RatSWD Working Paper Series

www.ratswd.de

RatSWD.
German Data Forum

245

Quality Standards for the Development, Application, and Evaluation of Measurement Instruments in Social Science Survey Research

Prepared and written by the Quality Standards Working Group

February 2015

Federal Ministry of Education and Research

Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the German section lists the complete contents of all issues in the series in chronological order.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD or of the Federal Ministry of Education and Research.

Quality Standards for the Development, Application, and Evaluation of Measurement Instruments in Social Science Survey Research

Prepared and written by the Quality Standards Working Group

Members of the working group:

- Beatrice Rammstedt (chairperson, GESIS – Leibniz Institute for the Social Sciences)
- Constanze Beierlein (GESIS – Leibniz Institute for the Social Sciences)
- Elmar Brähler (University of Leipzig)
- Michael Eid (Freie Universität Berlin)
- Johannes Hartig (German Institute for International Educational Research)
- Martin Kersting (University of Giessen)
- Stefan Liebig (University of Bielefeld)
- Josef Lukas (University of Halle)
- Anne-Kathrin Mayer (Leibniz Institute for Psychology Information)
- Natalja Menold (GESIS – Leibniz Institute for the Social Sciences)
- Jürgen Schupp (German Institute for Economic Research, Berlin)
- Erich Weichselgartner (Leibniz Institute for Psychology Information)

A German version of this paper is available as RatSWD Working Paper No. 230.

Table of Contents

1 Introduction

Since its establishment in 2004, the German Data Forum (RatSWD), which is funded by the German Federal Ministry of Science and Research (BMBF), has been advising the federal government and the governments of the federal states (*Länder*) on issues relating to the expansion and improvement of the research infrastructure for the empirical social, behavioral, and economic sciences (SBE). In late 2010, the RatSWD addressed the question of how the quality of survey instruments in the social and economic sciences, and especially in social and economic survey research, could be controlled and assured. At a meeting on November 9, 2012, the RatSWD therefore decided to set up a *Quality Assurance of Survey Instruments Working Group* under the leadership of Professor Beatrice Rammstedt. The establishment of the working group was prompted in particular by the desire to define quality standards in order to assure and optimize the quality of survey instruments. Hence, the working group made the formulation of these standards its primary objective. They are presented in this paper.

1.1 Relevance of the quality assurance of measurement instruments in social and economic survey research

Social and economic survey research focuses on social phenomena. Such phenomena include, for example, social inequality, migration, attitudes towards democracy, and quality of life. The investigation of these phenomena is prompted by the necessity to identify the mechanisms of perception and assessment of social phenomena in order, for example, to obtain data from which recommendations for the organization of social life can be derived. Both social and economic survey research are interdisciplinary research fields in which survey data are used to empirically answer research questions. Questionnaire responses, for example in a face-to-face or web-based survey, usually form the basis of these data. A questionnaire contains questions, or batteries of questions, aimed at eliciting a description of a single phenomenon or several different phenomena. In the present document, a question, or a set of questions, aimed at capturing a single, delimited phenomenon is referred to as a "measurement instrument." In other words, it is immaterial whether the phenomenon is measured with one or several questions.

Rather, the term “measurement instrument” is intended to make it clear that numerical, quantitative information on a specific phenomenon is obtained by means of data collection.[1] This numerical information (or quantitative data) is the basis of statements in survey research.

Many surveys are used not only as a basis for social stocktaking but also to identify areas where there is a need for political action and to develop political and social interventions. Particularly against the background of the considerable social policy importance of survey research, quality assurance and optimization of the measurement instruments applied is essential. The documentation of the process and results of quality evaluation facilitates the understanding of research results and/or the realization of secondary analyses of the existing data. In recent years, awareness of the importance of the quality of data has increased in the scientific community and in society in general. Recent reports on scientific misconduct have contributed to raising awareness. Not least for research-ethical and economic reasons, it is imperative to avoid using measurement instruments of unclear and in part questionable quality in large-scale and cost-intensive social science surveys.

In the present document, standards for the quality of measurement instruments are proposed on the basis of these considerations. The overall aim of these standards is to facilitate a comprehensive evaluation of the quality of measurement instruments and to contribute in the medium term to an increase in the quality of social and economic survey data. In principle, these standards are of benefit to all participants:

Developers of measurement instruments can use the standards as a guide to assuring the quality of their actions that covers all steps in the construction process—from the decision on the necessity to construct a new, or modify an existing, measurement instrument, through empirical quality evaluation, to adequate documentation.

(Potential) users of the measurement instruments or the data collected with these instruments receive information that facilitates the evaluation of the instruments at hand. For example, when planning their own data collections they can use the standards in order to choose from a

[1] In social science research, “measurement” is understood to mean a structurally accurate representation of the relationships between objects with the help of numbers (measurement values) (Schnell, Hill & Esser, 2011). “Structurally accurate” means here that the quantity of numbers (numerical relative) correctly represents a specific, previously defined relation between the objects (empirical relative). If, for example, the objects are ordered according to length, the numbers (smaller – bigger) should accurately represent the relation between the objects (shorter – longer) (cf. Schnell et al., p. 130).

number of different instruments the one that has proved its worth in the specific research and application context. Moreover, sound secondary analyses of the data collected using these measurement instruments are possible only on condition that the instruments (including the development procedure) have been thoroughly documented in accordance with the standards.

In addition to these participants, who are the main addressees of the present document, the *recipients of the research results obtained with the measurement instruments* (e.g. scientists and social and political decision-makers) also benefit. They can use the standards as an aid to appraising the informational value and relevance to action of the research results.

In survey methodology, the concept of *total survey error* (Groves et al., 2004) has emerged as a meaningful frame of reference for determining the quality characteristics of surveys. Basically, a distinction is made between the quality of the representation of the target population and the quality of the measurement (see Figure 1 in Chapter 2 below). The former refers to the quality of the sample, which is not the focus of the present document, and the question of whether, and if so to what extent, the survey results obtained from a sample can be generalized to the target population. The criterion of measurement quality, by contrast, refers to the measurement instruments deployed.

The methods of determining the quality of the measurement instruments within the total survey error framework originated mainly in psychological test theory (Lord & Novick, 1968; Lienert & Raatz, 1998; cf. also Groves et al., 2004; Schnell, Hill & Esser, 2011). In psychology, it is customary to examine the quality of tests and questionnaires and to publish information about their objectivity, reliability, and validity. In social and economic survey research, by contrast, such quality assurance rarely takes place. Hence, in surveys one finds hardly any information or documentation on what phenomena are measured with what questions and in what quality. This rather unsatisfactory situation may stem from the fact that—in contrast to psychology, where action-oriented guidelines for the assurance and evaluation of the quality of psychological tests and questionnaires exist (e.g. DIN, 2002; Häcker, Leutner & Amelang, 1998; Kersting, 2008)—such standards and guidelines are lacking in the social sciences.

The present quality standards are intended as a guide to assuring the quality of measurement instruments in the social sciences and in economics. They combine the perspective of total survey error from the field of survey methodology with that of psychological measurement from the field of test theory and diagnostics. The aim of the standards is to determine the quality of instruments from the procedural perspective of survey planning and implementation

and within the framework of the total survey error paradigm, and to present an action-oriented guideline on quality assurance and optimization. Specific assumptions about different types of error that may impair the quality of measurement instruments in survey research are derived from the total survey error concept. Based on these assumptions, quality standards are defined, which are intended as a guide for scientists and users to enable them to control, assure, and optimize the quality of the measurement instruments developed or applied by them.

1.2 What measurement instruments do the present standards cover?

Measurement instruments applied in surveys can take different forms. Some are realized in typical question form, for example the social-class question administered in the 2012 ALLBUS[2] ("Which class would you assign yourself to—the lower class, the working class, the middle class, the upper middle class, or the upper class?"), where the individual classes are read out as response categories by the interviewer. Other measurement instruments take the form of statements or items, for example the statement "Most politicians have actually no interest in the problems of ordinary people," with the response alternatives "agree" and "disagree" (ALLBUS 2012). Furthermore, surveys are sometimes used to collect test data, for example the basic skills assessed in the OECD's Programme for International Student Assessment (PISA; OECD, 2013a; Prenzel, Sälzer, Klieme & Köller, 2013) and the Programme for the Assessment of Adult Competencies (PIAAC; OECD, 2013b; Rammstedt, 2013), which are measured with the help of various tasks. Hence, a measurement instrument comprises one or several questions, items, or tasks including response categories and—as mentioned above—serves to capture a particular phenomenon.

For the most part, social and economic surveys measure attitudes (e.g. towards democracy, the European Union, or bringing up children), values (e.g. solidarity, tolerance, or hedonism), and behavior (e.g. religious practice, participation in elections, or health-related behavior), and facts (e.g. sex, age, education, and other socio-demographic information). Attitudes and values, in particular, are often complex and can rarely be captured with a single item or question. As phenomena (or constructs) that are not directly observable, they are usually

[2] German Social Survey (ALLBUS): http://www.gesis.org/en/allbus

measured with the help of several items or questions. On the basis of a test theory (classical or probabilistic), the responses obtained are converted into numerical values and compiled into composite indicators. In the Schwartz Value Survey (SVS), for example, Schwartz and Bilsky (1990) assume the existence of eleven universal basic human values. The importance attached to these values can differ across individuals and cultures. These basic values (e.g. self-direction, stimulation, and hedonism) are conceptualized in such a way that they each include a number of specific goals that are, in turn, measured with the help of several items.

As a rule, efforts are made to capture the phenomenon as efficiently as possible. When measuring directly observable facts or straightforward behavior—for example the number of children in order to determine the birth rate, or the number of times a respondent moved house within a certain period of time in order to determine mobility— it is often best to use a single question. Especially in the case of surveys of behaviors or facts that are typically of interest to researchers, standard instruments—such as the *Demographic Standards*[3]—have been developed and have established themselves in recent years, which are aimed at ensuring the comparability of official and academically driven surveys. In order to capture complex phenomena, they are usually measured with several questions. The data from these measurements are then aggregated into an index. One example is the construction of socio-economic status on the basis of educational, occupational, and income-related data (e.g. Ganzeboom, De Graaf, & Treiman's International Socio-Economic Index of Occupational Status, ISEI, 1992). The standards apply to individual measurement instruments—that is, questions, test tasks, or sets of items in a questionnaire, which are used to capture a distinct phenomenon. In other words, they do not serve to assess the survey questionnaire as a whole because, as mentioned above, such a questionnaire usually comprises several measurement instruments. When compiling a questionnaire composed of different measurement instruments, attention should also be paid to other aspects such as the order in which the instruments are presented, the use of headings to structure content, the length of the questionnaire, the appropriate placement of instructions, etc. However, these—very important—aspects of a questionnaire are not the subject of the present quality standards. For

[3] http://www.gesis.org/unser-angebot/studien-planen/demographische-standards (accessed on January 16, 2015).

information on the development and evaluation of questionnaires, the reader is referred to the relevant basic literature (Dillman, Smyth & Christian, 2009; Schnell, 2012).

Measurement instruments in survey research can be applied for different purposes (Hussy, Schreier & Echterhoff, 2010). The primary objective may be to *describe* and *quantify* a phenomenon, for example to determine electoral turnout or satisfaction with the work of the federal government, or to collect employment statistics. On the other hand, relations between different phenomena may be examined—for instance between unemployment and juvenile delinquency—or differences in certain phenomena across groups, for example gender differences in relation to income. Moreover, a measurement instrument may be applied in order to *explain* phenomena with the help of theories. This entails, first, formulating hypotheses about the causes of certain developments. The empirical validity of these hypotheses is then tested with the help of the measurement instrument. The results of the measurement may, in turn, necessitate the modification and amendment of the original theoretical assumptions and/or the concept of the phenomena. Similarly, *predictions* or prognoses about the way certain phenomena will develop in the future can be derived from the data. This, too, is done on the basis of explanations. And last, but not least, the aim may also be to investigate the consequences of *change*. This is the case, for example, with evaluation studies that accompany intervention, rehabilitation, or preventive measures in order to determine their effectiveness and impact.

The standards formulated in the present paper can be applied to the measurement instruments in question irrespective of the objective pursued when deploying the instruments, the research design, and the primary academic discipline in question.

1.3 How the quality standards should be used

The present quality standards reflect the minimum requirements that measurement instruments and their documentation must fulfill in order to enable their quality to be assessed. Specific guiding questions are formulated for each superordinate quality standard. Both researchers and practitioners can use these guiding questions to systematically examine which aspects of quality assurance were considered when developing and applying the instrument.

To justify these quality standards, background information from social science methodology and psychological test theory is used, and reference is made to empirical methods of, and

approaches to, quality assurance. However, as it is not possible to provide a detailed description of the individual methods, the relevant literature mentioned in each case should be consulted if necessary.

When using the standards, it is important to avoid viewing them in a mechanical and isolated way. Rather, the objectives and distinctive characteristics of the study in which the measurement instrument in question is being applied should be taken into account.

And finally, it should be noted that, in relation to the aforementioned objectives of social and economic survey research, the quality of the instruments used is an essential—but not the only—prerequisite to a study's informational value. Rather, attention must also be paid to the quality of the research design and the underlying theory, and to the adequate representation of the target population of interest.

1.4 Organization of this document

The present paper is divided into three chapters. In Chapter 2, which follows this introductory chapter, quality concepts of social and economic measurement instruments are defined within the framework of the total survey error paradigm. In the initial section of that chapter, central aspects of the total survey error approach are presented, from which the quality assurance steps that serve to structure the quality standards presented in Chapter 3 are then derived. Because a well-founded quality evaluation is made possible only by systematic and thorough documentation, the present quality standards pay particular attention to the documentation of the measurement instruments, which includes the theoretical and empirical arguments designed to justify their construction and application.

2 Quality characteristics of social science measurement instruments from a total survey error perspective

In this chapter, the total survey error approach is presented. From this presentation, which is based mainly on Groves et al. (2004) and Groves and Lyberg (2010), we will derive steps for the quality assurance of social science measurement instruments.

The total survey error approach takes as its starting point typical steps in the process of survey planning and implementation, and assumes a number of different errors that can occur during this process and can impair the quality of the collected data. These errors may be due to the

faulty measurement of a phenomenon of interest or the lack of representativeness of the sample. Measurement errors, or errors relating to the representation of the target population in the sample, may occur at different stages of the survey planning and implementation process. Figure 1 illustrates the various steps in this process (represented as rectangles). The errors that may occur during the process—and therefore the quality concepts to be ensured—are represented by ovals.

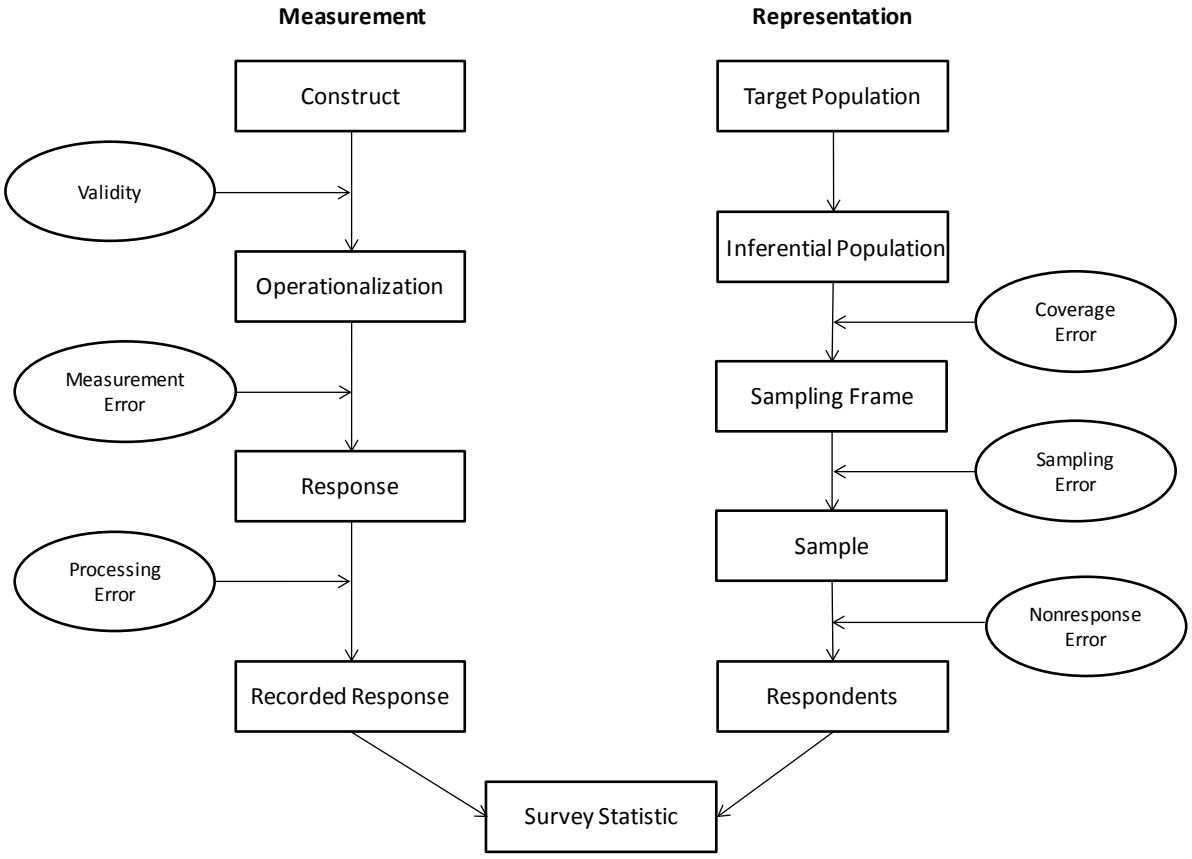

Figure 1. Total survey error components based on Groves and Lyberg (2010). Our translation.

The present document concentrates exclusively on the quality assurance of the measurement instruments. Hence, in what follows, only the *measurement* dimension (left side of Figure 1) is addressed. In international and national survey research, detailed and established standards already exist for the representational dimension. References are provided in the appendix to this document.

In the first step of the *measurement* process (cf. Fig. 1), the research question is formulated and the phenomena to be investigated are specified. For example, the research question might focus on how conservative CDU (Christian Democratic Union of Germany) voters are; how strongly Germans support the idea of democracy; or how environmentally conscious the Germans are. As a rule, the phenomena to be investigated are not directly observable. Initially, therefore, their designation has the status of a (general) concept. With regard to their measurement, such concepts are called constructs (e.g. "conservatism," "pro-democratic attitude," or "environmental consciousness"). At first, therefore, constructs are ideas that designate the phenomena to be investigated. The specification of the *construct* to be investigated (see Fig. 1) concludes the first step of the measurement.

In the second step, the constructs specified are defined more exactly and delimited, and the indicators (questions or items, observable reactions of the respondents) to be used to capture the construct are specified. In Figure 1, this step is referred to as *operationalization.*[4] This step entails defining how phenomena that cannot be directly observed should be measured within the framework of the survey. Therefore, "operationalization" refers to a set of adequately precise instructions on how the respondents, as bearers of characteristics that distinguish the construct, can be described with the help of the *survey data* (Figure 1) (cf. Diekmann, 2007, Schnell et al., 2011). Hence, the aim of the operationalization step is to ensure that the construct is measured as accurately as possible.

In the context of the total survey error framework, *validity* refers to the correspondence between the construct of interest and its operationalization. From this perspective, validity is achieved if it is possible to make inferences about this construct on the basis of its operationalization. For example, it would have to be empirically proved that the item "I am proud to be German" is an indicator of the construct "national pride" and that it does not measure another characteristic, such as nationalism. Conversely, validity is reduced if by virtue of the operationalization the items or questions capture the construct only partially or not at all. This correspondence between the operationalization and the construct must be investigated and proved.

---

[4] Groves et al. (2004) and Groves and Lyberg (2010) use the term "measurement" to refer both to this step in the survey implementation process and to the measurement process as a whole. We use the term "operationalization" because it seems to be a more apt designation of the corresponding step—development of the measures (questions, items, tasks) to capture the construct of interest.

Although a general definition of validity in psychological test theory reflects a broader understanding of the concept, it corresponds to a large extent to the validity concept presented here. According to this concept, validity is given when intended interpretations of the measurement results (in the present case, based on the operationalization) can be justified by theoretical and/or empirical evidence (Kane, 2013). Operationalization forms the basis of the remaining steps in the measurement process. By conducting the measurement as accurately as possible in each step of the process, and by empirically investigating the respective error influences, it can be ensured that, when using survey data, broadly accurate inferences can be made about the construct under investigation.

In the third step of the survey planning and implementation process, the questions developed in Step 2 are answered by the survey respondents (*Response* in Fig. 1). Response errors may occur, in particular, when questions have been worded in such a way that they are difficult to understand; when they are too abstract or not explicit enough; or when they are difficult to answer for some other reason. Moreover, respondents may differ in terms of their cognitive abilities, their motivation, or with regard to other characteristics that are irrelevant for the measurement but that may distort the measurement results. Survey mode is a further source of error: measurement instruments can be administered via different media (e.g. paper-and-pencil, telephone, face-to-face by the interviewer, electronically on a screen), and the survey mode per se, or the use of different modes within the one study, may influence response behavior because, for example, participants may have different levels of experience with the various media, or the various modes may be associated with different degrees of anonymity.

Measurement error (Fig. 1) is defined as the difference between the response expected on the basis of the operationalization of the construct (i.e. the questions or items formulated in the questionnaire) and the response that has been expressed verbally or has arisen in the form of a mental representation. The errors or discrepancies that arise while the question is being answered may be either systematic or random. Systematic errors—also known as bias—occur when the responses of all the respondents, or of a defined sub-group of respondents, are similarly inaccurate. One type of bias is social desirability. It occurs, for example, when respondents are asked whether they use drugs, and they consistently underreport consumption for reasons of social desirability. Random measurement errors, by contrast, occur when responses differ inconsistently across respondents because of various effects of the survey situation, for instance. For example, responses may be inaccurate if the interview takes place in noisy, distracting surroundings, which may disturb respondents in different ways. Random

measurement error corresponds to the definition of measurement error in classical test theory and will be investigated in connection with the determination of reliability (see Section 3.4).

In the final step of the survey implementation, the response is recorded, either by the respondent him- or herself or by the interviewer. This yields the *recorded response* (Fig. 1), or in the terminology of Groves and Lyberg (2010), the *edited response*. As a result of editing during the recording process, the recorded response may deviate from the reported or intended response. For example, a respondent may report that his highest level of educational achievement is *Abitur* (the general higher education entrance qualification in Germany) but the interviewer may mistakenly mix up the codes and record the code for *Fachabitur*, a subject-restricted higher education entrance qualification. The response categories in self-administered questionnaires may not be self-explanatory, mutually exclusive, or collectively exhaustive, which can also lead to the incorrect recording of the response. Further deviations can arise as a result of data cleaning; corrections resulting from plausibility checks (e.g. when a 12-year-old reports that he has a university degree or a male reports a pregnancy); summarizing values; or imputing missing values. *Processing error*, as a quality concept (Fig. 1), is defined as a systematic discrepancy between (a) a person's intended or reported response and (b) the recorded response, and, afterwards during data processing, the modified numerical value (see *Survey Statistic* in Fig. 1).

In sum, from the total survey error perspective, high survey measurement quality is assured if the (survey) data enable the most accurate inferences possible to be made about the true values of the construct of interest among the persons studied (see Fig. 1).

The following aspects of measurement quality assurance can be derived from the total survey error framework:

1) Definition and operationalization of the construct in the course of the development of the instrument
2) Determination of validity
3) Identification of measurement error, divided into:
    a. Identification and minimization of systematic effects that result from specific characteristics of the survey situation, for example the mode of data collection (method effects)
    b. Determination and minimization of the extent of random measurement error by means of the determination of reliability.

4) Determination and minimization of processing errors that occur when recording and processing the data.

In the following chapter, quality standards will be formulated on the basis of the aspects of quality assurance defined here.

3 Quality standards

The six quality standards presented in this chapter (for an overview, see Box 1) were developed on the basis of established methods and best practices in survey research. They are supplemented by procedures for quality assurance and optimization.

Box 1: Overview of the six quality standards

Standard 1: Instrument development

The object and purpose of the development of the instrument should be stated and the development methodology should be documented.

Standard 2: Validity

The interpretation of the values measured with the instrument should be explicitly formulated and evidence in support of this interpretation should be provided.

Standard 3: Minimization of method effects

When developing the instrument, possible method biases that systematically influence the response behavior of survey participants should be addressed and investigated, and the results of this investigation should be documented.

Standard 4: Reliability

Reliability should be assessed, the choice of assessment method should be justified, and the reliability parameters should be documented and evaluated.

Standard 5: Minimization of processing error

Standardized instructions for data collection and analysis should be provided and justified.

Standard 6: Other quality characteristics

Information should be provided about the economic efficiency and reasonableness of the instrument and the up-to-dateness of the psychometric parameters.

3.1 Instrument development

Background

The first step toward improving the quality of instruments used in survey research involves extensively and precisely documenting the methodology used to develop the instrument. The individual development steps and the decisions taken at each step should be carefully documented and justified, and the purpose and the target population for which the instrument is being developed and the questions with regard to which it is being tested should be outlined. Moreover, the survey modes for which the procedure is suitable or has been tested should be described. This information is of great importance, not only for the evaluation of the results achieved with a measurement instrument but also when re-using the measurement instrument or conducting secondary analyses of the original data.

As a rule, for the sake of better comparability of findings and for economic reasons, existing instruments are used where possible. Such tried-and-tested and well-established instruments can be found, for example in scientific databases[5] and in digital archives of social and psychological survey procedures.[6] When re-using an existing measurement instrument, its quality can be evaluated with the help of the present standards. If existing instruments are used, their origin and any modifications that may have been undertaken should be carefully documented.

If the existing instruments are available only in another language and they have to be translated, the available standards or guidelines for translating survey instruments should be

[5] For example PSYNDEX Tests: http://www.zpid.de/index.php?wahl=PSYNDEX&uwahl=Tests (accessed on January 15, 2015).

[6] For example ZIS, GESIS' compilation of social science items and scales (http://www.gesis.org/unser-angebot/daten-erheben/zis/; accessed on January 15, 2015) and the ZPID's Electronic Test Archive (http://www.zpid.de/index.php?wahl=products&uwahl=frei&uuwahl=userlog: accessed on January 15, 2015).

observed.[7] As a general rule, if substantial modifications are carried out, or if an instrument is translated, the quality of the modified instrument should be empirically proved once again (also theoretically, if necessary).

If no instruments are available for the object of research, or if existing instruments cannot be used for methodological or content-related reasons, new instruments must be developed. In this case, the necessity of developing a new instrument must be established. A new development is indicated if no suitable instruments are available, or if existing instruments are not suitable for the study context in question because, for example, they address other target populations, comprise too many questions, do not fulfill certain quality requirements, or because the questions do not capture topical issues.

In the course of the development of a new measurement instrument, the phenomenon of interest must first be described and defined (definition of the construct). Where possible, the definition of the construct should be derived from a substantive social or economic theory (core theory). For example, the construct "basic human values" (Schwartz Value Survey, Schwartz & Bilsky, 1990) is based on theories about formal characteristics of the values as abstract, substantively distinct motives. The content of the values are derived from motivation theories.

In addition to the core theory, an auxiliary theory (cf. Schnell et al., 2011) is needed that indicates how the abstract concepts can be translated into observable phenomena. This auxiliary theory contains statements about the operationalization of the construct, which can, in turn, be empirically evaluated. If the researcher's knowledge about the phenomenon of interest is not detailed enough to derive substantiated statements about the operationalization of the phenomenon of interest from it, this knowledge can be gained, for example, within the framework of preliminary qualitative studies such as observation, qualitative interviews, etc. Taking the auxiliary theory into account, a suitable measurement model (within the framework of classical or probabilistic test theory) is decided on at the beginning of the instrument development phase and an item-development or scaling method is chosen.

[7] See the Translation Guidelines for Best Practices in Cross-Cultural Surveys, which we will not treat in the remainder of this document (http://ccsg.isr.umich.edu/; accessed on January 15, 2015).

Construction principles for the development of the items or questions of the measurement instrument should be defined and justified. A widely used procedure consists of finding, where possible, several indicators for every dimension of the measurement instrument. For example, ethnic stereotypes and discriminatory behavior are deemed to be dimensions of xenophobia. A set of questions—the so-called item pool—is formed on the basis of these dimensions. Researchers can choose from a range of methods in order to select the most suitable questions or items from this item pool. One approach is to pretest the items with a sample of the target population. Item analyses are then carried out and items are selected on the basis of their item parameters (difficulty, item-total correlation). Alternatively, the suitability of the choice of items can be assessed discursively by selected experts on the basis of the following questions: Do the items adequately represent the range of possible items? Are the relationships among the items appropriate? Is the content of some items irrelevant? The expert judgements can be collected in individual interviews or in focus groups. Combinations of the said methods may also be used.

In addition to justifying the development and selection of the questions or items, the choice of response categories should be justified. (Why, for example, was a five-point rating scale used? Why were certain response options selected for the measurement instrument?) Care should be taken to ensure that the response categories are appropriate for the target population and that they are unequivocal, mutually exclusive, and collectively exhaustive (Dillman et al., 2009; Krosnick & Fabrigar, 1997).

The developers of the measurement instrument should furnish empirical evidence of its quality. When developing a new measurement instrument, empirical studies (such as pretests, for example) to develop and validate the instrument should be carefully planned. Ideally, studies to validate the interpretation of measures should not be conducted with the same sample that was used to develop the instrument (development sample), but rather with one or more separate samples (validation sample(s)). If validation studies cannot be carried out with new samples because of a lack of resources, this should be stated in the documentation of the measurement instrument.

Standard 1

The object and purpose of the development of the instrument should be stated and the development methodology should be documented.

Guiding questions for the development of a measurement instrument

1. Is the objective of the development of the measurement instrument stated? (Basic research/applied research; to describe, explain, or predict a phenomenon, or to measure changes in the phenomenon over time.)
2. Is (Are) the target population(s) for which the measurement instrument was developed stated?
3. Is the methodology used to develop the measurement instrument described? Is information provided regarding the sample of persons to whom the measurement instrument was administered and the mode of data collection?
4. Is a definition given of the phenomenon to be measured, or the construct? Has this phenomenon been delimited from adjacent/similar phenomena?
5. Is the core theory specified on the basis of which the phenomenon was defined and its delimitation undertaken? If so, are relevant studies that support the statements of the core theory mentioned? Is an overview provided of the results of these studies that are of relevance to the definition of the constructs?
6. Is mention made of the test theory that was chosen? Is the choice of test theory justified with reference to the auxiliary theory? If the measurement model is described, are existing studies to test the measurement model mentioned and are their key findings described?
7. Is justification of the necessity to develop a new instrument provided?
8. If an existing instrument has been modified: Is the original source stated? Is an exhaustive list provided of the modifications that were carried out? Is justification of the modifications provided?
9. If an existing measurement instrument has been translated: Is a description of the translation procedure provided? What translation standards were used?
10. Were the items generated according to stipulated rules? Is justification of the chosen scaling method (e.g. Likert-, Thurstone-, Guttman-, or Rasch scales) provided or can the choice be understood on the basis of the information provided on the auxiliary theory? Are the rules for the systematic selection of the items stated and is justification of the rules provided?

11. Were expert judgments taken into account when selecting the items? Are details of the discipline-specific qualifications and experience of these experts given in the documentation of the measurement instrument? Are the assessments of the experts outlined and is the degree of consensus among the experts stated?
12. Are the response categories described? Is justification of their selection provided? If so, are relevant studies and their findings stated?
13. Were validation studies conducted when developing the measurement instrument? Are the results of these studies presented?
14. Were the development and validation of the instrument conducted with different samples?
15. Are the scientific purposes for which the measurement instrument can be used stated? If so, is the interpretability of the results restricted as a result of limitations within the framework of the operationalization, the development of the measurement instrument, and the results of the validation?

3.2 Validity

Background

As a quality criterion, validity refers to the *adequacy and appropriateness of the interpretations and uses of the values measured with* a specific measurement instrument. From an overall perspective, validity can be regarded as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores" (Messick, 1989, p. 13 [emphasis in original], cited in Hartig et al., 2012, p. 114). Older approaches to this quality criterion regard it as an attribute of a measurement method. Contemporary approaches, by contrast, do not assume that validity is an attribute of a measurement method per se but rather that specific interpretations and uses of the resulting values may be valid. In the course of the development of the validity concept, various aspects of validity have been differentiated—in particular, *criterion validity*, *content validity* and *construct validity* (for the history of the concept, see Hartig et al., 2012; Kane, 2001, 2013). While these aspects were regarded for a time as separate alternatives to validity, *construct validity* (Cronbach & Meehl, 1955) gained acceptance as an integrative, overarching concept. More recent approaches no longer

differentiate “types of validity.” Rather, they focus on the quality of the *arguments* advanced to justify specific interpretations of measured values (*Validity as evaluation argument*, Cronbach, 1988; *An argument-based approach to validation*, Kane, 1992).

Following Kane (2001), interpretation of measured values may refer, for example, to

1. the *evaluation* of these values
2. the *generalization* of the values beyond the content of the concrete questions or items used
3. *extrapolation* in the sense of inferences about external phenomena of interest
4. the (causal) *description* of the data with reference to an underlying theoretical construct, and
5. *decision-making* on the basis of the data.

In the context of the validation of inferences from measured values, Kane (2013) uses the term *interpretation/use argument* (or IUA). The IUA specifies the way in which these data should be interpreted and used and the arguments that can be put forward to justify this particular interpretation. Following Kane (ibid., p. 9), a proposed interpretation of a measured value would be considered valid if the argument was coherent and complete and the inferences and assumptions were either supported by strong evidence or were highly plausible a priori.

Hence, within the framework of survey research, the interpretations and uses of the values collected with a particular measurement instrument that are of greatest relevance to the objective of the study should first be *specified*. Building on this, empirical and/or theoretical *evidence* should be provided to support these interpretations. As a quality criterion, validity is characterized by the fact that its evaluation is not a routine procedure in which the same methods can always be used. Rather, the *validation* of the interpretation of measured values constitutes theory-led research with which specific interpretations can be supported or falsified (Hartig et al., 2012; Kane, 2001).

From the total survey error perspective, the explanatory, theory-based interpretation of measured values as indicators of underlying constructs (see point 4 above) is of central relevance to survey research in the social sciences. This interpretation is closely related to Cronbach and Mehl’s conceptualization of *construct validity* (1955). Theory-based interpretations of values can be validated by testing theory-derived predictions about the correlations between these values and those of other variables. Numerous experimental and correlative research designs can be used for this purpose. An ideal-typical example is the theory-derived *nomological network* proposed by Cronbach and Mehl (1955), which

describes relationships among constructs. According to the social science literature, such descriptions can be found in the form of statements in auxiliary theories (Schnell et al., 2011; Krebs & Menold, 2014). Hence, it is possible to assess whether the measured values, which are supposed to represent a particular construct, empirically display the expected correlations with the values measured for the other constructs.

Even though the focus of social and economic survey research is on construct-related interpretation, other interpretations of values may also be important and therefore require validation. *Generalization* inferences from values presuppose that the questions or items employed adequately represent a broad content domain (and therefore all conceivable questions or items). The fulfillment of this requirement, which in the earlier literature is often associated with the term *content validity*, is frequently assessed by having the content of items/questions reviewed by subject-matter experts. This is the case, for example, with assessments of student performance, which are supposed to adequately reflect curriculum content or specific educational standards (e.g. Harsch, Pant & Köller, 2010; Pant, Stanat, Pöhlmann & Böhme, 2013). Demographic standards are an example from the field of social research. In Germany, these standards are developed by a group of experts from social research methodology, private sector social- and market research, and the Federal Statistical Office. Although this approach is often the only way to obtain empirical evidence for generalization inferences, judgment-based validity studies are often viewed critically (Guion, 1977; Kane, 2001; cf. also Krebs & Menold, 2014).

Extrapolation inferences from measured values are also of importance for social and economic survey research—namely in cases where the aim of the survey is to make predictions. Extrapolation from values to phenomena of interest outside the context of a specific measurement instrument is associated with one of the oldest validity concepts, namely *criterion validity* (cf. also Rammstedt, 2010). Empirical evidence of the validity of such extrapolation inferences from measurement values is often examined by measuring the correlations between these values and the relevant external criteria. For example, one can measure the correlation between the results of a performance assessment and academic success at school, or one can use external criteria from official statistics to examine the quality of predictions made on the basis of questionnaire data. By the same token, a specific behavior (e.g. voting for a conservative party) can be used as a criterion for political attitudes collected with a measurement instrument. In each case, there must be a meaningful relationship between the selected criteria and the intended interpretation of the values, and the

criteria themselves must be measurable in the specific context in a reliable and valid way. Box 2 provides an overview of a selection of empirical methods with which evidence supporting specific interpretations of values can be generated.

Box 2: Methods of generating evidence in support of specific interpretations of measured values

Construct-related explanation inferences

- **Examination of correlative relationships** with instruments that measure the **same or theoretically related constructs** and that should therefore yield values that correlate strongly with the measured values (*convergent validity*). For example, the measurement results yielded by Schwartz's Political Alienation Scale (Schwartz, 1973) correlate positively and substantially with the results obtained with scales for the measurement of political inefficacy (overview in Robinson, Shaver & Wrightsman, 1999).
- **Examination of correlative relationships** with measurement instruments **that measure constructs from which the target construct is supposed to be distinct** *(discriminant validity*). For example, the correlation between the results of the measurement of conservatism with an appropriate instrument and the results obtained using instruments designed to measure liberalism should be low or negative (e.g. Robinson, Shaver & Wrightsman, 1999).
- Use of the **multitrait-multimethod approach** (MTMM) (Campbell & Fiske, 1959) to examine the consistency of different methods of measuring the same construct (e.g. paper-based and web-based surveys, self-report measures, and interviewer assessment; dissimilar measurement instruments for the measurement of similar constructs). This approach entails examining and comparing correlations between measures that assess similar and dissimilar traits with similar and dissimilar methods (measurement instruments).
- Examination of the **dimensional structure of a measurement instrument** derived from a theory using confirmatory factor analyses or item response theory methods (e.g. the two pre-defined facets of national pride—nationalism and patriotism—are replicated in a two-factor structure).

Generalization inferences

- **Expert judgments** of the extent to which the questions or items used represent the construct or subject area of interest.
- **Analyses of documents** (e.g. curricula) in order to determine whether a content domain is covered by the questions or items used.

Extrapolation inferences

- **Cross-sectional examination of the correlations** between values and external criteria (e.g. respondents' self-reported healthy eating habits match other indicators of health measured at the same point in time; or whether a right-left scale distinguishes between those who vote for right-wing and left-wing parties; Groves et al., 2004; *concurrent validity*).
- **Prediction of forthcoming events** using the measured values (e.g. party preference measured at one point in time matches voter behavior at a later point in time; *predictive validity*).
- **Retrospective examination of the correlations** of measured values with actual events that occurred at a previous point in time (e.g. voter behavior at the last general election; *retrospective validity*).
- **Examination of the stability of correlations** between measured values and external criteria in various sub-populations (e.g. is the correlation between academic success and motivation to perform equally strong for boys and girls; *differential validity*).

In sum, the choice of validation method(s) should be justified in detail by the interpretation of the measured values that is most important for a specific study. If, for example, the focus is on the description of a population in relation to a certain construct, evidence must be provided that the measured values can in fact be interpreted as indicators of this construct. If the aim is to make predictions, it must be proved that the measured values used for this purpose are indeed associated with the phenomenon to be predicted.

Standard 2

The central interpretation of the measured values collected with the measurement instrument should be explicitly formulated, and evidence in support of this interpretation should be provided.

Guiding questions concerning validity

1. Does the accompanying documentation specify the way in which the values measured with the instrument are to be interpreted and used?
2. Is the argument on which this interpretation is based presented clearly and understandably in the accompanying documentation?
3. Is empirical and/or theoretical evidence furnished for the assumptions on which the argument is based?
4. Were empirical findings that are used as evidence obtained in studies of the same target population as that to which the current proposed interpretation refers?
5. Do the authors provide a reasoned statement as to which assumptions in the validity argument are least supported by the evidence?

3.3 Minimization of method effects

Background

This section deals with the estimation and minimization of systematic measurement errors against the background of the total survey error paradigm. Systematic measurement errors arise when the responses of all survey participants, or a subgroup thereof, systematically deviate from the response that one would expect as the "ideal response."

Basically, different elements of the realization of the measurement instrument, for example the wording of the items or the choice of response formats, can lead to undesirable systematic effects (e.g. Dillman et al., 2009; Faulbaum, Prüfer & Rexroth, 2009; Krosnick & Presser, 2010, Tourangeau, Rips & Rasinski, 2000). Systematic effects also include so-called context effects, which may be caused by the order of the questions in the questionnaire, or by other elements of the questionnaire, such as pictures (Sudman, Bradburn, & Schwarz, 1996). Moreover, the way the items and questions are formulated may not appeal to certain sub-groups of the target population, for example because words are used that violate the precept of unprejudiced research.

Different modes of data collection (face-to-face, telephone, mail (post), web-based) are also a source of systematic deviations. Instructions and guidance for the interviewers and the respondents serve to minimize undesirable effects. Furthermore, attention should be paid to

the layout of the measurement instrument. The instructions and the layout should be optimally suited to the chosen mode of data collection (for requirements and implementation examples, see Schnell, 2012). In so-called mixed-mode surveys, care should be taken to ensure uniform presentation in the different modes. However, this requirement cannot always be fulfilled. For example, depending on the mode, specific modifications have to be carried out in order to guarantee that the question-and-answer process within the mode is optimal (Dillman et al., 2009).

Systematic errors can be investigated and uncovered with the help of cognitive pretests or split-ballot experiments, which enable effects of question wording and response formats, and also possible context effects, to be studied. In cognitive pretesting, qualitative methods are used to examine the way in which respondents understand and answer the items or questions. This approach enables the researcher to determine whether distortions arise because of comprehension problems or during the response process. Split-ballot experiments are randomized experiments in which different variants of a measurement instrument are assigned to different experimental groups, and the resulting empirical values are compared in such a way that possible biases are revealed in the form of differences in these values across the experimental groups.

In empirical social research, cognitive pretesting is now a well-established method of testing questionnaires (Faulbaum et al., 2009; Schuman & Presser, 1981). Hence, the application of a cognitive pretest constitutes a minimum standard. An additional test of the measurement instrument with the help of split-ballot experiments can be conducted within the framework of standard quantitative pretests. Resources permitting, it can also be carried out in addition to cognitive pretests.

Standard 3

When developing the instrument, possible method biases that systematically influence the response behavior of survey participants should be addressed and investigated, and the results of this investigation should be documented.

Guiding questions to minimize method bias

1. Does the measurement instrument documentation indicate how the optimal formulation of items/questions was ensured?

2. Are mode-specific instructions provided for the measurement instrument? Is information provided on the layout of the instrument in a specific data collection mode?
3. Are details provided of results of studies that justify the necessity for mode-specific differences in the layout of the measurement instrument or in the accompanying instructions?
4. Was the understandability of the items tested, for example by means of cognitive pretests, when the instrument was being developed? Are the methods and results of these tests and the resulting modifications to the instrument documented and justified?
5. Were split-ballot experiments conducted when the instrument was being developed? Are the methods and results of these experiments and the resulting modifications to the instrument documented and justified?

3.4 Reliability

Background

Unsystematic variance across measurements of a characteristic is an indication of the inaccuracy of the measurement and, thus, of the poor reliability of the measurement instrument. Hence, reliability describes the precision with which an instrument measures an attribute. In classical test theory, reliability is defined as the ratio of true-value variance to observed-value variance. When reliability is high, results obtained from repeated measures hardly differ at all. Therefore, if reliability was perfect, the correlation between repeated measures of the same characteristic would be 1. If, by contrast, the instrument was not reliable at all, the correlation would be zero.

Reliability is a measure that indicates the extent to which observable differences between persons can be attributed to true (i.e. measurement-error-free) differences (Eid & Schmidt, 2014). Therefore, if the focus of a study is on the investigation of inter-individual differences, correspondingly high demands will be made on the reliability of the measurement instrument. The reliability of a measurement instrument is also an essential precondition for its validity (see Section 3.2) because the interpretation of values depends on their accuracy (Kane, 2013).

The assessment of reliability presupposes repeated measurements of the characteristic in question. Several different methods can be used (Eid & Schmidt, 2014; Groves et al., 2004; Rammstedt, 2010; Schnell et al., 2011):

- The test-retest method (or repeated interviews with the same respondents): The measurement instrument is administered twice to the same sample under comparable conditions and after an adequate time interval. This presupposes that the characteristic to be measured is stable because otherwise the unsystematic "unreliability" of the measure is confounded with the systematic instability of the characteristic.
- The parallel forms method: Here, the characteristic is measured with two different, yet parallel (or equivalent) instruments.
- The split-half method (only in the case of measurement instruments that comprise several items): Here, the items of the measurement instrument are divided into two halves and the correlation between these halves is computed.
- Internal consistency (only in the case of measurement instruments that comprise several items). The analysis of internal consistency is a refinement of the split-half method because not only is the instrument split into two halves, but also the intercorrelation of all the items in the measurement instrument is assessed.

The diversity of these methods takes account of the diversity of the attributes to be measured and the differences in the realization of the measurement instruments. On the basis of these procedures, reliability can be assessed with different coefficients. These coefficients presuppose that the repeated measures fulfill certain requirements. These requirements are formulated in measurement models and can be verified in empirical tests (Eid & Schmidt, 2014). The choice of measurement model depends on the scale level of the items and on that of the construct to be measured (latent variable). Item response theory models (e.g. the latent trait model), latent class models, or confirmatory factor analysis models can be used (Eid & Schmidt, 2014; Rost, 2004). If, for example, the repeated measures are at least essentially τ-parallel (i.e. they measure a single facet or dimension of a characteristic with the same discrimination and error variance), the test-retest-, parallel-test-, and split-half reliability can be assessed on the basis of the correlation between both measures, alone. The test-retest- and the parallel forms method are suitable for assessing the reliability of measurements of a characteristic with just one question or item. The limitation of the test-retest method lies in the fact that it can be used only in the case of characteristics that are stable over time, while the difficulty posed by the parallel forms method is that an additional question or item must be constructed that enables an equivalent measurement to be conducted. As a minimum standard, reliability should be verified and reported using at least one method and the preconditions for the application of the chosen method should be fulfilled. More detailed presentations of

reliability assessment methods can be found, for example, in Raykov & Marcoulides (2011) and Eid and Schmidt (2014).

Standard 4

Reliability should be assessed, the choice of assessment method should be justified, and the reliability coefficients should be documented and evaluated.

Guiding questions concerning reliability

1. Have reliability values been reported for the measurement instrument?
2. Is justification provided for the choice of reliability assessment method?
3. To what extent is it certain that the preconditions for the application of a chosen reliability assessment method are fulfilled?
4. Is a description provided of the reliability assessment study and of the sample that was used to assess reliability?
5. Are the results of the reliability assessment study evaluated and are the implications of these results for the use of the measurement instrument and the interpretation of the values obtained addressed? In the case of low reliability, are limitations on the use of the measurement instrument stated?

3.5 Processing errors

Background

Processing errors that can have a detrimental effect on the quality of a measurement instrument can arise either during data collection or when transferring and processing the data. Generally, a distinction is made between interviewer effects, effects of the data collection situation due to the mode of collection chosen, coding errors, and errors that occur when cleaning and weighting the data (Groves et al., 2004). In this context in psychology, the term used is “objectivity” rather than “processing error.”

A reduction of processing errors, or an increase in objectivity, is achieved by means of the comprehensive standardization of the measurement instrument (including the standardization

of instructions for data collection, analysis and interpretation) and the conditions under which the data are collected. Therefore, when conducting interviews care should be taken to ensure that the survey situation varies as little as possible across respondents. When using measurement instruments with multiple indicators, data analysis instructions should be provided that describe in detail the way in which the various indicators should be aggregated into an index. When carrying out validation and reliability assessment, the developers of the measurement instrument should ensure that the prerequisites for the formation of indices are fulfilled. For example, when aggregating the individual indicators into a scale value (mean or sum), the precondition of one-dimensionality should be fulfilled (Bühner, 2011; Schnell et al., 2011).

Standard 5

Standardized instructions for data collection and analysis should be provided and justified.

Guiding questions to check for processing errors

1. Are *mode-specific* instructions provided for implementing data collection with the measurement instrument?
2. Are the instructions for data collection with the measurement instrument formulated in such a way that different people are in a position to apply the instrument in a comparable way with the help of the instructions alone?
3. Are rules provided for dealing with questions that are expected to be asked by respondents or interviewers?
4. Where applicable, is a list of specific technical specifications/requirements provided (in the case of computer assistance: computer assisted personal interviewing [CAPI], computer assisted telephone interviewing [CATI], computer assisted self-interviewing [CASI], computer assisted web interviewing [CAWI])?
5. Where applicable, are the equipment-related and mode-specific problems addressed that may occur during data collection, and are possible solutions given?
6. Are instructions for index formation provided in the case of measurement instruments comprising multiple indicators? Is the extent to which the preconditions for aggregating indicators or forming indices are fulfilled reported?

3.6 Other quality characteristics

Background

In addition to the quality criteria presented above, other quality characteristics—so-called secondary quality criteria, which are well-established in the field of psychology—are also of importance for instruments that are applied in social and economic surveys. These secondary quality characteristics do not focus on the quality of the measurement but rather on the requirements relating to the application of the instrument—namely, the efficiency, reasonableness, and up-to-dateness of the instrument[8].

The *economic efficiency* of a measurement instrument is determined on the basis of the time it takes to administer and its ease of handling. Hence, if one had a choice between two comparable instruments with a similar content domain and similar measurement quality, preference would be given to an instrument that contained less items and therefore had a shorter response time. In the context of social and economic survey research, the economic efficiency aspect is of central importance because the implementation of population surveys involves high costs and respondents have to invest a considerable amount of time in the interests of research. Where possible, therefore, several content domains must be surveyed within the very limited time frame. This calls for measurement instruments that guarantee high measurement quality despite a reduced number of items and questions.[9]

Reasonableness refers to the fact that the measurement instrument does not impose an unnecessary psychological or mental burden on the survey participants. To ensure reasonableness, designers of measurement instruments should be guided by established ethical standards in the social sciences.[10]

---

[8] A further secondary quality criterion is usefulness. It refers to the extent to which a practical need exists for such a measurement instrument. The usefulness of a new instrument is low if it measures content that can be measured with an existing survey instrument. This quality standard has already been introduced in the present document—namely in connection with the development of the instrument (see Standard 1, Guiding Question 7).

[9] In collaboration with the German Institute for Economic Research (DIW), standardized short scales for the measurement of psychological traits have been developed by GESIS – Leibniz Institute for the Social Sciences within the framework of a research project. These short scales (including documentation) are freely available to interested scientists and can be found at the following address: http://www.gesis.org/kurzskalen-psychologischer-merkmale/kurzskalen (Accessed on January 16, 2015).

[10] Code of ethics of the German Sociological Association (DGS) and the Berufsverband Deutscher Soziologen (Professional Association of German Sociologists; BDS): http://www.soziologie.de/index.php?id=19; ADM Codex: http://www.adm-ev.de/fileadmin/user_upload/PDFS/Erklaerung_2008.pdf (accessed on January 16,

The *up-to-dateness* of measurement instruments relates to the empirical results of the assurance of measurement quality. Details of the validity and of other aspects of the measurement quality of an instrument should be up-to-date or should be updated at regular intervals. As a rule, these intervals should not exceed eight years (cf. DIN 33430 in DIN 2002; Kersting, 2008).

Standard 6

Information should be provided about the economic efficiency and reasonableness of the instrument and the up-to-dateness of the psychometric parameters.

Guiding questions to check economic efficiency, reasonableness, and up-to-dateness

1. Is the time required to administer the instrument and analyze the data stated? To what extent does this expenditure of time bear a rational relationship to the purpose of the measurement?
2. Can data collection with the measurement instrument be deemed harmless to respondents? Does data collection with the measurement instrument impose any physical or mental burdens on the respondents?
3. Are details provided about when the last measurement quality studies were conducted?

2015). For international codes, see for example the AAPOR Code of Ethics: http://www.aapor.org/AAPORKentico/Standards-Ethics/AAPOR-Code-of-Ethics.aspx (accessed on January, 16, 2015).

4 References

Bühner, M. (2011). Einführung in die Test- und Fragebogenkonstruktion. Munich: Pearson Studium.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the Multimethod-Multitrait Matrix. *Psychological Bulletin, 56*, 833-853.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Diekmann, A. (2007). *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen* (18th edition). Reinbek bei Hamburg: Rowohlt.

Dillman, D. A., Smyth, J. D. & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys. The tailored design method*. Wiley: New Jersey.

DIN. (2002). DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen. Berlin: Beuth.

Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion.* Göttingen: Hogrefe.

Faulbaum, F., Prüfer, P. & Rexroth, M. (2009). Was ist eine gute Frage? Die systematische Evaluation der Fragenqualität. Wiesbaden: VS Verlag.

Ganzeboom, H. B. G., De Graaf, P. M. & Treiman, D. J. (1992): A Standard International Socio-Economic Index of Occupational Status. *Social Science Research, 21* (1), 1-56.

Groves, R. M, Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2004). *Survey methodology*. New Jersey: Wiley.

Groves, R. M. & Lyberg, L. (2010). Total Survey Error: Past, present and future. *Public Opinion Quarterly, 74*, 849-879.

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement, 1*, 1–10.

Häcker, H., Leutner, D. & Amelang, M. (Eds) (1998). *Standards für pädagogisches und psychologisches Testen.* Supplementum 1/1998 der Diagnostica und der Zeitschrift für Differentielle und Diagnostische Psychologie. Bern: Hogrefe.

Harsch, C., Pant, H. A. & Köller, O. (Eds) (2010). Calibrating standards-based assessment tasks for English as a first foreign language: standard-setting procedures in Germany. Münster: Waxmann

Hartig, A., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Eds), *Testtheorie und Fragebogenkonstruktion. 2nd edition* (pp. 143-171). Heidelberg: Springer Verlag.

Hussy, W., Schreier, M. & Echterhoff, G. (2010). *Forschungsmethoden in Psychologie und Sozialwissenschaften*. Heidelberg: Springer-Verlag.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319-342.

Kane, M. T. (2013). Validating the interpretations and the uses of test scores. *Journal of Educational Measurement, 50,* 1-73.

Kersting, M. (2008). Qualität in der Diagnostik und Personalauswahl - Der DIN Ansatz. Göttingen: Hogrefe.

Krebs, D. & Menold, N. (2014). Gütekriterien quantitativer Sozialforschung. In J. Blasius & N. Baur (Eds). *Methoden der Sozialforschung*, (pp. 425-438). Berlin: Springer.

Krosnick, J. A. & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds), Survey measurement and process quality (pp. 141-164). New York: Wiley.

Krosnick, J. A. & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds), *Handbook of Survey Research* (2nd edition, pp. 263-313). West Yorkshire, England: Emerald Group.

Lienert, G. A. & Raatz, U. (1998) *Testaufbau und Testanalyse.* Weinheim: Beltz PVU.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Massachusetts: Addison-Wesley.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd edition, pp.13-103). New York: American Council on Education/Macmillan.

OECD (2013a). PISA 2013 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy. Paris: OECD.

OECD (2013b). OECD Skills Outlook 2013: First results from the Survey of Adult Skills. Paris: OECD.

Pant, H. A., Stanat, P., Pöhlmann, C. & Böhme, K. (2013). Die Bildungsstandards im allgemeinbildenden Schulsystem. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Eds), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (pp. 13–22). Münster: Waxmann.

Prenzel, M., Sälzer, C., Klieme, E. & Köller, O. (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland.* Münster: Waxmann.

Rammstedt, B. (2010). Messen und Skalieren in den Sozialwissenschaften, Gütekriterien (Reliabilität, Validität, Objektivität). In C. Wolf & H. Best (Eds) *Handbuch der sozialwissenschaftlichen Datenanalyse* (pp. 239-258). Wiesbaden: VS Verlag.

Rammstedt, B. (Eds) (2013). Grundlegende Kompetenzen Erwachsener im internationalen Vergleich - Ergebnisse von PIAAC 2012. Münster: Waxmann.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* New York: Taylor & Francis.

Robinson, J. P., Shaver, P. R., Wrightsman, L. S. (1999). *Measures of political attitudes*. San Diego: Academic Press.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2nd edition.). Bern: Huber.

Schnell, R. (2012). Survey-Interviews. Methoden standardisierter Befragungen. Wiesbaden: VS-Verlag.

Schnell, R., Hill, P. B. & Esser, E. (2011). *Methoden der empirischen Sozialforschung.* Munich: R. Oldenbourg Verlag.

Schuman, H. & Presser, S. (1981). Questions and answers in attitude surveys: Experiments in question form, wording, and context. New York: Academic Press.

Schwartz, D. C. (1973). Political alienation and political behavior. Chicago: Aldine.

Schwartz, S. & Bilsky, W. (1990). Toward a theory of the universal content and structure of values: Extensions and cross cultural replications. *Journal of Personality and Social Psychology, 58,* 878-891.

Sudman, S., Bradburn, N. M. & Schwarz, N. (1996). Thinking about answers: The application of cognitive processes to survey methodology. San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L. J. & Rasinski, K. (2000). *The psychology of survey response.* Cambridge: Cambridge University Press.

Appendix A

Sources of information on quality assurance in the area of the representativeness of survey data

Coverage error and sampling error

Bertino, S. (2006). A measure of representativeness of a sample for inferential purposes. *International Statistical Review, 74*, 149-159.

Kruskal W. & Mosteller F. (1979) Representative sampling, I: Non-scientific literature. *International Statistical Review 47*, 13-24.

Kruskal W. & Mosteller F. (1979) Representative sampling, II. Scientific literature, excluding statistics. *International Statistical Review 47*, 111-127.

Kruskal W. & Mosteller, F. (1979) *Representative sampling, III: The current statistical literature. International Statistical Review 47*, 245-265.

Kruskal W. & Mosteller, F. (1980) Representative sampling, VI: the history of the concept in statistics, 1895-1939. International Statistical Review 48, 169-195.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press

Särndal, C.-E., Swensson, B & Wretman, J. (1992). Model Assisted Survey Sampling. Berlin et al.: Springer.

Särndal, C.-E. & Lundström, S. (2005). Estimation in Surveys with Nonresponse. Chichester: John Wiley & Sons.

Schouten, B., Cobben, F. & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology, 35,* 101-113.

Valliant, R., Dever, J. A. & Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples. Statistics for Social and Behavioral Sciences,* Vol. 51. Berlin et al.: Springer.

Nonresponse error

AAPOR (Ed.) (2011). *Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys.* Available at: http://aapor.org/Content/NavigationMenu/AboutAAPOR/StandardsampEthics/StandardDefinitions/StandardDefinitions2011.pdf [Accessed on January 1, 2014].

Groves, R. M. & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.