes that its exchange partner is honest and/or benevolent (Geyskens et al. 1998). This definition clearly indicates what should be mentioned in the questions used to measure trust (honesty and benevolence). After researchers have defined what they want to measure, questions have to be developed that relate closely to the definition. Consequently, content validity is mostly achieved prior to the actual measurement.
- **Predictive validity** requires a measure to be highly correlated (see Chap. 5 for an introduction to correlations) with an outcome variable, measured at a later point in time, to which it is conceptually strongly related. For example, loyalty should lead to people purchasing a product in the future. Similarly, a measure of satisfaction should be predictive of people not complaining about a product or service. Assessing predictive validity requires collecting data at two points in time and therefore requires a greater effort. If both measures (i.e., the one to be evaluated and the outcome variable) are measured at the same point in time, we call this **criterion validity**.
- **Discriminant validity** ensures that a measure is empirically unique and represents phenomena of interest that other measures in a model do not capture. For example, customer satisfaction and customer loyalty are two distinct latent concepts. Discriminant validity requires the constructs used to measure these two concepts to also be empirically distinct (i.e., they should not correlate too highly).
- **Nomological validity** is the degree to which a construct behaves as it should in a system of related constructs. For example, customer expectations, perceived quality, and value have a significant influence on customer satisfaction. Similarly, satisfaction generally relates positively to customer loyalty. As such, you would expect the measure of satisfaction that you are evaluating to correlate with these measure.

3.7.2 Types of Reliability

How do we know if a measure is reliable? Three key factors are used to assess reliability: test-retest reliability, internal consistency reliability, and inter-rater reliability (Mitchell and Jolley 2013).

Test–retest reliability means that if we measure something twice (also called the *stability of the measurement*), we expect similar outcomes. The stability of measurement requires a market researcher to have collected two data samples, is therefore costly, and could prolong the research process. Operationally, researchers administer the same test to the same sample on two different occasions and evaluate how strongly the measurements are correlated. We would expect the two measurements to correlate highly if a measure is reliable. This approach is not without problems, as it is often hard, if not impossible, to survey the same people twice. Furthermore, the respondents may learn from past surveys, leading to *practice effects*. In addition, it may be easier to recall events the second time a survey is administered. Moreover, test–retest approaches do not work if a survey concentrates on specific time points. If we ask respondents to provide information on their last restaurant experience, the second test might relate to a different restaurant experience. Thus, test–retest reliability can only be assessed in terms of variables that are stable over time.

Internal consistency reliability is by far the most common way of assessing reliability. Internal consistency reliability requires researchers to simultaneously use multiple items to measure the same concept. Think, for example, of the set of questions commonly used to measure brand trust (i.e., "This brand's product claims are believable," "This brand delivers what it promises," and "This brand has a name that you can trust"). If these items relate strongly, there is a considerable degree of internal consistency. There are several ways to calculate indices of internal consistency, including split-half reliability and Cronbach's α (pronounced as alpha), which we discuss in Chap. 8.

Inter-rater reliability is used to assess the reliability of secondary data or qualitative data. If you want to identify, for example the most ethical organizations in an industry, you could ask several experts to provide a rating and then calculate the degree to which their answers relate.

3.8 Population and Sampling

A **population** is the group of units about which we want to make judgments. These units can be groups of individuals, customers, companies, products, or just about any subject in which you are interested. Populations can be defined very broadly, such as the people living in Canada, or very narrowly, such as the directors of large hospitals in Belgium. The research conducted and the research goal determine who or what the population will be.

Sampling is the process through which we select cases from a population. The most important aspect of sampling is that the selected sample is representative of the population. Representative means that the characteristics of the sample closely match those of the population. In Box 3.4, we discuss how to determine whether a sample is representative of the population.

Box 3.4 Do I Have a Representative Sample?
It is important for market researchers that their sample is representative of the population. How can we determine whether this is so?

- The best way to test whether the sample relates to the population is to use a database with information on the population (or to draw the sample from such databases). For example, the Amadeus database (www.bvdinfo.com) provides information on public and private companies around the world at the population level. We can (statistically) compare the information from these databases to the selected sample. However, this approach can only support the tested variables' representativeness; that is, the *specific representativeness*. Conversely, *global representativeness*—that is, matching the distribution of all the characteristics of interest to the research question, but which lie outside their scope—cannot be achieved without a census (Kaplan 1964).
- You can use (industry) experts to judge the quality of your sample. They may look at issues such as the type and proportion of organizations in your sample and population.
- To check whether the responses of people included in your research differ significantly from those of non-respondents (which would lead to your sample not being representative), you can use the **Armstrong and Overton procedure**. This procedure calls for comparing the first 50% of respondents to the last 50% in respect of key demographic variables. The idea behind this procedure is that later respondents more closely match the characteristics of non-respondents. If these differences are not significant (e.g., through hypothesis tests, discussed in Chap. 6), we find some support for there being little, or no, response bias (see Armstrong and Overton 1977). When the survey design includes multiple waves (e.g. the first wave of the survey is web-based and the second wave is by phone), this procedure is generally amended by comparing the last wave of respondents in a survey design to the earlier waves. There is some evidence that this procedure is better than Armstrong and Overton's original procedure (Lindner et al. 2001).
- Using follow-up procedures, a small sample of randomly chosen non-respondents can again be contacted to request their cooperation. This small sample can be compared against the responses obtained earlier to test for differences.

When we develop a sampling strategy, we have three key choices:

- census,
- probability sampling, and
- non-probability sampling.

Box 3.5 The US Census

https://www.youtube.com/user/uscensusbureau

If we are lucky and somehow manage to include every unit of the population in our study, we have conducted a **census**. Thus, strictly speaking, this is not a sampling strategy. Census studies are rare, because they are very costly and because missing just a small part of the population can have dramatic consequences. For example, if we were to conduct a census study of directors of Luxemburg banks, we may miss a few because they were too busy to participate. If these busy directors happen to be those of the very largest companies, any information we collect would underestimate the effects of variables that are more important at large banks. Census studies work best if the population is small, well-defined, and accessible. Sometimes census studies are also conducted for specific reasons. For example, the US Census Bureau is required to hold a census of all persons resident in the US every 10 years. Check out the US Census Bureau's YouTube channel using the mobile tag or URL in Box 3.5 to find out more about the US Census Bureau.

If we select part of the population, we can distinguish two types of approaches: probability sampling and non-probability sampling. Figure 3.3 provides an overview of the different sampling procedures, which we will discuss in the following sections.

3.8.1 Probability Sampling

Probability sampling approaches provide every individual in the population with a chance (not equal to zero) of being included in the sample (Cochran 1977, Levy and Lemeshow 2013). This is often achieved by using an accurate *sampling frame*, which is a list of individuals in the population. There are various sampling frames, such as Dun & Bradstreet's Selectory database (includes executives and

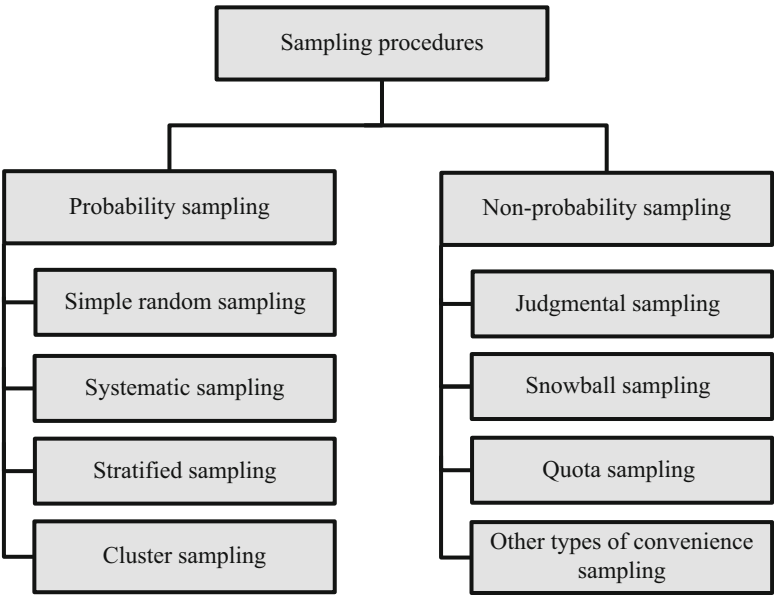

Fig. 3.3 Sampling procedures

companies), the Mint databases (includes companies in North and South America, Italy, Korea, the Netherlands, and the UK), and telephone directories. These sampling frames rarely cover the population of interest completely and often include outdated information, but are frequently used due to their ease of use and availability. If the sampling frame and population are very similar, we have little **sampling error**. Starting with a good-quality sampling frame, we can use several methods to select units from it (Sarstedt et al. 2017).

The easiest way is to use *simple random sampling*, which is achieved by randomly selecting the number of cases required. This can be achieved by using specialized software, or using Stata.[2]

Systematic sampling uses a different procedure. We first randomize the order of all the observations, number them and, finally, select every n^{th} observation. For example, if our sampling frame consists of 1000 firms and we wish to select just 100 firms, we could select the 1^{st} observation, the 11^{th}, the 21^{st}, etc. until we reach the end of the sampling frame and have our 100 observations.

Stratified sampling and cluster sampling are more elaborate techniques of probability sampling requiring us to divide the sampling frame into different groups. When we use *stratified sampling*, we divide the population into several different homogenous groups called *strata*. These strata are based on key sample

[2]See www.stata.com/support/faqs/statistics/random-samples for details. Stata will be discussed in detail in Chap. 5 and beyond.

characteristics, such as different departments in organizations, or the areas in which consumers live. Subsequently, we draw a random number of observations from each stratum. While stratified sampling is more complex and requires accurate knowledge of the sampling frame and population, it also helps ensure that the sampling frame's characteristics are similar to those of the sample.

Cluster sampling requires dividing the population into different heterogeneous groups, with each group's characteristics similar to those of the population. For example, we can divide a country's consumers into different provinces, counties, and councils. Several of these groups could have key characteristics (e.g., income, political preference, household composition) in common, which are very similar (representative of) to those of the population. We can select one or more of these representative groups and use random sampling to select observations that represent this group. This technique requires knowledge of the sampling frame and population, but is convenient because gathering data from one group is cheaper and less time consuming.

Generally, all probability sampling methods allow for drawing representative samples from the target population. However, simple random sampling and stratified sampling are considered superior in terms of drawing representative samples. For a detailed discussion, see Sarstedt et al. (2017).

Stata features several advanced methods to deal with sampling. A few are discussed on http://www.ats.ucla.edu/stat/stata/library/svy_survey.htm. For detail, please see http://www.stata.com/manuals13/svy.pdf

3.8.2 Non-probability Sampling

Non-probability sampling procedures do not give every individual in the population an equal chance of being included in the sample (Cochran 1977, Levy and Lemeshow 2013). This is a drawback, because the resulting sample is most certainly not representative of the population, which may bias the subsequent analyses' results. Nevertheless, non-probability sampling procedures are frequently used as they are easily executed, and are normally less costly than probability sampling methods. Popular non-probability procedures include judgmental sampling, snowball sampling, and quota sampling (Sarstedt et al. 2017).

Judgmental sampling is based on researchers taking an informed guess regarding which individuals should be included. For example, research companies often have panels of respondents who are continuously used in research. Asking these people to participate in a new study may provide useful information if we know, from experience, that the panel has little sampling frame error.

Snowball sampling involves existing study participants to recruit other individuals from among their acquaintances. Snowball sampling is predominantly used if access to individuals is difficult. People such as directors, doctors, and high-level managers often have little time and are, consequently, difficult to involve. If we can ask just a few of them to provide the names and the details of others in a

similar position, we can expand our sample quickly and then access them. Similarly, if you post a link to an online questionnaire on your LinkedIn or Facebook page (or send out a link via email) and ask your friends to share it with others, this is snowball sampling.

Quota sampling occurs when we select observations for the sample that are based on pre-specified characteristics, resulting in the total sample having the same distribution of characteristics assumed to exist in the population being studied. In other words, the researcher aims to represent the major characteristics of the population by sampling a proportional amount of each (which makes the approach similar to stratified sampling). Let's say, for example, that you want to obtain a quota sample of 100 people based on gender. First you need to find what proportion of the population is men and what women. If you find that the larger population is 40% women and 60% men, you need a sample of 40 women and 60 men for a total of 100 respondents. You then start sampling and continue until you have reached exactly the same proportions and then stop. Consequently, if you already have 40 women for the sample, but not yet 60 men, you continue to sample men and discard any female respondents that come along. However, since the selection of the observations does not occur randomly, this makes quota sampling a non-probability technique. That is, once the quota has been fulfilled for a certain characteristic (e.g., females), you no longer allow any observations with this specific characteristic in the sample. This systematic component of the sampling approach can introduce a sampling error. Nevertheless, quota sampling is very effective and inexpensive, making it the most important sampling procedure in practitioner market research.

Convenience sampling is a catch-all term for methods (including the three non-probability sampling techniques just described) in which the researcher draws a sample from that part of the population that is close at hand. For example, we can use *mall intercepts* to ask people in a shopping mall if they want to fill out a survey. The researcher's control over who ends up in the sample is limited and influenced by situational factors.

3.8.3 Probability or Non-probability Sampling?

Probability sampling methods are recommended, as they result in representative samples. Nevertheless, judgmental and, especially, quota sampling might also lead to (specific) representativeness (e.g., Moser and Stuart 1953; Stephenson 1979). However, both methods' ability to be representative depends strongly on the researcher's knowledge (Kukull and Ganguli 2012). Only when the researcher considers all the factors that have a significant bearing on the effect under study, will these methods lead to a representative sample. However, snowball sampling never leads to a representative sample, as the entire process depends on the participants' referrals. Likewise, convenience sampling will almost never yield a representative sample, because observations are only selected if they can be accessed easily and conveniently. See Sarstedt et al. (2017) for further details.

3.9 Sample Sizes

After determining the sampling procedure, we have to determine the **sample size**. Larger sample sizes increase the precision of the research, but are also much more expensive to collect. The gains in precision decrease as the sample size increases (in Box 6.3 we discuss the question whether a sample size can be too large in the context of significance testing). It may seem surprising that relatively small sample sizes are precise, but the strength of samples comes from selecting samples accurately, rather their size. Furthermore, the required sample size has very little relation to the population size. That is, a sample of 100 employees from a company with 100,000 employees can be nearly as accurate as selecting 100 employees from a company with 1,000 employees.

There are some problems with selecting sample sizes. The first is that market research companies often push their clients to accept large sample sizes. Since the fee for market research services, such as those offered by Qualtrics or Toluna, is often directly dependent on the sample size, increasing the sample size increases the market research company's profit. Second, if we want to compare different groups, we need to multiply the required sample by the number of groups included. That is, if 150 observations are sufficient to measure how much people spend on organic food, 2 times 150 observations are necessary to compare singles and couples' expenditure on organic food.

The figures mentioned above are net sample sizes; that is, these are the actual (usable) number of observations we should have. Owing to non-response (discussed in Chaps. 4 and 5), a multiple of the initial sample size is normally necessary to obtain the desired sample size. Before collecting data, we should have an idea of the percentage of respondents we are likely to reach (often high), a percentage estimate of the respondents willing to help (often low), as well as a percentage estimate of the respondents likely to fill out the survey correctly (often high). For example, if we expect to reach 80% of the identifiable respondents, and if 25% are likely to help, and 75% of those who help are likely to fully fill out the questionnaire, only 15% (0.80·0.25·0.75) of identifiable respondents are likely to provide a usable response. Thus, if we wish to obtain a net sample size of 100, we need to send out $\left(\frac{desired\ sample\ size}{likely\ usable\ responses}\right) = 100/0.15 = 667$ surveys. In Chap. 4, we will discuss how we can increase response rates (the percentage of people willing to help).

3.10 Review Questions

1. Explain the difference between items and constructs.
2. What is the difference between reflective and formative constructs?
3. Explain the difference between quantitative and qualitative data and give examples of each type.
4. What is the scale on which the following variables are measured?
   - The amount of money a customer spends on shoes.

- A product's country-of-origin.
- The number of times an individual makes a complaint.
- A test's grades.
- The color of a mobile phone.

5. Try to find two websites offering secondary data and discuss the kind of data described. Are these qualitative or quantitative data? What is the unit of analysis and how are the data measured?
6. What are "good data"?
7. Discuss concepts reliability and validity. How do they relate to each other?
8. Please comment on the following statement: "Face and content validity are essentially the same."
9. What is the difference between predictive and criterion validity?
10. Imagine you have just been asked to execute a market research study to estimate the market for notebooks priced $300 or less. What sampling approach would you propose to the client?
11. Imagine that a university decides to evaluate their students' satisfaction. To do so, employees issue every 10th student at the student cafeteria on one weekday with a questionnaire. Which type of sampling is conducted in this situation? Can the resulting sample be representative of the student population?

3.11 Further Readings

Mitchell, M. L., & Jolley, J. M. (2013). *Research design explained* (8th ed.). Belmont, CA: Wadsworth.

The book offers an in-depth discussion of different types of reliability and validity, including how to assess them.

Churchill, G. A. (1979). A paradigm for developing better measures for marketing constructs. *Journal of Marketing Research*, *16*(1), 64–73.

A landmark article that marked the start of the rethinking process on how to adequately measure constructs.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: John Wiley and Sons.

This is a seminal text on sampling techniques, providing a thorough introduction to this topic. However, please note that most descriptions are rather technical and require a sound understanding of statistics.

Diamantopoulos A, Winklhofer HM (2001) Index construction with formative indicators: an alternative to scale development. *Journal of Marketing Research*, *38*(2), 269–277.

In this seminal article the authors provide guidelines on how to operationalize formative constructs.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Thousand Oaks, CA: Sage.

This is a very accessible book which guides the reader through the classic way of developing multi-item scales. The text does not discuss how to operationalize formative constructs, though.

Marketing Scales Database at www.marketingscales.com/search/search.php

This website offers an easy-to-search database of marketing-related scales. A description is given of every scale; the scale origin, reliability, and validity are discussed and the items given.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage.

Like DeVellis (2017), *this book presents an excellent introduction to the principles of the scale development of measurement in general.*

References

Armstrong, J. S., & Overton, T. S. (1977). Estimating nonresponse bias in mail surveys. *Journal of Marketing Research, 14*(3), 396–403.

Bearden, W. O., Netemeyer, R. G., & Haws, K. L. (2011). *Handbook of marketing scales. Multi-item measures for marketing and consumer behavior research* (3rd ed.). Thousand Oaks: Sage.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*(1), 605–634.

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods, 22*(3), 581–596.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Thousand Oaks: Sage.

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61*(12), 1203–1218.

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science, 40*(3), 434–449.

Erdem, T., & Swait, J. (2004). Brand credibility, brand consideration, and choice. *Journal of Consumer Research, 31*(1), 191–198.

Geyskens, I., Steenkamp, J.-B. E. M., & Kumar, N. (1998). Generalizations about trust in marketing channel relationships using meta-analysis. *International Journal of Research in Marketing, 15*(3), 223–248.

Kaplan, A. (1964). *The conduct of inquiry*. San Francisco: Chandler.

Kukull, W. A., & Ganguli, M. (2012). Generalizability. The trees, the forest, and the low-hanging fruit. *Neurology, 78*(23), 1886–1891.

Kuppelwieser, V., & Sarstedt, M. (2014). Confusion about the dimensionality and measurement specification of the future time perspective scale. *International Journal of Advertising, 33*(1), 113–136.

Levy, P. S., & Lemeshow, S. (2013). *Sampling of populations: Methods and applications* (5th ed.). Hoboken: Wiley.

Lindner, J. R., Murphy, T. H., & Briers, G. E. (2001). Handling nonresponse in social science research. *Journal of Agricultural Education, 42*(4), 43–53.

Mitchell, M. L., & Jolley, J. M. (2013). *Research design explained* (8th ed.). Belmont: Wadsworth.

Moser, C. A., & Stuart, A. (1953). An experimental study of quota sampling. *Journal of the Royal Statistical Society. Series A (General), 116*(4), 349–405.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks: Sage.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review, 81*(12), 46–55.

Sarstedt, M., & Schloderer, M. P. (2010). Developing a measurement approach for reputation of nonprofit organizations. *International Journal of Nonprofit and Voluntary Sector Marketing, 15*(3), 276–299.

Sarstedt, M., Diamantopoulos, A., Salzberger, T., & Baumgartner, P. (2016a). Selecting single items to measure doubly-concrete constructs: A cautionary tale. *Journal of Business Research, 69*(8), 3159–3167.

Sarstedt, M., Diamantopoulos, A., & Salzberger, T. (2016b). Should we use single items? Better not. *Journal of Business Research, 69*(8), 3199–3203.

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016c). Estimation issues with PLS and CBSEM: Where the bias lies! *Journal of Business Research, 69*(10), 3998–4010.

Sarstedt, M., Bengart, P., Shaltoni, A. M., & Lehmann, S. (2017, forthcoming). The use of sampling methods in advertising research: A gap between theory and practice. *International Journal of Advertising*.

Stephenson, C. B. (1979). Probability sampling with quotas: An experiment. *Public Opinion Quarterly, 43*(4), 477–496.