Doing Research in Political Science

An Introduction to Comparative Methods and Statistics

Paul Pennings, Hans Keman
and Jan Kleinnijenhuis

4

Concepts, cases, data and measurement

CONTENTS

This chapter focuses on the measurement of political concepts. A concept has been measured whenever data have been found that indicate whether, or to what degree, the concept applies to an observed case. A measurement is simply defined as an assignment of a value (or datum) to an observed case (or an observed unit) on a variable (a concept). One measurement of the concept bilateralism, for example, is obtained by assigning the value 'yes' to Germany, another by assigning the value 'no' to New Zealand.

Whether it makes sense to ask not only whether a concept applies, but also which degree of the concept applies, depends not only on its definition but also on its level of measurement. The previous chapter showed also that the units (cases) to which a concept applies are by no means trivial in comparative political science.

Measurements always presume the availability of data. Various types of available data – e.g. data from statistical agencies, or survey data – will be discussed in Section 4.1. Separate measurements might be represented as the entries (or cells) in a rectangular data matrix with the units (cases) as rows and the variables (concepts) as columns (see Figure 1.1). This rectangular data matrix which brings together the various measurements for a set of concepts with respect to a set of units, is treated in Section 4.3. The problem of generalizability of research findings, which arises when the available data constitute only a subset of all conceivable data, is introduced in Section 4.2. Often a variety of data is useful to judge whether a single concept applies to a given unit. Scalability analysis (Sections 4.4 and 4.5) can be used to test the reliability of multiple indicators.

4.1 Data and Data Collection in Political Science

Political science is in our view an empirical science. Its inspiration may well hinge on philosophies of the good world, but more or less irrefutable facts constitute its basis. The relevant facts can be gathered from different sources.

4.1.1 Data obtained from official statistical agencies

An obvious source for comparative information on political processes is the data published on a yearly or quarterly basis by national and international statistical agencies, such as the IMF, the World Bank and the OECD, although the focus of these data is on economics. The statistical yearbooks from the *Encyclopaedia Britannica,* the yearbooks from SIPRI on military expenditures and warfare, and the Yale University *World Handbook of Political and Social Indicators* are additionally useful. All types of data sets with respect to political and social indicators compiled by political scientists and sociologists have been made publicly available. Some journals in the field of political science, for example the *European Journal of Political Research* publish data sets collected by political scientists also. Table 4.1 gives an overview of available data sets for comparative political science. Most university libraries provide online access to these databases.

Table 4.1 *Commonly used data sets from statistical agencies in political science*[*]

IMF (www.imf.org)	International financial statistics Direction of trade statistics
OECD (www.oecd.org)	Historical statistics Employment outlook OECD economic surveys (country reports)
ILO (www.ilo.org)	Labour force statistics
Encyclopaedia Britannica (DVD, also www.eb.com)	Yearbooks, Statistical Addendum (comparative data on government, elections, economics and demography)
Worldwide Elections (http://sshl.ucsd.edu/election/world.html)	Comparative data on parties contesting elections and election outcomes
SIPRI (Stockholm International Peace Research Institute) (http://databases.sipri.se)	Yearbook of world armaments and disarmament
ICPSR (www.icpsr.umich.edu)	Archive of (partly comparative) data sets gathered by political scientists
LexisNexis (www.lexis.com)	Archive of textual accounts of the political process (e.g. newspapers, magazines)
IPU (www.ipu.org)	Comparative database on the features of parliaments and electoral systems around the world
Parties and elections (www.parties-and-elections.de)	Database of all elections in Europe since 1945 and on political parties
Comparative political data sets (http://www.ipw.unibe.ch/mitarbeiter/ru_armingeon/CPD_Set_en.asp)	Data on politics and expenditures in all OECD countries and other central and eastern European countries

[*]Most university libraries have licenses to access these databases.

The compilers of data sets that enable comparisons between nations have usually obtained their data from national statistical agencies. Third World countries, in particular, do not have the statistical agencies to deliver the required data. When data from national agencies are available, they might not match the definitions of the international agencies precisely. Often the data obtained from statistical agencies do not allow for the distinctions desired by political scientists. The data set NIAS.SAV, which is used throughout this book, was compiled by a group of researchers visiting the Netherlands Institute for Advanced Study in the Humanities and Social Sciences in 1995/1996 and updated afterwards by the first author of this book.

4.1.2 Verbal and visual accounts, content analysis

Verbal accounts from politicians, eyewitnesses, journalists and contemporary historians constitute an important source of information for political scientists. These verbal accounts are accompanied in a growing number of cases by visuals on photographs, films and video. Verbal and visual accounts of the political process are provided by the participants in the process as well as by observers and interpreters.

Many contributions of the participants in the political process towards decision-making are recorded officially (e.g. party programmes, parliamentary proceedings). Politicians will use the media to pursue their ends, and will use press conferences, press reports, and 'sound bites' in television programmes to provide additional evidence, or at least additional images, of their daily pursuits.

Altogether the amount of available verbal and visual accounts from the political sphere is overwhelming. Citations, paraphrases and sound bites are the traditional means of mastering, or at least reducing, this overwhelming excess of information. It often remains an open question, however, whether the same citations, or even citations with the same purport, would also have been selected by other citation experts when complex policy documents, party programmes or parliamentary debates are at stake. The *reliability* of citations is low.

The term 'content analysis' refers to 'any technique for making inferences by objectively and systematically identifying specified characteristics of messages' (Holsti, 1969: 14). Content analysis thus aims at data with respect to verbal and visual messages that are more reliable than citations and paraphrases. Content analysis data typically enable systematic comparisons of verbal and visual accounts delivered by one actor at various points in time, or between various sources. Two basic types of content analysis can be distinguished: thematic content analysis and relational content analysis (Roberts, 1997; Popping, 2000).

Thematic content analysis aims at an assessment of the (frequency of the) presence of specified themes, issues, actors, states of affairs, words or ideas in the texts or visuals to be analysed. Which themes, issues or actors are sought depends completely on the theoretical concepts to be operationalized. The themes, issues or actors sought should be mutually exclusive (no overlaps). The complete set should be exhaustive (no unclassified texts). A mutually exclusive set of themes, issues or actors constitutes a nominal variable, since it does not exhibit a rank order. The frequency distribution of such a nominal variable indicates which themes, issues, facts or actors were mentioned more or less frequently in the texts or visuals being analysed. In the Manifesto research project (Budge et al., 2001), for example, a thematic content analysis has been performed of more than a thousand party programmes from industrialized countries (1945–98). Sentences from party programmes were classified into 54 predetermined issue areas, such as 'social justice', 'military positive', 'military negative' or 'economic orthodoxy'. Data from this content analysis will be used in many places in this book.

Relational content analysis aims at an assessment of the relations between actors, issues, ideas, etc., according to the texts or visuals being analysed.

For example, relations between nations are being sought in a content analysis project (COPDAB) by analysing newspaper articles. Its aim is to reconstruct the 'real events' underlying them. Roughly 350,000 events from the period 1948–1978 were construed on the basis of news reports in 77 international newspapers and news magazines, predominantly from the USA and the Middle East. The database consists of subject-nation/predicate/object-nation relationships. Hence, by classifying this type of information it is possible to compare the degree of cooperation or conflict between actors (here: national states) in a reliable and valid fashion.

4.1.3 Questionnaires and surveys

When the personal experiences, perceptions, opinions, attitudes and reported behaviours of persons are crucial to answering a research question, questionnaires and surveys come into play. In questionnaires and surveys the unit of measurement is usually an individual. Influential individuals might be asked, however, to act as the mouthpiece of their company, their party, or even their nation. In the latter case these organizations will usually become the units of analysis.

Here we will use the term *questionnaire* to denote a set of personalized questions that will be posed to a single actor on the basis of a preliminary investigation with respect to the actor's experiences, policy and world view. Usually the interview design allows subsequent questions to be asked that were not foreseen in the interview script. Subsequent questions will depend on the answers of the subject that are the starting point for an interview with a person. Questionnaires and interviews are at the heart of journalism. Political scientists will use them to reveal inside views of the political process. The reliability of answers obtained during an interview relies on an exchange between the interviewer and the respondent. Elite subjects willing to give an interview often want to stress their policy views once more, whereas the interviewer wants to have answers to preconceived questions. Friction in elite interviews is often enhanced by abstract, overarching questions that do not account for the multitude and diversity of daily experiences of elite persons on the basis of which answers to these questions have to be assembled. The question 'how much power has A in your opinion?', for example, is a confusing question. Policy experts might be as confused with respect to the various faces of 'power' as political scientists. Abstract, ambiguous and vague questions evoke abstract, ambiguous and vague answers.

The term *survey* is used to denote a standard list of questions that will be posed to a great number of individuals. Usually not the population of all individuals, but a sample from it will be interviewed. Interviews might be conducted by telephone or in a personal setting with an interviewer, usually at the homes of the interviewed persons. Examples of surveys in many countries are the National Election Studies. Commercial marketing agencies conduct surveys on a regular (daily or weekly) basis so as to monitor trends in opinions and behaviours on the basis of which their clients – firms, ministries, and to a minor extent also political parties – base their marketing decisions. A *panel survey* is a special type of survey where the

same respondents are interviewed repeatedly over time. Comparative surveys in several countries are relatively rare. A sociological example, which is also useful in the context of comparative research of political values, is provided by the world value survey designed by Inglehart and colleagues (Inglehart, 1997). Eurobarometer provides comparative data on political attitudes and political behaviour in the European Union. Since many textbooks are available on survey research, we will not delve into it here.

4.2 Sampling and the Basics of Statistical Testing

Usually it is unnecessary to gather measurements on all the empirical cases to which a theory applies. Efficient research bears on a few crucial cases only or on a sample of cases from the population of all cases to which a theory applies. We will start the discussion of sampling here, before the statistics comes in. Sampling inevitably gives rise to the generalizability question. Is it reasonably safe to infer that the research results with respect to the sample will hold for the population of all cases to which the theory applies? An answer to this question depends, of course, on known characteristics of the relationship between the sample and the population.

In a *random sample* every individual from a given population has the same probability of being sampled. Most statistics presume random samples, although random sampling is an ideal type only. Research results that hold for a random sample may not hold for the population as a whole. Interesting research results on the basis of a sample are matched against a dull *null hypothesis* maintaining that in the population as a whole the result does not hold. A first type of error (*type I error*) is to keep maintaining that the interesting result holds for the population as a whole, whereas actually the null hypothesis holds. The aim of statistical testing is to reduce the probability of a type I error to less than a specified level, commonly set at 5 per cent. A *type II error* is made when interesting research results on the basis of a sample are discarded in favour of the null hypothesis, but the null hypothesis is false after all. The so-called 'power' of statistical tests is their ability to reduce type II errors. The power of various statistical tests is too complicated a subject to be discussed in this book.

4.2.1 Statistical inference from a random sample

If in the population the numbers '0' and '1' (e.g. representing 'girls' and 'boys') occur with the same frequency, then selecting a sample of 4 elements from this sample will definitely result in one of 16 sequences with equal probability: 1111, 1110, 1101, … , 0000. Each of these 16 sequences has a probability of 1/16. By counting aspects of these 16 sequences it is easily verified that the probability of getting a sample distribution of either boys only or girls only is 1/8 (1/16 for the sequence 1111 + 1/16 for 0000). Although girls occur precisely as often in the

population as boys, the chance of encountering an equal number of boys and girls in a sample of 4 amounts to 3/8 only (6 of 16 sequences only, namely 1100, 1010, 1001, 0110, 0101, 0011). One is more likely to obtain three times as many exemplars of the one sex than of the other (Probability 1/2, corresponding to 8 from 16 sequences, namely 0001, 0010, 0100, 1000, 0111, 1011, 1101, 1110). If one has found either no girls at all or no boys at all in a sample of 4, and one is willing to accept erroneous assertions one out of five times (type I error of 20 per cent), then statistically speaking the conclusion is warranted that boys and girls do not appear equally frequently in the population, since the chance of finding no boys at all or no girls at all amounted to 1/8 (= 12.5 per cent) only. Statisticians are usually more conservative in the sense of accepting erroneous assertions with respect to the population distribution for less than 5 per cent of the possible number of samples only (type I error < 0.05).

Let us emphasize three aspects of the statistician's line of thought in this simple example. First it should be noted that the statistician's tests are based on counts in an imaginary universe of all conceivable samples that might have been drawn. The second aspect to be noticed is that an important ingredient in the calculus of the statistician is the *sample size.* As long as the number of children in the sample is limited, giving birth to children of the same sex only is no reason to falsify the hypothesis that the odds of getting boys and getting girls are equal.

The third aspect to be aware of is that counts in an imaginary population to which the null hypothesis applies mount up to a *probability distribution* of all counts. Selecting at random sets of children from a school class of boys and girls gives a Newtonian or binomial distribution of the numbers of each gender in the sets. Once the probability distribution is known, statistical testing is straightforward from a mathematical point of view. The question of which probability distribution is appropriate under which circumstances will recur in Section 5.6. Distributions such as the Gaussian or normal distribution, the *t*-distribution, the chi-square distribution and the *F*-distribution play a central role in these sections. Why each distribution applies is a matter for mathematical statistics. Here we will use specific probability distributions on the authority of mathematical statisticians.

4.2.2 Random samples and non-random samples

Most samples are not random. Two types of non-random samples will be discussed here: the stratified sample and the cluster sample. The *stratified sample* intends to be more representative of the population as a whole than a random sample would be. Statistical tests based on random-sample assumptions will be too conservative for a stratified sample. The key to stratified sampling is the use of known population distributions in the sampling plan. If it is known that 50 per cent of mankind are women, and that 20 per cent of men and 22 per cent of women are older than 65, then it is quite natural to draw a stratified sample with 10 per cent of elderly men, 11 per cent of elderly women, 40 per cent of men under 65, and 39 per cent of women under 65. One should keep in mind, though, that the variables of

interest are often not the variables on which the sample is stratified. Samples are usually stratified with respect to demographic characteristics, but the advantage of a demographically stratified sample over a random sample vanishes when the variables of interest are related only remotely to demography.

The *cluster sample*, or multi-level sample, is less representative of the population than a random sample. At the first level, clusters are selected, e.g. municipalities within a nation. At the second level, individuals within the first-level clusters, e.g. inhabitants of a selected municipality, are selected. A special type of a cluster sample is the snowball sample, where a set of individuals is sampled randomly and next the population of relatives of the interviewed person is asked to participate in the interview. The statistical inference problem is double-edged now. In principle one has to infer whether results holding for a sample of inhabitants would hold for the municipality as a whole and next whether results that hold for the sample of municipalities hold for the population as a whole. Cluster sampling is often preferred for pragmatic reasons over random sampling. Progress has been made during the last decade with respect to statistics for multi-level samples (Bryk and Raudenbush, 1995; Snijders and Bosker, 1999; Snijders, 2003; Skrondal and Rabe-Hesketh, 2004), but in this book we will only deal with statistics that assume random samples.

Many economists and political scientists will even perform statistical tests that assume a random sample, when the units of analysis at their disposal amount to the complete population. Economists studying quarterly data will perform statistical tests that assume a random sample, although the population from which these quarters are randomly drawn is metaphysical. Political scientists using data on all democracies for which data are available (western democracies) will perform statistical tests also. The attraction of statistical tests is their property of taking research results more seriously as the number of units of analysis increases. Since increasing the number of units of analysis will also be a means to cancel out random measurement errors and casual interpretation errors, statistical tests that assume a random sample are often used even when this assumption is obviously false.

4.3 Operationalization and Measurement: Linking Data with Concepts and Units

The *operational definition* of a concept prescribes which measurements are appropriate to measure a theoretical concept. The operational definition of a concept bridges the gap between the general definition of a concept and the available data (see Section 3.5). Concept definition is the first filter in the funnel from concepts to data, as Figure 4.1 depicts. *Operationalization* is defined as the set of efforts to obtain an acceptable operational definition, which renders a *valid* transformation that can be *reliably* measured.

The operational definition embedded in the measurement procedure is the next filter. Separate measurements have to be in accordance with the operational

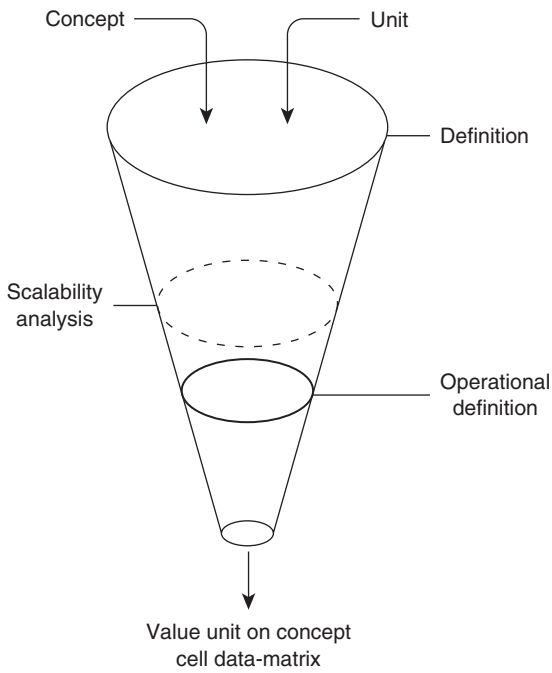

Figure 4.1 *The funnel of operationalization: from a concept and a unit towards a value*

definition, whereas the operational definition has to match the definition of the theoretical concept. The 'salience of an issue for a party', for example, might be defined on a theoretical level as the importance of an issue relative to the importance of other issues according to the policy statements of a party (Budge et al., 2001). If party manifestoes are used as the single source of available data to measure 'issue saliency', then an operational definition might be 'the percentage of sentences in a party manifesto devoted to a given issue'. Usually, various data and, as a consequence, various operational definitions can be imagined to measure a theoretical concept. Alternative operational definitions of issue salience, for example, could refer to speeches in parliament. As compared with the concept definition the operational definition is restricted to the specific method for data collection to be used. An operational definition of policy viewpoints designated to be used in a content analysis of party platforms will differ significantly from an operational definition designated to perform an elite survey among party officials.

Sometimes operational definitions are provided implicitly in the form of an elaborated measurement procedure, coding scheme, or classification scheme. Operational definitions may well include additional guidelines to apply general definitions to a specific empirical context.

Box 4.1 Levels of measurement

Levels of measurement are important for judging what type of statistics can be used or not. Without a proper understanding of these levels a correct choice of technique is impossible. These techniques will be discussed in Chapters 5 and 6 and applied in Part III in examples of existing research.

Measurement level	**Meaning of numbers assigned to categories**	**Examples to be treated in this book**
Dichotomous, binary	Different number = different category	Either use data analysis techniques for nominal scales or use (possibly adjusted) techniques for interval scales
Nominal	As dichotomous, binary	Frequency table or contingency tables
Ordinal	Higher number = higher rank	Either use data analysis techniques for nominal scales (e.g. if number of categories less than 5) or use techniques for interval scales
Interval	Equal interval between numbers = equal difference between categories	Frequency distribution, (rank order) correlation and regression analysis
Ratio	*x* times as far from zero = x times more	Logit and probability statistics
Absolute	Number = number	

Measurement is defined as the assignment of a value on a variable to a unit of measurement in accordance with an operational definition. The measurements within comparative political science map its theoretical concepts into databases that are accessible for data analysis. The assigned values may be visual (e.g. colour graphs on a monitor representing real-time approval of political speeches by members of a focus group), nominal (e.g. yes/no, communist/socialist…/ conservative) or numerical. Length, for example, is measured in numbers of metres and centimetres, the gravity of a war is measured by the number of deaths, and political participation by the number of distinct types of activities aimed at political influence. Distinct visual and nominal codes can be represented as distinct numbers also. The visual, verbal and numerical values for separate units of measurement form a measurement scale with nominal, ordinal, interval or ratio level of measurement (see Section 4.1.2).

Table 4.2 *Data matrix of countries (units of analysis) by population characteristics (columns)*

Country	Year	Population (000s)	Turnout (%)
Italy	1960	50,198	93.7
Italy	2002	57,474	81.2
Sweden	1960	7,480	85.9
Sweden	2002	8,925	80.1
UK	1960	52,373	78.7
UK	2002	60,242	59.4

Each measurement fills in a slot in a data matrix with units (of measurements) in the rows, and variables (indicators of concepts) in the columns. As an example, part of a data matrix with 'population' and 'turnout' as variables and stacked country–year combinations is presented in Table 4.2. The value for 'turnout' in Italy in 2002 was measured as 81.2 per cent, for example.

Putting units of measurement in the rows and not in the columns is a matter of convenience reinforced by statistical packages. Successful measurements result in a completely filled rectangular array, since, for each combination of a unit of measurement and an indicator, a value will be obtained.

The reader should keep in mind that the data matrix in the final analysis often results from data at a lower level. The value of turnout for the Italian population as a whole (unit of analysis), for example, is actually an aggregation of the voting behaviour of individual Italians (unit of measurement in the first stage). The ultimate data matrix with units of analysis in the rows and concepts in the columns often results from a (rowwise) aggregation of data on units of measurement and/or a (columnwise) combination of indicators of the ultimate concepts (see Table 4.2).

4.3.1 Handling missing data

Measurements should ideally result in a completely filled rectangular data matrix. However, often many values in the data matrix remain missing.

Many data are simply not available. In the comparative research of nations it may be impossible to retrieve (recent) data on specific economic or political indicators for the complete set of countries. Next, not all indicators may apply to all units of measurement. Survey interviews often have filter questions, e.g. 'did you vote at the last elections?'. The follow-up question – which party was voted for – will be posed only to respondents who answered that they did indeed cast their vote. A third type of missing value results from rest categories in the measurement process. Substantial hypotheses on parties belonging to one of the ten ideological 'party families' distinguished by Gallagher et al. (2001) do not apply to parties which were coded as 'other parties'. A content analysis classification of issues raised in party programmes may have 'uncoded' as a category. Many questions

in survey research allow for 'don't know' as an answer. Four strategies to deal with missing values will be discussed here.

Inclusion in tables as missing values is appropriate when the number and distribution of missing values is interesting. To answer the simple question 'have the poor a greater propensity to vote leftist?', it would be a good idea to include in the cross-table to answer this question the percentages of the poor and of the wealthy who abstained from voting, since, for the poor, abstention might be an alternative for a vote for the left.

Listwise deletion means that units of measurement with a missing value on one or more of the variables relevant for an analysis are excluded from the analysis. Listwise deletion is appropriate when the excluded units are not extremely important in the research design. When the number of units of measurement is large compared with the number of missing values, this solution is often preferred.

Pairwise deletion is an alternative to listwise deletion in multivariate data analysis when more than two variables with missing values enter the data analysis. As a first step, the bivariate relationships between separate variables might be based on all the cases with non-missing values for the two variables. Next the multivariate analysis will be performed on the bivariate relationships. The advantage is that fewer units of measurement will be discarded. The disadvantage is its obscurity. It is not always easy to reconstruct which units of measurement bear a special weight for the outcomes of data analysis.

Substitution of the missing values by approximations is a third possibility when it is known that a value for the variable must exist. The missing values might be filled in by predicting the true scores on the basis of causal relationships, by intrapolation or extrapolation, or by cross-sectional mean substitution. If, for example, the exact amount of military expenditure of a specific country is unknown, but the gross national product and the number of military personnel are known, and causal relationships between gross national product, military personnel and military expenditures are also known, then an estimate of military expenditure might be given. The estimated expenditures might be predicted from gross national product and military personnel. Intrapolation and extrapolation are obvious means to fill in the gaps in time series. A warning is, however, in order. Intrapolation and extrapolation may result in erroneous estimates of the statistical properties of time series models: data based on intrapolation and extrapolation give rise to a serious underestimation of the jerkiness of changes (see Section 6.7.5).

In sum: missing values create problems. Each treatment has pros and cons. It depends on the research question and the research design which treatment is to be preferred.

4.4 Criteria to Evaluate the Quality of Operationalization and Measurements

Many criteria may be applied to judge the quality of the measurements of a concept. The *efficiency* of measurements relates the quality of measurements to the

time and money invested in getting the data. The *compatibility* of the measurements refers to their usefulness not only in the main research project but also in related research projects that use slightly different data (other nations, other time periods, slightly different data collection methods). The major criteria to judge the quality of measurements are *validity* and *reliability*, however. Measurements that are not valid or not reliable cannot be efficient or compatible with other data either.

The *validity* of measurements, often referred to as *construct validity*, is defined as the degree to which one actually measures whatever concept (or 'construct') the measurement procedure purports to measure. It refers to the closeness of the correspondence between the measurements and the concept being measured. But how to establish this correspondence?

Measurements possess *face validity* when they are perceived as indisputable facts with respect to the measured concept in the scientific community. Assessments of face validity are often based on the agreement of measurement results with common-sense expectations, regardless of the precise definitions of the concept.

Correlational validity (or 'internal validity') is obtained by using a traditional, but imprecise, measurement device as a yardstick to verify the correspondence between the measurements and the concept being measured. Newer measurement devices, e.g. an electron microscope, should be able to reproduce the measurements of the older ones, e.g. a lens microscope, albeit with greater precision. The refined results should, however, correlate highly with the old results.

The *predictive validity* (or 'external validity') of measurements refers to their usefulness in making correct predictions about real-world phenomena. A judgement with respect to external validity presupposes a causal theory with the concept being measured as an independent variable. Let us give an example. One might doubt whether counting the attention given to various issues in party programmes (e.g. Budge et al., 2001) renders valid measurements of the party agenda. An empirical demonstration that government expenditures on issues correspond to the attention given to these issues in the programmes of the governing party (but not with the attention given in the programmes of the opposition parties) renders an external validation for the measurements. Predictive validity is probably the most important hallmark of validity, since it relates the usefulness of the obtained measurements to the context of prevailing theories.

Students will notice that the word 'validity' is not only used in the context of the validity of measurements, but also in the context of the *validity of theories.* As was stated in Section 1.2, a theory is said to be 'internally valid' when it holds for the cases being investigated. A theory is said to be 'externally valid' when the theory also holds for the cases to which the theory applies which were not included in the data analysis. External validity of research findings is a synonym of generalizability of research findings.

Measurements are reliable to the extent that measurements with respect to the same units deliver consistent results. Reliability, however, cannot compensate for low validity. The *reliability* of measurements is related to the validity of measurement in the same way as a standard deviation from the mean is related to the mean. Measurements are not reliable when separate measurements have

a large variance, i.e. when the precise measurement results for a given unit of measurement at a given time are shaky. It should be noted that a negligible variance of separate measurements does not imply that the measurements are valid: they may all be far from the truth collectively. Two varieties of reliability should be distinguished.

- *Intra-observer reliability* refers to the consistency between repeated measurements by the same observers using the same measurement devices with respect to the same units of measurement. Low intra-observer reliability is either a sign of a less than perfect task performance by the observer or a result of faulty, ambiguous or contradictory instructions with respect to the observation task.
- *Inter-observer reliability* refers to the agreement between measurements of different observers with respect to the same units of measurement. A lack of inter-observer reliability may indicate that the measurement procedure is too superficial (leaving room for additional interpretations of observers) or too complicated (encouraging personal heuristics) to overcome the subjective insights of observers. A mismatch between the phenomena to be observed and the concepts to be measured may also be at the heart of low inter-observer reliability. This type of mismatch will occur when classifications which were appropriate to the study of one specific country are transferred thoughtlessly to other countries.

Measures for the assessment of intra-observer reliability and inter-observer reliability are available for each level of measurement. Reliability measures start from ordinary measures of agreement between observers, but these measures have to be adjusted for agreement on the basis of mere chance. As an example, Scott's π (pi), a reliability measure for nominal variables, will be considered. In our example, Scott's π is equivalent to Cohen's (κ kappa), which is included in SPSS. As a starting point one can use the percentage of cases agreed upon as a first measure. If 100 cases are observed by two coders and identical observations show up for 98 cases then the agreement according to this intuitive measure would amount to 98 per cent. This intuitive measure does not take into account, however, that agreement may result from chance. If coders have two choices of code, then the probability of their agreeing by chance amounts to 50 per cent ($0.5 \times 0.5 + 0.5 \times 0.5$), at least when they make choices equally often. Things are even worse when they do not. Let us give a policy example. Suppose a new law is promulgated with rather vague criteria on special tax reliefs for firms stimulating environmental investments. Suppose that 100 firms demand special tax reliefs, but the civil servants enacting this law judge that only two firms deserve tax relief, because they know that enough money is available to grant two tax subsidies only. Agreement by chance as to whether the 100 firms should be granted tax relief now amounts to $0.98 \times 0.98 + 0.02 \times 0.02 = 0.96$. According to Scott's π the percentage of decisions agreed upon should be adjusted for agreement on the basis of mere chance:

$$\pi = \frac{\%\text{agreements} - \%\text{agreements expected}}{100\% - \%\text{agreements expected}}$$

Scott's π has a maximum value of 100 per cent. If the two civil servants pick out precisely the same two firms for tax relief, then this maximum will be reached. If they agree on 96 cases, but disagree precisely on the question of which two firms deserve tax relief, then Scott's π amounts to 0.49 only. This figure reflects common sense, since the civil servants disagree where the crucial question of which firms deserve tax relief is concerned, notwithstanding their amazing agreement that 98 out of 100 firms do not deserve tax relief.

When multiple indicators are available for one concept, the reliability of the measurements can be assessed by computing one way or another the agreement between these indicators. In the context of multiple-indicator research or 'scalability analysis' or 'item reliability research', which will be discussed in the next sections, the term scalability is used as a synonym for reliability.

4.4.1 Multiple indicators: the scalability (reliability) problem

Often a variety of related indicators of a concept can be imagined. One may choose one of these indicators as the best indicator on theoretical reasons. Often one will use *multiple indicators* to reconstruct a concept. In party manifesto research, for example, references to 'crime', negative references to 'social security' and references to 'economic orthodoxy' may be considered as signs of a rightist party ideology. In survey research, answers to a number of indicative questions will be combined to arrive at measurements of an abstract concept such as 'political efficacy'. To measure this single concept the survey respondent is asked whether he or she agrees or disagrees with a number of related statements such as 'Members of Parliament do not care about the opinions of people like me', 'Political parties are only interested in my vote and not in my opinions', 'People like me have absolutely no influence on governmental policy' and 'So many people vote in elections that my vote does not matter'. The operational definition of a concept should clarify whether a specific pattern is expected in the data with respect to the multiple indicators of the concept.

Multiple indicators may simply be intended as a *repeated measurements scale* of precisely the same concept. In survey research, several questions can be posed with respect to slightly different aspects of the concept (e.g. questions with respect to newspaper reading, watching television news and participating in political discussions to measure 'political interest'). In the case of repeated measurements one expects that each indicator gives rise to almost the same results.

Indicators may also build up to a *cumulative measurements scale*, however. The concept of 'political participation', for example, can be measured both with 'easy' indicators such as voting at elections (many citizens participate to this degree) and with 'difficult' indicators such as running for a political function (only a few citizens participate to this degree). Cumulative measurement scales resemble long jumping. An 'easy' indicator of one's jumping capacities is whether one can leap over a ditch 1 metre wide (many will pass this easy test),