Reliability - the Precision of a Measurement (Version 2.0)

Danner, Daniel

Erstveröffentlichung / Primary Publication
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Danner, D. (2016). *Reliability - the Precision of a Measurement (Version 2.0).* (GESIS Survey Guidelines). Mannheim: GESIS - Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.15465/gesis-sg_en_011

gesis Leibniz-Institut für Sozialwissenschaften

GESIS Survey Guidelines

Reliability – The Precision of a Measurement

Daniel Danner

December 2016, Version 2.0

Abstract

Reliability describes the precision of a measurement. The present contribution begins by defining the concept of reliability and explaining why the reliability of a measurement is relevant. It then discusses the model assumptions that must be made in order to estimate the reliability of a measurement and presents five methods of estimating reliability: the test-retest method, the parallel-forms method, the split-half method, the internal consistency method, and the estimation of reliability using structural equation modelling. The contribution concludes with a brief outline of the commonalities and differences between classical test theory and item response theory and the importance of these theories for the estimation of reliability.

Citation

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

1. What is reliability?

Individual differences are frequently measured in the social sciences. The German General Social Survey (ALLBUS; e.g., Koch & Wasmer, 2004) and the European Social Survey (ESS; European Social Survey, 2014), for example, regularly measure attitudes to political and social issues. Some surveys, such as the International Social Survey Programme (Haller, Jowell, & Smith, 2009), also measure personality traits of the respondents. And studies such as the Programme for International Student Assessment (PISA; OECD, 2011), the Programme for the International Assessment of Adult Competencies (PIAAC; Rammstedt, 2013), and the German National Educational Panel Study (NEPS; Blossfeld, von Maurice, & Schneider, 2011) measure participants' cognitive abilities.

Reliability describes the precision of measurements such as these. From a formal point of view, the reliability coefficient, *R*, is the ratio of true score differences, τ, to observed score differences, *Y*:

$$R = \frac{Variance\ (\tau)}{Variance\ (Y)}$$

True score differences are systematic differences between individual personalities, attitudes, or abilities. Observed score differences may also be influenced by unsystematic factors such as situational disturbances or random measurement errors.

The aim of a measurement is to capture the true differences between individuals. This succeeds when the reliability of a measurement is high. A high level of reliability means that a large proportion of the observed differences is attributable to true differences. A low level of reliability, on the other hand, means that the observed differences are significantly "contaminated" by measurement errors. There is no binding threshold above which the estimated reliability of a measurement is adequate. A reliability of 0.70 is often considered adequate for group studies (Rammstedt, 2004), 0.80 is generally described as good (Nunnally & Bernstein, 1994; Weise, 1975), and a reliability coefficient of over 0.90 is deemed to be high (Weise, 1975).

Reliability is always a property of a measurement rather than of a measurement instrument. An instrument may yield measurements of different levels of reliability in different samples. In a very homogeneous sample, in which there are hardly any true differences between individuals, reliability may be lower than in a heterogeneous sample, in which there are significant interindividual differences. For example, when the same instrument is used to measure political attitudes in extreme groups and in a heterogeneous sample, the measurement carried out in the extreme groups may be less reliable. In practice, therefore, researchers often endeavour to estimate reliability on a representative sample as it can then be assumed that the reliability in the population is comparable.

2. Why is the reliability of a measurement relevant?

The reliability of a measurement is relevant when relationships between different variables are examined or when a single individual's value is the focus of interest.

Many research questions address the relationship between different constructs. In the political sciences, researchers study the relationship between attitudes and voting behaviour, for example (e.g., Wüst, 2002); in psychology, they examine the relationship between cognitive abilities and occupational

success (e.g., Schmitt & Hunter, 2004) and between personality and behaviour (e.g., Hossiep & Mühlhaus, 2005). Correlations are often used to assess the strength of these relationships. However, the reliability of a measurement limits the correlation that can be measured between two variables: If reliability is high, the maximum correlation is also high; if reliability is low, so too is the maximum correlation. This can be quantified as follows: The maximum correlation (r_{max}) between one variable and another variable is the square root of their reliability (R):

$$r_{max} = \sqrt{R}$$

For example, if attitudes to politicians are measured, and the reliability of the measurement is $R = 0.90$, the maximum correlation between the measured attitude and another variable is $\sqrt{0.90} = 0.95$. If the reliability is only 0.50, the maximum correlation between the measured attitude and another variable is $\sqrt{0.50} = 0.70$.

The reliability of a measurement is also of relevance when a single individual's value is considered. A precise measurement with a high reliability enables an individual's value to be precisely estimated. The true value and the observed value are then very similar. An imprecise measurement with a low reliability enables only an imprecise estimation. The true value and the observed value may then differ significantly. The difference between the true value and the observed value can be quantified on the basis of the confidence interval of a measurement. The confidence interval describes the range within which an individual's true value falls when the observed value, *Y*, the reliability of the measurement, *R*, and the standard deviation, *SD*, of the test are known. The 95% confidence interval, *CI*, of a measurement can be estimated using the following formula:

$$CI = Y \pm 1.96 * SD * \sqrt{1 - R}$$

For example: A person takes an intelligence test and the result of the measurement is an intelligence quotient (IQ) of 111. The reliability of the measurement, *R*, is 0.90, the standard deviation, *SD*, is 15. Hence, the 95% confidence interval ranges from $111 - 1.96 * 15 * \sqrt{1 - 0.90} = 102$ to $111 + 1.96 * 15 * \sqrt{1 - 0.90} = 120$. If the reliability of the test was only 0.50, the measurement would be less precise, the confidence interval would be wider and would range from $111 - 1.96 * 15 * \sqrt{1 - 0.50} = 90$ to $111 + 1.96 * 15 * \sqrt{1 - 0.50} = 132$.

3. What model assumptions must be made in order to estimate reliability?

The reliability of a measurement describes the ratio of true value differences, τ, to observed value differences, *Y*. The true value of a measurement cannot be observed. However, the variance of the true values can be estimated when certain model assumptions are made. The starting point for various measurement models is classical test theory, which essentially states that an observed value, *Y*, is composed of a true value, τ, and a measurement error, ε (Bühner, 2011; Lord & Novick, 1968; Steyer & Eid, 2001):

$$Y = \tau + \varepsilon$$

A number of different measurement models can be distinguished within classical test theory:

3.1 The parallel measurement model

The most parsimonious measurement model is the parallel measurement model, which specifies that an observed value, *Y*, is composed of a true value, τ, and a measurement error, ε. The model specifies,

further, that several measurements, *i*, (of items or tests) have the same true value, τ_i ($\tau := \tau_i$), and that the error variances of several measurements are identical ($s^2_\varepsilon := s^2_{\varepsilon_i}$). An example of a parallel measurement model with two measurements is shown in Figure 1. Although this measurement model is parsimonious, it makes the restrictive assumption that the variances of several measurements are identical. However, this assumption must not be satisfied.

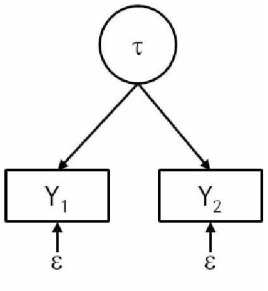

Figure 1: Parallel measurement model with two measurements

3.2 The tau-equivalent measurement model

The tau-equivalent measurement model is less restrictive than the parallel measurement model. It specifies that several measurements, *i*, have the same true value, τ_i ($\tau := \tau_i$), but that the variances of the measurement errors ($s^2_{\varepsilon_i}$) may differ. A tau-equivalent measurement model has several parameters that must be estimated. At least three measurements are therefore necessary to estimate these model parameters. An example with three measurements is shown in Figure 2. This model is less restrictive because it allows the measurement errors to differ in size across different measurements. However, it requires that the true value of different measurements be identical. This assumption, too, may be violated.

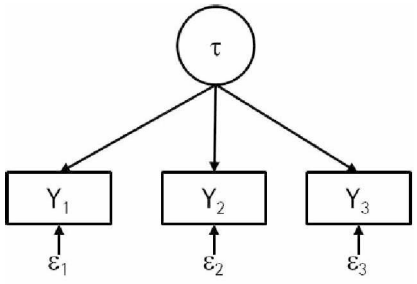

Figure 2. Tau-equivalent measurement model with three measurements

3.3 The tau-congeneric measurement model

The tau-congeneric measurement model is the least restrictive of the models presented here. It specifies that the variances of the measurement errors ($s^2_{\varepsilon_i}$) may differ and that the true values of several measurements are linear functions of each other ($\tau := \lambda_i * \tau_i$). Two true values are linear functions of each other if one value can be transformed into the other value through multiplication (e.g., $\tau_2 = 0.75 * \tau_1$). A tau-congeneric measurement model has more parameters than a tau-equivalent measurement

model. Therefore, at least four measurements are necessary to estimate the model parameters. An example with four measurements is shown in Figure 3. This model allows different measurements to reflect the true value to a different extent. It also allows the extent of the influence of measurement errors to differ in different measurements.

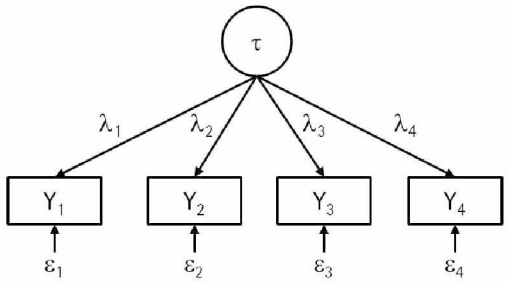

Figure 3. Tau-congeneric measurement model with four measurements

The various measurement models enable different methods of estimating the reliability of a measurement to be applied. Structural equation models (e.g., Bollen, 1989; Eid, Gollwitzer, & Schmitt, 2010; Tabachnick & Fidell, 2013) can be used to test which measurement model applies.

4. How can the reliability of a measurement be estimated?

Various methods can be used to estimate the reliability of a measurement. In what follows, the most commonly used estimation methods are presented: the test-retest method, the parallel-forms method, the split-half method, the internal consistency method, and the estimation of reliability using structural equation modelling.

4.1 The test-retest method

The test-retest method can be used when an instrument is administered to the same person on two separate occasions. The reliability, R, of the two measurements, Y_1 and Y_2, can then be estimated on the basis of the correlation, r, between them:

$$R = r_{Y_1,Y_2}$$

The test-retest method yields a reliable estimate of reliability when the parallel measurement model applies. That means that the same true value must be measured on both measurement occasions. In practice, it means that the true value may not change in the interval between the two measurements. This assumption is plausible in the case of constructs that remain stable over time (e.g., intelligence or extraversion). In the case of constructs that may change over time (e.g., moods or attitudes), the assumption is less plausible. If the test-retest method is used, it must be assumed that there are no practice- or recollection effects between the measurements. The second assumption that must be made is that the variance of the measurement error is identical in the case of both measurements. In practice, this means that interference effects, such as noise during the administration of the instrument, or tiredness on the part of the test person, are the same for both measurements. These assumptions can be tested using structural equation modelling (e.g., Bollen, 1989; Eid, Gollwitzer & Schmitt, 2010;

Tabachnick & Fidell, 2013). An example of such a structural equation model is shown in Figure 1. In practice, however, such a test is frequently omitted.

4.2 The parallel-forms method

The parallel-forms method can be used when two parallel forms of an instrument are available, and both forms are administered to the same test persons. Parallel forms of an instrument measure the exact same true value and are influenced to the same extent by measurement errors. Examples of parallel forms are the A and C forms of the intelligence structure test (Liepmann, Beauducel, Brocke, & Amthauer, 2007) and the two forms of the vocational aptitude tests developed by Schmale (2001). If two parallel forms are available, the reliability, R, of the two measurements, Y_1 and Y_2, can be estimated on the basis of the correlation, r, between them:

$$R = r_{Y_1,Y_2}$$

The parallel-forms method yields a reliable estimate of reliability when the parallel measurement model applies. This means that both parallel forms must measure the exact same value and must be influenced to the same extent by random measurement errors. In practice, it is often difficult to construct parallel versions of an instrument because different items often capture different aspects of a construct. The fact that two instruments allow parallel measurements must therefore be well substantiated. As shown in Figure 1, the parallelism of two measurements can also be tested by using a structural equation model.

4.3 The split-half method

The split-half method can be used to estimate reliability when an instrument is administered only once. This method entails splitting the instrument into two halves; it is assumed that both halves will provide a parallel measurement. The split-half method is frequently applied when an instrument comprises several items. In this case, the instrument is split into two equal parts. This can be done in several ways. It is customary either to divide the items according to even and odd item numbers (odd-even split), to divide them into a first and a second test half, to divide them according to item characteristics (statistical twins method), or to divide them randomly. The correlation between the two test halves is used to estimate the reliability of the halves. In order to obtain an estimate of the reliability of the instrument as a whole, this estimate is then corrected by applying the Spearman-Brown formula (see also Amelang & Schmidt-Atzert, 2006; Bühner, 2011). The reliability, R, of a measurement, Y, that comprises the scores for the two test halves (Y_1,Y_2) can then be estimated using the following formula:

$$R = \frac{2 * r_{Y_1,Y_2}}{1 + r_{Y_1,Y_2}}$$

The Spearman-Brown correction does not have to be carried out manually, because statistical packages such as SPSS automatically provide the corrected reliability estimate, which is known as the Spearman-Brown coefficient (SPSS).

The split-half correlation can also be computed by applying the maximal split-half coefficient method (e.g., Callendar & Osburn, 1979; Hunt & Bentler, 2012). This method is based on a proposal by Guttman (1945) to split tests into all possible halves (according to even and odd item numbers) and to compute all the correlations. If the test halves are parallel, the highest correlation will offer the best estimate of the reliability of the overall test. One practical problem with this approach is that dividing the instrument into all possible test halves is very computation-intensive. In the case of a ten-item instrument, there are $2^{10-1} - 1 = 511$ possible combinations; a 25-item instrument yields $2^{25-1} - 1 =$

16,777,215 possible combinations. For this reason, Hunt and Bentler (2012) suggested creating a sample of possible combinations (e.g., 10,000 combinations) and estimating reliability on the basis of that sample. The method is described in detail in Hunt and Bentler (2012). The R package 'Lambda4' (Hunt, 2013) can be used to estimate reliability on the basis of the maximal split-half coefficient.

Irrespective of the way in which the test halves are formed, the split-half method presupposes that both halves allow parallel measurements. This assumption can be tested with a structural equation model, as shown in Figure 1 above.

4.4 The internal consistency method

The internal consistency method enables reliability to be estimated when an instrument is administered only once. The first step in estimating reliability on the basis of internal consistency is to divide the test up into individual items. Reliability can then be estimated on the basis of the variances (s^2) of the items, i, and of the variance of the sum of the items. Cronbach's alpha is the coefficient most commonly used to estimate internal consistency (e.g., Amelang & Schmidt-Atzert, 2006; Bühner, 2011). The reliability, R, of a measurement, Y, that comprises k items can then be estimated using the following formula:

$$R = \frac{k}{k-1} * \left(1 - \frac{\sum_{i=1}^{k} s_i^2}{s_Y^2}\right)$$

As a rule, it is not necessary to manually compute Cronbach's alpha as it can be automatically computed with statistical software such as SPSS, SAS, or STATA.

Cronbach's alpha is a reliable estimator of the reliability of a measurement when the items are tau-equivalent – that is, when all the items reflect the same true value. If this precondition is violated, Cronbach's alpha underestimates reliability and is thus only a lower-bound estimate of reliability (Cortina, 1993; Lord & Novick, 1968). In practice, this means that the reliability of a measurement may be higher than Cronbach's alpha. Moreover, a tau-equivalent measurement model requires that the covariance of the items is due only to their true values. The measurement errors of the individual items may not covary. In other words, the items must display a unidimensional structure – that is, individuals' true values are the only factor that explains the covariance between the items. If this assumption is violated, Cronbach's alpha is not an unbiased estimator of reliability. Structural equation modelling can be used to test whether the items of an instrument are tau-equivalent.

4.5 Structural equation modelling

If measurements are neither parallel nor tau-equivalent but rather only tau-congeneric, structural equation modelling can be used to estimate their reliability. The advantage of this method is that reliability can be estimated and the underlying model assumptions can be tested at the same time. The estimation of composite reliability following Raykov (1997) is presented in what follows. In order to estimate the reliability of tau-congeneric measurements with structural equation models, the instrument in question must comprise at least four items. Moreover, structural equation models require relatively large samples – at least $N = 200$ persons (Hoyle, 1995). If these preconditions are satisfied, a structural equation model as shown in Figure 4 can be used to estimate reliability. Besides the measurement model, this structural equation model contains a phantom variable, M, that corresponds to the sum or the mean of the items. This phantom variable facilitates the estimation of reliability.

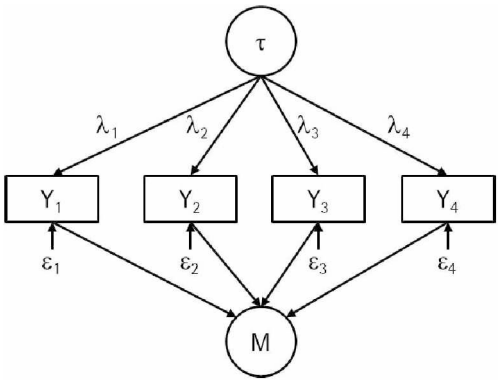

Figure 4. Tau-congeneric measurement model with four items and one phantom variable

The reliability of this sum or mean can then be estimated using the model parameters. To this end, the variance of the latent variable, τ, is weighted with the loads, λ_i, of the items and divided by the variance, s^2, of the phantom variable, M:

$$R = \frac{\sum(\lambda_i)^2 * s_\tau^2}{s_M^2}$$

The parameters of a structural equation model can be estimated with statistical programmes such as Amos, Mplus, SAS, or R. The estimated parameters can then be entered into the formula. The procedure is described in detail in Raykov (1997).

The composite reliability method (Raykov, 1997) yields a reliable estimate of reliability when the tau-congeneric measurement model applies (and even when the tau-equivalent or the parallel measurement model applies). This means that the relationships between the items may be explained only by their true values, and the measurement errors of the items may not correlate. The goodness of fit of the structural equation model can be used to test whether a tau-congeneric measurement model applies. The most commonly used fit indices are the root mean square of approximation (RMSEA) and the comparative fit index (CFI). A RMSEA < .06 and a CFI > .95 indicate good model fit (Hu & Bentler, 1998).

4.6 Comparison of different methods

Different methods can be used to estimate the reliability of a measurement. Which method is the most suitable depends on which measurement model can be assumed. The test-retest method can be used to estimate reliability if an instrument is administered on two separate occasions and if it can be assumed that both measurements are parallel – that is, that they reflect the same true value and have the same error variance. If two parallel measurement instruments are available, reliability can be estimated using the parallel test method. And if an instrument can be split into two parallel halves, the split-half method of estimating reliability can be applied. If the items of an instrument reflect the same true value (i.e., are tau-equivalent), the internal consistency method can be used to estimate their internal consistency. In many cases in which the items of a test are not tau-equivalent but rather tau-congeneric, reliability can be estimated using structural equation modelling. The various estimation methods make different assumptions, which is why the estimates that they yield may differ. However, this does not mean that a measurement has different reliabilities. Rather, it has only one reliability. Therefore, the estimation method that is most suitable for the data in question should always be employed. Structural equation modelling can be used to test which measurement model applies and

which method is the most suitable (e.g., Bollen, 1989; Eid, Gollwitzer, & Schmitt, 2010; Tabachnick & Fidell, 2013).

5. Can the reliability of a measurement be estimated even if an instrument was developed using item response theory?

Yes, it can. Classical test theory and item response theory hail from different research traditions. The two theories offer different perspectives on the quality of a measurement. However, these perspectives are not contradictory, but rather they complement each other.

Classical test theory describes an observed value, Y, as a combination of a true value, τ, and a measurement error, ε

$$Y = \tau + \varepsilon$$

As a rule, the observed value is the sum or the mean of a scale. This value is treated as an interval-scaled, normally distributed variable. The true value, τ, describes a personal characteristic, an ability, or an attitude. The precision of a measurement can then be determined on the basis of the reliability coefficient, which is defined as the ratio of true variance to observed variance.

Item response theory does not describe how an observed value is composed but rather the probability that a certain value will be observed. In the most simple item response theory measurement model, the Rasch model, this probability, P, depends on the difficulty of the item, σ, and the ability of the person, θ:

$$P(Y = 1|\theta, \sigma) = \frac{\exp(\theta - \sigma)}{1 + \exp(\theta - \sigma)}$$

The observed value is a response category of an item (e.g., $Y = 1$). This value is treated as a categorical (in the simplest case, dichotomous) variable. The item difficulty parameter, σ, describes the characteristic of an item. The person parameter, θ, describes a characteristic of the individual (e.g., an ability or an attitude). This parameter can be estimated on the basis of the observed data, and the precision of a measurement can be determined on the basis of the standard measurement error of this estimate. One special feature of item response theory is that the size of the standard measurement error may differ in different regions on the ability continuum. In practice, this means that the precision of a measurement may be greater around the middle of the ability continuum than in the extreme regions. The estimation of the person parameter is described in detail in Embretson and Reise (2000), for example.

Item response theory thus enables a differentiated estimate of the precision of a measurement to be made. However, as researchers are often more interested in the precision of several measurements in a sample than in the precision of the measurement of an individual person's score, the various software packages also compute the average variance extracted. The average variance extracted describes the average proportion of the variance of the manifest variable that can be explained by the person variable.

Even if a Rasch model was used to scale the instrument, "classical" methods, such as the test-retest method or the split-half method, can also be applied to estimate the reliability of the sum or the mean of the items.

6. References

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4th ed.). Heidelberg: Springer.

Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). *Grundidee, Konzeption und Design des Nationalen Bildungspanels für Deutschland* (NEPS Working Paper No. 1). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Bollen, K. A. (1989). *Structural equations with latent variables*. Oxford: John Wiley & Sons.

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

Callendar, J. C. & Osburn, H. G. (1979). An empirical comparison of coefficient alpha, Guttman's lambda – 2, and MSPLIT maximized split-half reliability estimates. *Journal of Educational Measurement, 16*, 89-99. doi: 10.1111/j.1745-3984.1979.tb00090.x

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

Eid, M., Gollwitzer, M., & Schmitt, M. (2010). *Statistik und Forschungsmethoden*. Weinheim: Belz.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

European Social Survey (2014). *ESS Round 6 (2012/2013)* Technical Report. London: Centre for Comparative Social Surveys, City University London.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.

Haller, M., Jowell, R., & Smith, T. W. (2009). *The International Social Survey Programme*. New York: Routledge.

Hoyle, R. H. (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA US: Sage Publications, Inc.

Hossiep, R. & Mühlhaus, O. (2005). *Personalauswahl und -entwicklung mit Persönlichkeitstests*. Göttingen: Hogrefe.

Hu, L.-t. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3(4)*, 424-453. doi: 10.1037/1082-989x.3.4.424

Hunt, T. D. & Bentler, P. (2012). *Quantile Lower Bounds to Reliability Based on Splits*. Retrieved from the University of California, Los Angeles Website: http://statistics.ucla.edu/preprints/uclastat-preprint-2012:5

Hunt, T. (2013). *Package 'Lambda4'*. Retrieved from http://cran.r-project.org/web/packages/Lambda4/Lambda4.pdf

Koch, A. & Wasmer, M. (2004). Der ALLBUS als Instrument zur Untersuchung sozialen Wandels: Eine Zwischenbilanz nach 20 Jahren. In: R. Schmitt-Beck, M. Wasmer, & A. Koch (Eds.): *Sozialer und politischer Wandel in Deutschland. Analysen mit ALLBUS-Daten aus zwei Jahrzehnten*. 2004, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 13-41, Blickpunkt Gesellschaft, Band 7.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *I-S-T 2000 R - Intelligenz-Struktur-Test 2000 R* (2nd ed.). Göttingen: Hogrefe.

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

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.

OECD (2011). *PISA 2009 Results: Overcoming Social Background – Equity in Learning Opportunities and Outcomes* (*Volume II*). doi: 10.1787/9789264091504-en

Rammstedt, B. (2004). *Zur Bestimmung der Güte von Multi-Item-Skalen: Eine Einführung* (ZUMA How-to Series No. 12). Mannheim: ZUMA.

Rammstedt, B. (Ed.). (2014). *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich: Ergebnisse von PIAAC 2012*. Münster: Waxmann.

Raykov, T. (1997). Estimation of Composite Reliability for Congeneric Measures. *Applied Psychological Measurement, 21*, 173-184. doi: 10.1177/01466216970212006

Schmale, H. (2001). *Berufseignungstest (BET). Tabellenband* (4th revised and enlarged ed.). Bern: Hans Huber.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*, 162-173. doi: 10.1037/0022-3514.86.1.162

Steyer, R. & Eid, M. (2001). Messen und Testen. Berlin: Springer.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Boston: Allyn and Bacon.

Weise, G. (1975). *Psychologische Leistungstests*. Göttingen: Hogrefe.

Wüst, A. (2002). *Wie wählen Neubürger? Politische Einstellungen und Wahlverhalten eingebürgerter Personen in Deutschland*. Opladen: Leske+Budrich.