gesis Leibniz Institute for the Social Sciences

ZIS Publication Guide (Version 2.1)

Guideline for Documenting Instruments in the Open Access Repository for Measurement Instruments (ZIS)

Désirée Nießen, Katharina Groskurth, Isabelle Schmidt, Matthias Bluemke

Citation

Nießen, D., Groskurth, K., Schmidt, I., & Bluemke, M. (2020). ZIS Publication Guide (Version 2.1) – Guideline for documenting instruments in the open access repository for measurement instruments (ZIS). https://doi.org/10.6102/pubguide2.1.English

ZIS

Open access measurement instruments for the social science
www.gesis.org/en/zis

gesis Leibniz Institute for the Social Sciences

Table of contents

The Documentation of Measurement Instruments in ZIS

For publishing your instrument, use the template (English) to ease the documentation and publication process. Any publication in ZIS is structured in six sections: overview, instrument, theoretical background, scale development, quality criteria, and literature and data sources. Due to the consistent format, users can compare different instruments easily and quickly. This section follows the structure laid out in the template and includes information to help you optimally design your documentation using the guidelines of the German Council for Social and Economic Data (Rat für Sozial- und Wirtschaftsdaten; RatSWD, 2014). Likewise, it is helpful to browse ZIS to familiarize yourself with the structure and scope of the publications.

1. Overview

Abstract

Please use maximum 120 words to provide a summary of your instrument, its construct, and its application field. You should additionally point out subscales of your construct and well-known population surveys that included your instrument (if applicable).

2. Instrument

Instruction

In this section, you should describe the instructions for scale use.

Items

Please present your items in the table below. All items should be listed and numbered. If applicable, you should also indicate the subscale to which an item refers. When an item needs to be recoded so that a higher value represents a higher expression of the construct or the subscale, you should indicate it with a minus in the column “polarity.””

Table 1

Items of the Scale …

No.	Item	Polarity	Subscale
1	Here, the first item is listed.	+	A
2	Here, the second item …	–	B

Response specifications

Here, you should specify the possible response categories. Additionally, you should describe numerical codes and labels for the response categories.

Scoring

This section describes how numerical values are assigned to different response categories. It also mentions when items are reverse-coded, which items can be combined to form a scale score, whether

there are subscales that must be analysed separately, and whether simple or weighted total scores are formed. In addition, please recommend how to handle unanswered items or (sub-)tasks and how to interpret the results.

Application field

This section describes

- the purpose of the instrument
- the survey mode typically used for the instrument (e.g., web-based, paper-and-pencil, or verbal interviewing; [PAPI: paper-and-pencil personal interviewing; CAPI: computer-assisted personal interviewing; PASI: paper-and-pencil self-administered interviewing; PATI: paper-and-pencil telephone interviewing; CATI: computer-assisted telephone interviewing; CASI: computer-assisted self-administered interviewing])
- the target population(s) for whom the instrument has been developed
- whether the instrument is also suitable for individual diagnostics

If special professional qualifications are required for using the instrument, please mention these.

3. Theory

This section describes the theoretical background of the instrument. Please explain why the instrument is relevant and from which theory you derived the construct. This section should include a definition of the underlying construct, a description of the subscales (if applicable), and a description of the instrument's relations to other constructs (interdependencies and demarcations). Make sure to reference any relevant literature. The section should be used to answer the following questions:

- Which construct (and, if applicable, subscales) should be captured by the items?
    - Definition of the construct

      Example (Schwartz et al., 2015):

      "Human values refer to what is important to people in their lives and the goals they strive to attain. According to Schwartz (1992), values "(1) are concepts or beliefs, (2) pertain to desirable end states or behaviors, (3) transcend specific situations, (4) guide selection or evaluation of behavior and events, and (5) are ordered by relative importance" (p. 4). Values differ based on the specific goals they express and motivate people to pursue. Schwartz (1992) specified ten different types of values, which are recognized across cultures, which are operationalized in the Human Values Scale: conformity, tradition, benevolence, universalism, self-direction, stimulation, hedonism, achievement, power, and security."

    - Relations to other constructs (interdependencies and demarcations)

      Example (Breyer & Danner, 2015):

      "Grit is related to self-control (ability to resist temptation and control impulses) and need for achievement (implicit pursuit of moderately difficult goals, followed by immediate feedback). However, it differs from these concepts due to its long-term-intensity, awareness and consistency despite missing feedback. According to Duckworth et al. (2007), grit is highly correlated with Big Five Conscientiousness, but has an incremental validity of beyond conscientiousness and IQ regarding success measures."

- Why is the instrument relevant?

  Example (Breyer, 2015a):
  "The Left-Right Self-Placement Scale is often used for self-placement and external assessment of political attitudes and to position political parties on the left-right spectrum."

4. Scale development

Item generation and selection

This section is aligned with the formal guidelines of the German Council for Social and Economic Data (Rat für Sozial- und Wirtschaftsdaten; RatSWD, 2014), standard 1 (instrument development). Here, you should describe how the items were generated and according to which criteria items were selected. In case of revisions, the description should include the date the items started being used in their current format. Please indicate who developed your items, according to which construction principle, and which items represent which subscale. If expert judgements have been used to determine whether the scale represents the defined content area, please describe the specialized levels of education, experience, and the qualifications of the experts involved. You can also explain how the experts made their judgments and to what extent the judges agreed. If you conducted a pilot study, please provide descriptive statistics on sample size, age, and gender. Changes from previous versions of a scale should be noted, for example, when you reworded or discarded some items. Please answer the following questions in this section:

- Who developed the items? When were the items developed?

  Example (Schwartz et al., 2015):
  "The items of the ESS Human Values Scale are based on the Schwartz Value Survey (SVS; Schwartz, 1992) and the Portrait Values Questionnaire (PVQ; Schwartz, 2003)."

- When was the scale introduced?

  Example (Breyer & Voss, 2016):
  "The Happiness and Satisfaction Scale has been used since 2002 in the Family and Changing Gender Roles module of the ISSP."

- According to which construction principle the items were developed?

  Example (Schwartz et al. 2015):
  "The ESS board representing each country allocated 21 items to measure the ten values, three for universalism and two for the remaining values."

- Did the items change over time (e.g., items were reworded/discarded)?

  Example (Braun, 2014):
  "ISSP 1988 includes an additional dimension which relates to the "economic consequences (of female work)". The items asked in the ISSP 1988 study are: […]
  In the gender-role battery of ISSP 1994 (Zentralarchiv, 1997), which is documented here on the basis of data from East and West Germany and the United States, one item of the 1988 battery was dropped and three added: one for the economic consequences dimension and one item relating to the role of men for each of the two other dimensions."

If the items were translated or adapted, please describe how the translation process was carried out (e.g., single or double translation), by whom (e.g., professional translators, substantive experts, mem-

bers of research team), and the review and assessment steps implemented (e.g., team discussions, consultation with developers, expert reviews, or pretesting). Information on translation of instruments and best practice procedures can be found in International Test Commission (2017), Mohler et al. (2016) and Rios and Sireci (2014).

Please answer the following questions:

- How were the items or the scale translated (adapted) and reviewed (e.g., pretesting), and by whom?

  Example (Nießen et al., 2019):
  "To enhance the usability of the KSA-3, and to enable social surveys to use the KSA-3 in an English-language context, the scale was adapted to the English language and validated for the UK population. First, the nine items of the KSA-3 were adapted to English by translating the items following the TRAPD approach (Translation, Review, Adjudication, Pretesting, and Documentation; Harkness, 2003), whereby two professional translators (English native speakers) translated the items independently of each other into British English and American English, respectively. Second, an adjudication meeting was held where psychological experts, the two translators, and an expert in questionnaire translation reviewed the various translation proposals and developed the final translation."

- Were any items particularly challenging to translate, or did any items require adaptations (i.e., intentional deviations for cultural or design reasons)?

  Example challenge (Van Widenfelt et al., 2005):
  "For example, one of the items from the Negative Affect Self-Statement Questionnaire (NASSQ; Ronan, Kendall, & Rowe, 1994) that we translated, "I am a winner," was literally translated in a Belgian version of the questionnaire as "ik ben een winnaar." For the same item, it was clear from our pilot interviews with Dutch children and adolescents that such a literal translation was not culturally appropriate for use in the Netherlands, and thus risked insufficient endorsement and a skewed response pattern. Some children relayed to us that they understood the item but said that they would never say or think that, nor did they expect other Dutch children would. We discussed with the children similar translations such as "I am the best," but the children had the same reactions as they had to "I am a winner." Hence together with the children that we interviewed and later with our translation team we decided on "alles lukt me" (I will succeed in everything I do/I can do anything)."

  Example cultural adaptation (Arce-Ferrar, 2006):
  "Four of 16 items of the Spanish version of GERS [Greenleaf's (1992) Extreme Response Scale (GERS)] dealing with aspects relevant to the U.S. context but not to that of Mexico were modified. For example, an item that had content dealing with "family investing in the stock market" was judged irrelevant and rephrased as "family saving in certificate deposits," which is a more likely financial practice."

Detailed translation descriptions are provided for use as examples in Arnold et al. (2015) or Martinez et al. (2006).

Samples

Here, you should describe the samples used for the development and evaluation of your scale. Please describe the recruitment of the samples and characteristic features (e.g., gender, age, educational level, and potentially further relevant features such as mother language). Make sure to provide infor-

mation on the sampling (e.g., sampling plan, type of sample [random sample, stratified random sample, quota sample, stratified sample, ad hoc sample]), the participation rates (e.g., nonresponse), when the survey took place, the objective of the survey when it was conducted, whether participation was rewarded, and the survey mode. You should also describe whether your data include missing values and how missing values were handled (e.g., listwise deletion, pairwise deletion, imputation, or full information maximum likelihood).

Item analyses

Here, you should describe results that provide information about the dimensionality and item quality of the instrument, stating the statistical software used (e.g., SPSS, Mplus, Stata, R, SAS). The dimensionality can be tested via exploratory factor analysis (EFA), confirmatory factor analysis (CFA), or structural equation modeling (SEM) within the framework of the classical test theory. Alternatively, it can be tested by the item response theory (IRT). When you assume that the dimensionality of the items is the same between the included groups (e.g., age or educational groups), the data can be analysed without any group separation (pooled data). When you assume that the dimensionality of the items differs between the included groups, the different dimensionality should be considered and the analysis should be completed separately for the groups or via multi-group models (e.g., multi-group SEM; MG-SEM).

If you describe an EFA (principal axis factor analysis [PAF], principal component analysis [PCA]), you should include the extraction and the rotation method, at minimum the initial course of Eigenvalues (e.g., the screeplot), the factor loadings matrix (after the rotation) according to the PAF (or PCA), and the final communalities. SEM lends itself to present the results of the measurement model in a figure. For more detailed information, refer to appendix A1 and A2.

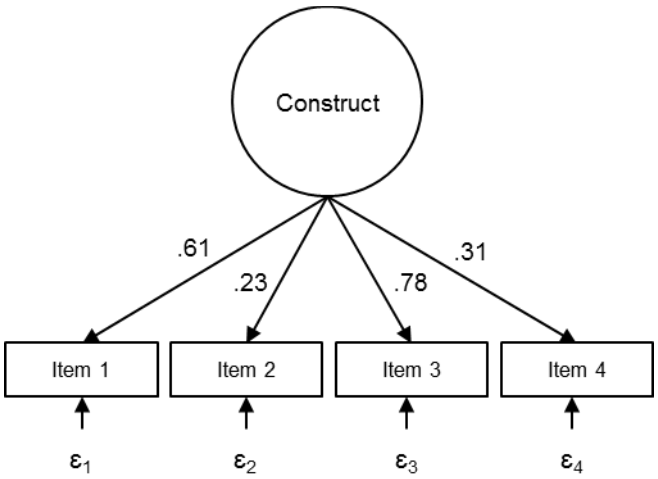

Figure 1. Tau-congeneric measurement model for the construct. Standardized path coefficients, RMSEA = .030, CFI = .980, SRMR = .020, $\chi^2(4) = 6.232$, $p = .183$, $N = 1{,}236$.

If you carried out item analyses within the IRT, show that there is evidence for the final model. For example, check if various parameter restrictions hold: test a one-parametric model (1PL) such as the Rasch model for binary data or the partial credit model for ordinal data against a two-parametric model (2PL) such as the Birnbaum model for binary data or the generalized partial credit model for ordinal

data. If applicable, please evaluate the multidimensionality using different models. We also recommend that the item characteristic curve (ICC) and the test information curve should be mapped for the final model.

Item parameter

Here, present characteristic parameters that allow the rating of the item quality. Please demonstrate (for continuous items) means, standard deviations, skewness, kurtosis, and selectivities (item-total correlations) of the manifest items. Alternatively, present path coefficients (from construct to item), means (of the items) of a SEM, or item discrimination parameters and the threshold of an IRT model.

In general, the following standards apply to the presentation: parameters as *M* (for mean) or *SD* (for standard deviation) are presented in italics except for Greek letters (e.g., χ^2). In principle, all numbers should be presented with two digits after the decimal point (e.g., $M = 4.23$). Probabilities (*p*-values) and fit indices (χ^2, RMSEA, SRMR, CFI) should have three digits after the decimal point (e.g., $p = .003$; RMSEA = .030). The decimal separator is a dot. Characteristic parameters that can only range between 0 and 1 are displayed without a leading zero (e.g., $p < .001$). Presenting the item values in a table may be useful.

Table 2

Means, Standard Deviations, Skewness, Kurtosis, and Selectivities of the Manifest Items

	M	*SD*	Skewness	Kurtosis	Selectivity
Item 1					
Item 2					
Item 3					
…					

Note. Scale from 0 (*not correct*) to 4 (*correct*), $N = 1{,}236$.

5. Quality criteria

Objectivity

This section is aligned with the formal guidelines of the German Council for Social and Economic Data (RatSWD, 2014), standard 5 (minimization of the process error). In this section, rate the objectivity of application, evaluation, and interpretation of the scale. An instrument can be regarded as an objective tool when it works independent of the administrator (*objectivity of application*) and independent of the evaluator of the test (*objectivity of evaluation*). Additionally, unambiguous and user-independent rules should be provided (*objectivity of interpretation*).

> *Example (Breyer & Bluemke, 2016):*
> "For the Work-Family Conflict Scale, there are several factors supporting objectivity. Firstly, the scale was administered in personally conducted face-to-face interviews in most countries, and the interviewers were specially trained. Secondly, objectivity is supported by the standardized questionnaire format and written instructions. Finally, as ordered and labelled categories in a fixed order are used to supply answers, and since norming data are available (see descriptive statistics), the application as well as the interpretation of the scale can be considered as very objective."

Reliability

This section mirrors official guidelines supplied by the German Council for Social and Economic Data (RatSWD, 2014), standard 4 (reliability). In this section, you should describe reliability estimates and may present confidence intervals. A test can be deemed reliable when it measures a specific construct within a small margin of measurement error. The more the reliability coefficient approaches unity, the higher the reliability. When using measurement models in line with classical test theory, Cronbach's alpha, split-half, or (test-)retest correlation can be computed. Due to the strong but often violated assumptions underlying alpha, only Raykov‘s rho or McDonald‘s omega appropriately reflect scale reliability. Use omega-h (ω_h) for the reliability of the (unweighted) scale score in terms of the general factor underlying all items (even when unidimensionality does not hold). Use omega (ω or ω_{total}) for the reliability of all systematic (reliable) aspects of item covariance in the scale score in terms of both general factor and specific factors (bifactor model) underlying the items or item-subsets. When using categorial (ordinal) indicators, less than five response options call for computing ordinal alpha or omega-categorical. It is recommended to explain the appropriateness of the reliability estimation method(s) being used.

In contrast to classical test theory, which describes the observed value as a composition of the true value and the measurement error, IRT focuses on the probability of observing the value for different skill levels. Thus, the accuracy of the measurement also varies depending on the skill level. The accuracy of the measurement is thus person-specific in the IRT context. Nevertheless, reliability estimators for the average reliability of the estimation of person parameters can be reported for IRT models as well. In this case, we recommend using the Andrich reliability or the scale reliability for binary variables according to Raykov et al. (2010). There is a description of the specific estimators in appendix A3. Reliability in the GESIS Survey Guidelines (Danner, 2016) provides a detailed overview of reliability estimation.

Validity

This section is aligned with the standard 2 (validity) and standard 3 (minimization of method-specific effects) of the formal guidelines of the German Council for Social and Economic Data (RatSWD, 2014). Here, you should present results that indicate the content validity, construct validity, or criterion validity of the instrument. An instrument can be viewed as valid when it captures the construct of interest (the construct that was intended to be measured). Please make sure to distinguish between different types of validity and give a clear and complete presentation of these aspects.

You may also clarify whether scores refer to an entire scale, to subscales, or to single items and whether scores refer to a raw value or a standardized value. If you used statistical adjustment/optimization methods to determine the validity (e.g., correction for attenuation, correction for variance restriction, or controlling for covariates in multiple regression), you should report both the uncorrected and the corrected values. Appropriately, label all the statistics used in connection with the adjustment and report simple parameter estimates (e.g., zero-order correlations) in addition to adjusted estimates (e.g., multiple regression coefficients). You may also support the validity claim by referencing the demonstrated validity in other researchers’ investigations.

For construct validity, please consider how the construct in question relates to (dis-)similar constructs. For criterion validity, argue why the selected criterion is appropriate and valid. Reflect on the objectivi-

ty, reliability, and demographic fit/applicability to the current population (e.g., educational level, age, and work experience) of each criterion measure, especially in light of the instrument's target population. Confidence intervals may be presented in addition to point estimates. More detailed information can be found in appendix A4.

In your documentation, you should comment on the variables you chose for investigating construct and criterion validity and provide the respective results in a table (see example below). The interpretation of validity coefficients may follow norms for interpreting effect sizes, such as Cohen's (1992) guidelines for interpreting group differences: small effect (r = .10), medium effect (r = .30), and large effect (r = .50). For individual differences, Gignac and Szodorai (2016) suggest interpreting correlation coefficients of r = .10, .20, and .30 as small, typical, and large.

Example (Breyer, 2015b):

Table 4

Correlation of the Social Trust Scale with Other Variables

	Social Trust
Political trust	.44
Participating voluntarily	.13
SES	.22
Happiness	.28
Optimism	.14
Security (HVS)	-.16
Country's health system	.21
Country's democracy	.25
Country's education	.19

Note. Pearson correlation coefficients. All correlations are significant, $p < .001$ (two-tailed). Observations are weighted based on design weights and population weights.

Descriptive statistics (scaling)

Here, provide details of the distribution of the scale score(s). For continuous scale score(s), please provide mean values, standard deviations, skewness, and kurtosis to serve as comparative data.

If intended, you can also provide information on standardization in the form of standard tables (i.e., norm tables). The requirements for a standardization sample (i.e., norming sample) and the preparation of standards are described in appendix A6. Standard tables can be made available as a separate document under the "Downloads" tab in ZIS.

Further quality criteria

This section is aligned to the formal guidelines of the German Council for Social and Economic Data (RatSWD 2014), standard 6 (further quality criteria). Please report results here that can be used to evaluate further quality criteria such as economy, falsifiability, and response bias or test fairness. For the evaluation of economy, please provide information on the processing time and indicate whether this is your estimate or a data-based median or mean value. You can indicate whether falsification is to be expected, whether and how falsification can be counteracted by the type of specifications and

implementation, and if falsification can be counteracted by the evaluation (if applicable).To assess test fairness, please investigate and ensure measurement invariance for samples that differ in age, gender, or nationality. You can investigate measurement invariance within the classical test theory or differential item functioning (DIF) within the IRT. Using SEM or IRT models, you can investigate different forms of measurement invariance/DIF. We provide a detailed description of the levels of measurement invariance and their investigation with SEMs in appendix A5. For various procedures to evaluate DIF as defined by IRT, we recommend the articles of Dimitrov (2017) and Montoya and Jeon (2019). Dimitrov offers a combined method for binary IRT and SEMs in the sense of a common DIF metric. Montaya and Jeon postulate a method for binary and ordinal models that can be applied in IRT as well as in the context of SEMs.

6. Literature and data sources

Further literature

In this section, you may refer to the relevant original literature.

Data sources

Here, you may present links to datasets that contain your instrument and are accessible online. In the spirit of Open Science, used data sets (including tables, outputs, and evaluation files [codes]) should be made publicly accessible. You can save and publish these for free in the GESIS datorium.

Acknowledgements

Here, you may thank colleagues for their help and support.

Contact details

Please provide your contact data in this section of the documentation.

References

The bibliography contains all literature referenced in the text. The citation guidelines of the American Psychological Association (2019) apply.

American Psychological Association (2019). *Publication manual of the American Psychological Association* (7th edition). American Psychological Association.

Arce-Ferrer, A. J. (2006). An investigation into the factors influencing extreme-response style: Improving meaning of translated and culturally adapted rating scales. *Educational and Psychological Measurement, 66*, 374–392. https://doi.org/10.1177/0013164405278575

Arnold, B., Mitchell, S. A., Lent, L., Mendoza, T. R., Rogak, L. J., Barragán, N. M., Willis, G., Medina, M., Lechner, S., Penedo, F. J., Harness, J. K., & Basch, E. M. (2016). Linguistic validation of the Spanish version of the National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). *Supportive Care in Cancer, 24,* 2843–2851. https://doi.org/10.1007/s00520-015-3062-5

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika. 74*, 137–143. https://doi.org/10.1007/s11336-008-9100-1

Braun, M. (2014). Gender-role attitudes (ISSP 94). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS).* https://doi.org/10.6102/zis223

Breyer, B. (2015a). Left-Right Self-Placement (ALLBUS). *Zusammenstellung sozialwissenschaftlicher Items und Skalen.* doi:10.6102/zis83

Breyer, B. (2015b). Social Trust Scale (ESS). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS).* https://doi.org/10.6102/zis235

Breyer, B. & Bluemke, M. (2016). Work-Family Conflict Scale (ISSP). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS).* https://doi.org/10.6102/zis243

Breyer, B. & Danner, D. (2015). Grit Scale for Perseverance and Passion for Long-Term Goals. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS).* https://doi.org/10.6102/zis237

Breyer, B., & Voss, C. (2016). Happiness and Satisfaction Scale (ISSP). Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS). https://doi.org/10.6102/zis240

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*, 230–258. https://doi.org/10.1177/0049124192021002005

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504. https://doi.org/10.1080/10705510701301834

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cortina, J. M. (1993). What is coefficient Alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Danner, D. (2016). *Reliability – The precision of a measurement. GESIS Survey Guidelines.* GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_011

Dimitrov, D. M. (2017). Examining differential item functioning: IRT-based detection in the framework of confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development*, *50*, 183–200. https://doi.org/10.1080/07481756.2017.1320946

DiStefano, C., Liu, J., Jiang, N., & Shi, D. (2018). Examination of the weighted root mean square residual: Evidence for trustworthiness? *Structural Equation Modeling: A Multidisciplinary Journal, 25*, 453–466. https://doi.org/10.1080/10705511.2017.1390394

Gabler, S., & Quatember, A. (2013). Repräsentativität von Subgruppen bei geschichteten Zufallsstichproben [Representativeness of subgroups in stratified random samples]. *AStA Wirtschafts-und Sozialstatistisches Archiv,* 7, 105–119. https://doi.org/10.1007/s11943-013-0132-3

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. doi:10.1016/j.paid.2016.06.069

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1–55. https://doi.org/10.1080/10705519909540118

International Test Commission (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.).

https://www.intestcom.org/

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*, 936–949. https://doi.org/10.3758/s13428-015-0619-7

Martinez, G., Marín, B. V., & Schoua-Glusberg, A. (2006). Translating from English to Spanish: The 2002 National Survey of Family Growth. *Hispanic Journal of Behavioral Sciences, 28*, 531–545. https://doi.org/10.1177/0739986306292293

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

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*, 568–592. https://doi.org/10.1037/0021-9010.93.3.56

Mohler, P., Dorer, B., de Jong, J. & Mengyao, H. (2016). *Translation:* Overview. *Guidelines for Best Practice in Cross-Cultural Surveys.* Survey Research Center, Institute for Social Research, University of Michigan. Cross-cultural survey guidelines. https://ccsg.isr.umich.edu/index.php/chapters/translation-chapter/translation-overview

Montoya, A. K., & Jeon, M. (2019). MIMIC models for uniform and nonuniform DIF as moderated mediation models. *Applied Psychological Measurement*, Advance online publication. https://doi.org/10.1177/0146621619835496

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers, 32,* 396–402. https://doi.org/10.3758/BF03200807

Nießen, D., Schmidt, I., Beierlein, C., & Lechner, C. M. (2019). An English-language adaptation of the Authoritarianism Short Scale (KSA-3). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS).* https://doi.org/10.6102/zis272

Pokropek, A., Davidov, E., & Schmidt, P. (2019): A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 26*, 724–744. https://doi.org/10.1080/10705511.2018.1561293

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163. https://doi.org/10.2307/271063

RatSWD (Ed.). (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 survey research].* RatSWD. http://www.ratswd.de/themen/qualitaetsstandards

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173–184. https://doi.org/10.1177/01466216970212006

Raykov, T. (2001). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25*, 69–76. https://doi.org/10.1177/01466216010251005

Raykov, T. Dimitrov, D., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 17*,

265–279. https://doi.org/10.1080/10705511003659417

Raykov, T., & Marcoulides, G. A. (2011*). Introduction to psychometric theory.* Routledge.

Rios, J. A., & Sireci, S. G. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing, 14,* 289–312. https://doi.org/10.1080/15305058.2014.924006

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*, 31–57. https://doi.org/10.1177/0013164413498257

Schwartz, S. H., Breyer, B., & Danner, D. (2015). Human Values Scale (ESS). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS).* https://doi.org/10.6102/zis234

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality, 13(5, Spec Issue),* 389–408. https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0.CO;2-A

Van Widenfelt, B. M., Treffers, P. D., De Beurs, E., Siebelink, B. M., & Koudijs, E. (2005). Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families*. Clinical Child and Family Psychology Review, 8*, 135–147. https://doi.org/10.1007/s10567-005-4752-1

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Doctoral dissertation). SemanticScholar. https://pdfs.semanticscholar.org/7a22/ae22553f78582fc61c6cab4567d36998293b.pdf

Appendix

A1: Exploratory factor analysis

With exploratory factor analysis, one tests whether some factor(s) might underlie the items (variables) even if there are no previously existing assumptions about item relationships. To conduct an exploratory factor analysis in SPSS, you can choose the items that are to be reduced in the dialog box “factor analysis.” These specifications will influence the results:

- Selection of the extraction method:

  Principal component analysis: PCA aims at reducing multivariate data to a few components. Factors (principal components) are derived from linear combinations of the items. The factor loadings represent as much item variance as is possible to predict from the components, that is, true variances as well as measurement errors are included. As a result, factor loadings are usually higher than those of principal axis factor analysis.

  Principal axis factor analysis: PFA aims at understanding the latent constructs underlying the data. The extracted factors do not merely explain the item variance but the covariances among all items. Unlike PCA, PFA takes measurement error into account. PFA does not predict all the item variance, only the shared variance that can be explained by the factors. Often, both PCA and PAF obtain similar results regarding the underlying factors.

- Selection of the number of extracted factors:

  Kaiser criterion: The Kaiser criterion is the default extraction criterion in SPSS. All factors with an Eigenvalue > 1 are extracted automatically. In the logic of PCA, an Eigenvalue > 1 just indicates that the factor explains more variance than an item explains all by itself on average. It is useful to include further criteria for reliable factor extraction.

  Theoretical considerations: When sound hypotheses about the number of underlying factors exist, one can specify the number of factors a priori (e.g., when a certain number of subscales is theoretically determined).

  The course of Eigenvalues: The scree plot (in SPSS under “extraction”) helps to identify the number of factors. Usually, factors are extracted by visually inspecting the “bend” after the steep curve in the course of Eigenvalues. Alternatively, parallel analysis can be run to compare the initial course of Eigenvalues with a second course Parallel analysis generates a course of Eigenvalues based on the same number of items. Unlike the former analysis, the second time the items are normally distributed, yet uncorrelated random variables (for a description and code for SPSS and SAS see O'Connor, 2000). The initial Eigenvalues are only meaningful if they are higher in magnitude than the second ones obtained from random noise.

  The pattern of factor loadings: The number of factors to extract may be supported by the pattern of factor loadings. Unambiguously interpretable loading patterns (e.g., content-related items load highly on the same factor but low on other factors) support the resulting factor structure.

- Selection of the rotation algorithm:

  Orthogonal ("Varimax"): The orthogonal (right-angled) rotation of factors is a reasonable choice when more than one factor exists but no relationships between the factors can be assumed. Hence, the underlying factors are substantively and statistically independent of each other. Within each factor, the variance of the loadings is maximized: items can be clearly assigned to a factor while they show hardly any loadings on other factors.

  Oblique ("Oblimin"): The oblique (oblique-angled) rotation is a reasonable choice when relationships between the factors cannot be denied. Thus, correlations between the rotated factors are permitted. After oblique rotation, factors usually lend themselves to easy interpretation.

A2: Structural equation modeling

Structural equation models are useful for testing hypotheses on whether latent variables underlie the manifest items (indicators). With SEM, one can test measurement models, that is, whether (or how well) the theoretical structure corresponds to the data.

The fit of the measurement model depends on the complexity of the model, the sample, and the distribution of scores for each item. Furthermore, the violation of the normality assumption influences the model fit. Apart from univariate normality, the multivariate normal distribution must hold. We recommend using a robust maximum likelihood estimation (MLR instead of ML) for estimating the model and evaluating its fit. If you decide to use ML estimation, we ask for proof that multivariate normal distribution holds (e.g., non-significant Mardia's Test or Small's Omnibus Test). The WLSMV estimator (weighted least squares means and variance adjusted) can be recommended for ordinal data (e.g., Li, 2016).

It is common to assess the model fit with the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the standardized root mean square residual (SRMR). With WLSMV estimation, the weighted root mean square residual (WRMR) may be used. The model is said to fit sufficiently well when all fit indices fall within the limits of specific cutoffs.

Hu and Bentler (1999) generally recommend ML-based CFI > .95, ML-based SRMR ≤ .08, and ML-based RMSEA ≤ .06 as indicating good model fit. Browne and Cudeck (1992) provide even more restrictive cutoffs for RMSEA, suggesting that RMSEA ≤ .05 points to close fit of the model, RMSEA ≤ .08 indicates reasonable model fit, and RMSEA > .10 indicates a poor fitting model.

Whereas previous suggestions are based on ML estimation, Yu (2002) replicated the formerly suggested cutoffs ($N \geq 250$: CFI > .95; RMSEA ≤ .05) for binary outcomes with WLSMV estimation. She further recommends a threshold of WRMR ≤ 1.0 as indicating good model fit. However, WRMR strongly depends on sample size and unexpectedly improves with severe model misspecification, non-normality, low loadings (.5), and dichotomous data (DiStefano et al., 2018). Therefore, be cautious when assessing fit with this index.

The model fit should not only be assessed by absolute norms but also by the relative merits in comparison to alternative measurement models. Information criteria such as the Bayesian information criterion (BIC) can help decide between competing models (especially when models are not nested but based on the same data set). BIC takes sample size and model complexity into account. The model with the lower BIC value is preferred. Differences in BIC (ΔBIC) ≥ 6 strongly support the model with the lower BIC value (Raftery, 1995).

In sum, SEM allows comparing alternative measurement models. In (at least essential) tau-equivalent measurement models, all items have equivalent unstandardized factor loadings (usually 1.0), indicating that all items reflect their latent variable to the same extent. In these models, Cronbach's alpha is an unbiased estimate of reliability. In tau-congeneric measurement models, items may exhibit different factor loadings in which all items reflect the same construct with different item reliability. For tau-congeneric measurement models, unbiased point estimates of reliability are Raykov's rho (1997) or McDonald's omega (1999). In your documentation, you can compare the fit of a tau-equivalent measurement model with the fit of a tau-congeneric measurement model.

A3: Reliability estimators

Cronbach's alpha: Cronbach's alpha is the most frequently used method for estimating the reliability of a scale. It determines the internal consistency by testing the relations among the items within a scale. If your instrument consists of several subscales, Cronbach's alpha needs to be estimated separately for every subscale. Here, only those items which belong to one subscale should be included in each computation. A fitting essential tau-equivalent measurement model (i.e., equal factor loadings for all items; see dimensionality) is necessary for the application of Cronbach's alpha.

Split-half method: If your measurement model consists of many items, the reliability can be estimated via the split-half method. The items will be separated into two parallel halves; the correlation of the two halves determines the reliability. By default, SPSS splits the scale in its middle: the first half of the items constitutes the first half for the test and the second half of the items constitutes the second half for the test. The SPSS default setting is acceptable when one has homogenous items. Alternatively, the first half of the test consists of all even-numbered items whereas the second half of the test consists of all odd-numbered items (odd-even method). Having heterogeneous items, the subdivision will usually be done according to the selectivity and item difficulty (method of item twins). A prerequisite for estimating reliability by the split-half method is parallel halves; the correlation between the two halves can then be used for estimating their reliability. The reliability of the total scale is commonly estimated by applying the Spearman-Brown prophecy to the split-half correlation to adjust for the different lengths of test-halves and the total scale.

Retest method: The correlation between two points of measurement (i.e., two points in time) gives the (test-)retest reliability. The retest method rests on the assumption that true scores as well as measurement errors do not change between the two points of measurement. To estimate the (test-) retest correlation, scale scores (sums or averages across items) must be computed for the first and second point in time.

As an alternative or additional method, the **composite reliability** according to Raykov (1997), rho ρ, or according to McDonald (1999), omega ω, may be presented. Especially when the results merely rest on a tau-congeneric model (different factor loadings), the composite reliability estimates need to be reported. Cronbach's alpha will either underestimate (heterogeneous factor loadings; Cortina, 1993) or overestimate (correlation of residual variances; Bentler, 2009; Raykov, 2001) the reliability. Having used measurement models of latent state-trait theory (Steyer et al., 1999), estimates of reliability, consistency, measurement specificity (and, if appropriative, method specificity) can be presented. When IRT models were used, Andrich's reliability or the scale reliability according to Raykov et al. (2010) can be reported.

A4: Types of validity

Content-related validity is given when the items represent the construct of interest. Content-related validity is usually substantiated by means of argumentation, for example, when item content is systematically derived from the definition of a construct.

Factorial validity is given when a measurement instrument is based on a correctly specified measurement model which fits the nature of the construct and when its items reflect the underlying dimensions with sufficient reliability (sufficient factor loadings). Depending on the nature of the construct and the conception of the instrument, different measurement models might be suitable. Whatever the suggested measurement model is, test it using a factor-analytic approach.

Construct validity is given when an instrument reflects all facets of a theoretical construct. Two aspects of construct validity can be distinguished:

- **Convergent validity** exists when the scale complies with variables or scales that capture the same or a related construct, as indicated by high correlations of the scale with corresponding measurements.
- **Discriminant validity** presupposes a low or null correlation of the scale of interest with measurements that capture unlike concepts.

Different methods to assess the convergent and discriminant validity exist. EFAs or SEMs indicate whether the number and content of assumed sub-dimensions are supported by the data (see dimensionality). Additionally, nomological networks help to describe the relations between theoretical constructs and observed alternative tests.

Criterion validity is given when relevant criteria from outside the test situation (such as behaviour) can be predicted statistically by the scale score. Depending on the point of measurement, one distinguishes concurrent from predictive (or prognostic) validity:

- **Concurrent validity** is given when the external criterion is measured roughly at the same point in time as when the assessment with the instrument of interest takes place (e.g., comparison of a standardized student performance test to teachers' markings/grades).
- **Predictive validity** is given when the results of the instrument predict behaviour prospectively (e.g., prediction of occupational success by a vocational aptitude test).

A5: Measurement invariance levels

In a CFA/SEM, measurement invariance refers to the comparability of a measurement model via grouping characteristics of test persons. Parameters that refer to the covariance structure are factor loadings, residual variances, variance and covariance of latent variables. Parameters that refer to the mean structure are intercepts and latent means. Factor loadings, intercepts, and residual variances are part of the measurement model and variances, covariances, and latent means are part of the structural model. Measurement invariance evaluations successively test the similarity of the measurement model across different groups. An equivalence of the measurement model across different groups implies the measurement error-free comparability of corresponding structural parameters of the particular invariance level across those groups.

Configural invariance is given when the number of factors and the loading patterns (assignment of items to factors) are equivalent across all groups. You may assess configural invariance via a multigroup structural equation model where factor loadings and intercepts are allowed to vary. If configural invariance is given, the construct has a comparable dimensional structure across groups.

Metric invariance is accepted when the number of factors, the loading patterns, and the factor loadings are equivalent across all groups. Metric invariance can be assessed by specifying a metric invariance model with factor loadings constrained to equality across groups. If metric invariance is given, correlations of latent variables can be compared across groups.

Scalar invariance is achieved when the number of factors, the loading patterns, the factor loadings, and the item intercepts are equivalent across all groups. Scalar invariance can be assessed by specifying a scalar invariance model and constraining the factor loadings and intercepts across groups to be equal. If scalar invariance is given, the relationship between the items and the captured construct is equivalent across groups in the sense that item difficulties are similar enough for latent means to be comparable across groups.

Strict invariance is achieved when the number of factors, the loading patterns, the factor loadings, the item intercepts, and the residual variances are equivalent across all groups. Strict invariance can be tested using a strict invariant model in which the factor loadings, item intercepts, and residual variances are equated across all groups. When strict invariance is present, the residual variances are comparable across groups, which allows for comparison of manifest (item or scale) statistics such as the scale mean/sum or the variance of the mean/sum. In addition, strict measurement invariance can be used to exclude distortion of metric testing and scalar invariance levels themselves.

Procedure for testing: Configural invariance is the lowest level of measurement invariance. It is a necessary precondition for testing metric invariance, which itself is a prerequisite for testing scalar invariance. For the evaluation of the measurement invariance, you can calculate χ^2 difference tests between the invariance models and use the differences in the various fit indices (CFI, RMSEA, SRMR). Note that the χ^2 difference test is also dependent on the sample size, which is not the case with CFI, for example. To evaluate invariance the following guidelines can be used:

- **Configural invariance** is accepted when the measurement model (without additional constraints, such as equal loadings and intercepts) shows good fit across all groups. Cutoffs to decide on overall fit are stated in appendix A2. Rutkowski and Svetina (2014) looked at overall fit in the measurement invariance context comparing large numbers of groups (10 vs. 20) and replicated findings on overall fit in the single-group confirmatory factor analysis context (Hu & Bentler, 1999).
- **Metric invariance** compared the metric to the configural measurement model. It is rejected if the χ^2 difference test is significant and/or there exist differences of $\Delta CFI \leq -.010$ in combination with $\Delta RMSEA \geq .015$ or $\Delta SRMR \geq .030$. It is accomplished if $\Delta CFI > -.010$ in combination with $\Delta RMSEA < .015$ or $\Delta SRMR < .030$ (Chen, 2007).[1]
- **Scalar invariance** compares the scalar to the metric measurement model. It is rejected if the χ^2 difference test is significant and/or there exist differences of $\Delta CFI \leq -.010$ combined with $\Delta RMSEA \geq .015$ or $\Delta SRMR \geq .010$ according to Chen (2007).[2]
- **Strict invariance** between the scalar (or metric model if no mean structure is co-modelled) and strict model can be rejected if a χ^2 difference test between the two models is significant and/or there exist differences of $\Delta CFI \leq -.010$ in combination with $\Delta RMSEA \geq .015$ or $\Delta SRMR \geq .010$ (Chen, 2007)[3].

Chen's (2007) guidelines are often used for measurement invariance testing, although other guidelines based on simulations have been recommended by other authors (e.g., Meade et al., 2008 recommend $\Delta CFI \leq -.002$). The above guidelines for change in fit indices refer to the comparison of two groups. However, Rutkowski and Svetina (2014) derived similar guidelines for ten or twenty group comparisons (metric invariance: $\Delta CFI \leq -.020$, $\Delta RMSEA \geq .030$; scalar invariance: $\Delta CFI \leq -.010$, $\Delta RMSEA \geq .010$).

Partial invariance: If "full" measurement invariance of a measurement invariance level cannot be achieved, there is the possibility of testing "partial" measurement invariance. Using modification indices and expected parameter changes (EPCs), non-invariant parameters are freely estimated sequentially. Partial measurement invariance is present if the model has a (group) equivalent and a (group) specific part. If at least two of the parameters of one invariance level (e.g., two factor loadings or two intercepts) are equivalent across the groups, corresponding structural parameters of the latent variables can still be compared across the groups (Byrne et al., 1989) because the bias in the structural parameters is usually negligible (Pokorpek et al., 2019).

[1] The cutoffs refer to the following conditions: large samples (total $N > 300$), equal sample size of the groups, and mixed non-invariance. For small samples (total $N \leq 300$), unequal sample sizes of the groups and uniform non-invariance, $\Delta CFI \leq -.005$ in combination with $\Delta RMSEA \geq .010$ or $\Delta SRMR \geq .025$ are recommended as guidelines for the rejection of metric invariance (Chen, 2007).

[2] The cutoffs refer to conditions of equal sample size across groups, large total sample size (total $N > 300$), and mixed non-invariance. When sample size is unequal across groups or rather small (total $N \leq 300$) and when non-invariance is uniform, $\Delta CFI > -.005$ in combination with $\Delta RMSEA < .010$ or $\Delta SRMR < .005$ indicate scalar invariance (Chen, 2007).

[3] The cutoffs refer to the following conditions: large samples (total $N > 300$), equal sample size of the groups, and mixed non-invariance. For small samples (total $N \leq 300$), unequal sample sizes of the groups, and uniform non-invariance, $\Delta CFI \leq -.005$ in combination with $\Delta RMSEA \geq .010$ or $\Delta SRMR \geq .005$ are recommended as guidelines for the rejection of strict invariance (Chen, 2007).

Note that the above guidelines refer to ML estimates. Guidelines for changes in fit indices when using the WLSMV estimator are not yet established.

A6: Norming sample

The sample used is a random sample or a stratified random sample (see Gabler & Quatember, 2013). All units of the population should have a computable, non-zero, and positive selection probability. This is a basic requirement for the sample to represent the population adequately with regard to the characteristic: “A sample (or a sample result) is representative of a population with regard to a distribution of interest or a parameter characterising this distribution if this distribution or the parameter can be estimated without distortion (at least approximately) and if the desired accuracy is maintained in this estimate” (Gabler & Quatember, 2013, p. 107; translated by the authors). In the case of non-response, this must be stated and analysed (e.g., with R-indicators).
When preparing the standard, the distribution of the data (raw values) must be considered. In principle, the preparation of all standards is permissible for normally distributed data. Common representations of norms include percentile ranks (PR)/percentiles (PC), normalized standard scores, and standard scores. In the case of non-normally distributed data, only specify the percentile rank and standards generated from this rank (e.g., T standard, stanine).
In the case of nominal and ordinal data, only the specification of the percentage rank standard is permissible. The standard used must correspond to the ability to differentiate, as can be seen from the description of the content. In addition, a standard that corresponds to the expertise of the user group must be used. In addition, please give a brief justification of the content and/or statistics concerning whether standard differentiation is necessary. If so, you can provide separate standards in tables for subgroups. In this case, the sample size of the subgroups, the significance of the differences between the subgroups, and the effect size of the differences must be indicated and described when drawing up the standards. Please provide information on how to decide which group of standards is to be used in which case and explain the effects of applying these group-specific standard values. For cross-cultural studies, you can provide separate tables of norm values for each country.
Depending on the degree of testing of a measurement instrument, we assign it to a "development status". The development status is divided into two categories: tried or validated.

- An instrument is considered **"tried"** if its reliability (if to be assessed) has been evaluated as satisfactory, but there is less evidence for construct and criterion validation. This means that in the scale development process, only the first step of item selection was carried out. After the final item selection, the validation of the construct is usually performed using one or more additional samples. If this step is completely missing or only isolated evidence of construct validity (e.g., only convergent or only discriminant validity but not both) and/or criterion validity (e.g., a criterion with weak content) is available, the instrument is classified as "tried."
- An instrument is considered **"validated"** if the reliability (if to be assessed) has been evaluated as satisfactory and the validation of the instrument has been carried out on one or more further samples after the final item selection. There is sufficient evidence of construct and/or criterion validity.

An instrument is marked as **"standardized"** if one or more samples exist that are random samples or stratified random samples. All units of the population should have a computable, non-zero, and positive selection probability. The samples have been described, and information on the distribution of the scale values has been provided.