gesis Leibniz Institute for the Social Sciences

Documenting Measurement Instruments for the Social and Behavioral Sciences

Isabelle Schmidt & Clemens M. Lechner

June 2020, Version 1.0

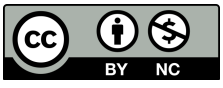
Abstract

Measurement is the key to any quantitative science. Both the validity and the replicability of research findings hinge on the quality of the measurement instruments used. This is especially true for the social and behavioral sciences. Here, researchers often investigate constructs such as personality, attitudes, values, intentions or behavior. Such 'latent' constructs cannot be directly observed but can only be inferred indirectly from the respondents' responses to a survey (i.e., items, scales, questionnaires or tests), which lends even greater importance to the quality of the measurement instruments used. For this reason, a thorough documentation of measurement instruments is an integral part of a transparent research practices. In this guideline, we list information that is crucial for the documentation of measurement instruments in surveys. The guideline is aligned with the quality standards for the documentation of measurement instruments in the social sciences (RatSWD, 2014) and with the standards for the documentation of psychological characteristics of the test evaluation system of the Test Board of the Federation of German Psychologists. Whereas some of these quality standards were developed for testing individuals in applied settings, our lists focuses specifically on documenting measurement instruments intended for social and behavior science research. In research, slightly different quality criteria apply compared to individual diagnostics.

Citation

Isabelle Schmidt & Lechner, Clemens M. (2020). Documenting Measurement Instruments for the Social and Behavioral Sciences. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS - Survey Guidelines).

DOI: 10.15465/gesis-sg_en_033



Measurement is key to any quantitative science. Both the validity and replicability of research findings hinge on the quality of the measurement instruments used. This is especially true in the social and behavioral sciences. Here, researchers often investigate latent, unobservable constructs that can only be inferred indirectly from respondents' answer to a survey (i.e., items, scales, questionnaires, or tests). The quality of social and behavioral science research, therefore, hinges critically on the quality of the measurement instrument. For this reason, a thorough documentation of measurement instruments is part and parcel of transparent research practices.

In the following, we list information that is crucial for the documentation of measurement instruments to assess characteristics of individuals (e.g., personality, attitudes, values, intentions and behavior) in surveys. The guideline is aligned with the quality standards for the documentation of measurement instruments in the social sciences (RatSWD, 2014) and with the standards for the documentation of psychological characteristics of the test evaluation system of the Test Board of the Federation of German Psychologists' Associations (in German: Testbeurteilungssystem des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen [TBS-DTK]). Whereas some of these quality standards were developed for testing individuals in applied settings, our lists focuses specifically on applications of measurement instruments in social and behavior science research. In research, slightly different quality criteria apply compared to individual diagnostics.

The application of the standards refers to a single measurement instruments, that is, questions, test items or item sets in a survey. The information that should be documented can be categorized according to the following four subject matters:

1. the measurement instrument itself
2. its theoretical background
3. its development and
4. standard quality criteria of survey instruments.

1. Information on the measurement instrument

Instructions and items

These checklist items focus on the core of the measurement instruments: the wording of instructions and items as well as how they are presented to respondents.

- ☐ Was the wording of the **introductory sentences** and **instructions** for the measurement instrument that participants receive specified?
- ☐ Was the exact **wording of the items** of the measurement instrument named?
- ☐ If applicable: Was the **order of presentation** of items described?

Response specifications

These checklist items focus on the response scale(s) on which respondents answer the items.

- ☐ What was the **nature of the response categories** (e.g., fully labelled, end-point only labelled scale, rating scale, frequency scale)?
- ☐ What was the **number of response categories**?
- ☐ If applicable: What were the **verbal labels** of the response categories?

- ☐ If applicable: What were the **numerical labels** of the response categories?
- ☐ If applicable: Which **visual stimuli labels** in the scale (e.g., smileys or abstract depictions of humans) were used?
- ☐ If applicable: Which **spatial layout** of the response options (e.g., horizontal or vertical; only relevant in print or web settings) was used?
- ☐ If applicable: Are screenshots or screen recordings (e.g., for interactive web-based surveys) made publicly available (e.g., at a public repository)?

Scoring

These checklist items focus on how the researchers arrive at final scores per respondent which can then be used for substantive analyses.

- ☐ What were the **numerical values** assigned to each response category?
- ☐ Was the **recoding** of individual items or scales necessary before scoring (e.g., inverting negatively keyed items; ipsatizing items)?
- ☐ How were items combined to compute **scale score(s)** (e.g., a weighted or unweighted sum score)?
- ☐ Was any **scaling or transformation** of scores recommended (e.g., to a mean of 100 and a standard deviation of 15)?
- ☐ If applicable: How were **items mapped to subscales**?
- ☐ How were **missing responses** (item non-response) handled?

Application field

These checklist items focus on the aims for which researchers can use the instrument.

- ☐ What was the **purpose** of the measurement instrument (i.e., what it intends to measure and why)?
- ☐ What was the **survey mode** for which the instrument was developed (e.g., web-based, paper & pencil, or verbal interviewing)?
- ☐ For which **target population** (e.g., "working-age women", "immigrant youth") was the measurement instrument developed and validated?

2. Information of the theoretical background

These checklist items focus on the theoretical tradition and concepts based on which the instrument was developed.

- ☐ Was the measurement instrument **relevant to the field** or research tradition?
- ☐ Was the **theoretical background** of the underlying construct clearly outlined?
- ☐ Was any **relevant literature** on which the instrument is based cited?

3. Information about the development

Item generation and selection

These checklist items focus on the way in which the instrument was developed.

- ☐ What **sources of information** were used during item generation (e.g., existing scales, literature research, focus groups, expert opinion)?
- ☐ Based on which **selection criteria** were items included in (or excluded from) the final version of the instrument (e.g., item characteristics and/or theoretical considerations)?
- ☐ Were **expert judgements** used in the selection of items during the development of the measuring instrument? If, yes:
  - ☐ Were the **technical level of training** and the **experience of the experts** indicated in the documentation for the measuring instrument?
  - ☐ Were the experts' assessments described and the **degree of agreement** between the experts indicated?
- ☐ How were **items generated** (e.g., four experts wrote them)?
- ☐ Were any **pretests** conducted (e.g., cognitive pretests, webprobing) to ensure the comprehensibility of the items?
- ☐ In case of an **adaption of an existing instrument** in another language:
  - ☐ Was the **translation procedure** described?
  - ☐ What **standards** were used for the translation (e.g., back-translation or the TRAPD approach)?

Sample(s) and data

These checklist items focus on the sample and data that were used to develop and validate the instrument.

- ☐ Were the **sample(s) used for the development** and the sample(s) for the **evaluation** of the measurement instrument with respect to the following points described?
  - ☐ Was the **recruitment** of the sample(s) (simple random sampling, stratified random sampling, cluster sampling, ad-hoc sample; participation with or without payment) described?
  - ☐ Was the **timing** of data collection (e.g., year and month) indicated?
  - ☐ Was the **composition** of the sample(s) (e.g., gender, age, educational achievement, mother language, socioeconomic status, geographic region) described?

Item analyses

These checklist items focus on how the instrument was analysed in the validation process.

- ☐ Which **item statistics** (e.g., frequencies, range, mean/median, variance/standard deviation, skewness, kurtosis, percent missing) were inspected to assess quality?
- ☐ Which **item parameters** were reported that allow to rate the item quality? (e.g., sign and size of path coefficients [from construct to item] or means [of the items] of a structural equation model,

the item discrimination parameters and threshold of an IRT [item response theory] model can be presented, or, alternatively, means, standard deviations, and selectivity of the manifest items.)?

- ☐ If applicable: Was the **dimensionality** of the measurement instrument tested and through what methods was it established?

Technical details

These checklist items focus on how the analyses were implemented and how secondary data users can replicate them.

- ☐ Which **statistical software** (e.g., Mplus, R packages, Stata, SPSS) and, if applicable, **which packages** and which version was used to run the analyses?
- ☐ Was there information about **missing data** patterns and the handling of missing data?
- ☐ Are **replication files** (data and code) available (e.g., from a repository)?

4. Information about quality criteria

These checklist items focus on the three main quality criteria for measurement instruments (objectivity, reliability, and validity) and a range of additional criteria that contribute to quality and transparency.

Objectivity

- ☐ Was the objectivity of **application** of the measurement instrument (e.g., availability of instrument administration guidelines or procedures for interviewer training) described?
- ☐ Was the objectivity of **evaluation** of the measurement instrument evaluated?
- ☐ Was the objectivity of **interpretation** of the scale-score(s) evaluated?

Reliability

- ☐ Which **reliability estimate** was reported (e.g., test–retest, split-half, internal consistency measures such as Cronbach's Alpha or McDonald's Omega, Andrich's reliability in item response theory models)?
- ☐ What was the **rationale or justification** for each reliability estimate?

Validity

- ☐ If applicable: **Content validity** – does the wording of each item match the definition of the concept/construct it is intended to measure; is it a "pure" indicator of the concept/construct and free of irrelevant content, etc.?
- ☐ If applicable: **Factorial validity** – does the actual (empirical) structure of the instrument matches the theoretical or intended structure of the instrument?
- ☐ If applicable: **Criterion validity** (concurrent or predictive validity) and/or **convergent and divergent validity** – was the measurement instrument related to external variables in plausible ways?

Further quality criteria

- ☐ **Reference Data** – Were descriptive statistics of the sample data (e.g., range, means/medians, variances/standard deviations, frequencies) presented?
- ☐ **Test Economy** – is the instrument with respect to the material **costs** (e.g., licencing fees, material costs, or complex test administration procedures) and the **duration** of the administration of the instrument efficient?
- ☐ **Fakeability** and **response bias** – is the instrument susceptible to untruthful or misleading answering?
- ☐ **Test Fairness** – have different groups of persons the same chance to of reaching comparable scores?
- ☐ **Reasonableness** - is the instrument mentally, physically and in temporal terms acceptable?

Further readings

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah; NJ: Erlbaum.

References

RatSWD. (2014). *Qualitätsstandards zur entwicklung, anwendung und bewertung von messinstrumenten in der sozialwissenschaftlichen umfrageforschung [quality standards for the development, application, and evaluation of measurement instruments in social science sur- vey research]*. Retrieved from http://www.ratswd.de/themen/qualitaetsstandards