# Current Challenges, New Developments, and Future Directions in Scale Construction

Danner, Daniel; Blasius, Jörg; Breyer, Bianka; Eifler, Stefanie; Menold, Natalja; Paulhus, Delroy L.; Rammstedt, Beatrice; Roberts, Richard D.; Schmitt, Manfred; Ziegler, Matthias



**Empfohlene Zitierung / Suggested Citation:**
Danner, D., Blasius, J., Breyer, B., Eifler, S., Menold, N., Paulhus, D. L., ... Ziegler, M. (2016). Current Challenges, New Developments, and Future Directions in Scale Construction. *European Journal of Psychological Assessment*, *32*(3), 175-180. https://doi.org/10.1027/1015-5759/a000375

This is the postprint of the article originally published in
**European Journal of Psychological Assessment, Vol. 32, No. 3, © 2016 by Hogrefe**
available online at: https://doi.org/10.1027/1015-5759/a000375
This version of the article may not completely replicate the final version published in European Journal of Psychological Assessment. It is not the version of record and is therefore not suitable for citation.

# Current Challenges, New Developments, and Future Directions in Scale Construction

**Daniel Danner,[1] Jörg Blasius,[2] Bianka Breyer,[1] Stefanie Eifler,[3] Natalja Menold,[1] Delroy L. Paulhus,[4] Beatrice Rammstedt,[1] Richard D. Roberts,[5] Manfred Schmitt,[6] and Matthias Ziegler[7]**

[1]GESIS – Leibniz-Institute for the Social Sciences, Survey Design and Methodology, Mannheim, Germany, [2]University of Bonn, Institut für Politische Wissenschaft und Soziologie, Germany, [3]Catholic University of Eichstätt-Ingolstadt, Sociology Institute, Germany, [4]University of British Columbia, Department of Psychology, Vancouver, Canada, [5]Center for Innovative Assessments, Professional Examination Service, New York, NY, USA, [6]University of Koblenz-Landau, Fachbereich Psychologie, Germany, [7]Institut für Psychologie, Humboldt Universität zu Berlin, Germany

Measurement instruments are the foundation for empirical research in the social sciences. Instruments are necessary for measuring latent constructs such as cognitive and non-cognitive skills in the Programme for the International Assessment of Adult Competencies (PIAAC), personality characteristics in studies such as the International Social Survey Program (ISSP), or attitudes in international studies such as the European Social Survey (ESS). Measurement instruments also allow researchers, practitioners, and policy makers to describe individuals, groups, or societies, to assess patients in clinical settings, or to select, classify, or assist in the remediation of workers and students. Many policy, research, and applied decisions depend on measurement instruments and the quality of these decisions depends on the quality of the instruments, which is closely entwined with the scale development process (Ziegler, 2014). The aim of this editorial is to describe challenges and new developments in scale construction and discuss how they can facilitate the quality of measurement instruments.

Developing measurement instruments is a multi-step approach (American Educational Research Association, 2014; Rammstedt et al., 2015; Ziegler, 2014). First, the construct, the intended use of the instrument, and the targeted population should be defined and documented (e.g., in a test blueprint). Subsequently, items are generated, and selected, and finally psychometric qualities such as reliability, validity, and fairness of score interpretations resulting from the measures are evaluated. If necessary, norms are provided and then, the development process is documented. Each step in the scale construction process helps to shape and sharpen the instrument but also brings challenges that can compromise the psychometric properties of the instrument in question. Usually, we cannot optimize reliability, validity, and fairness to the same extent, because the most reliable items may not yield the most valid or fairest test score interpretations. However, the intended use of the measure should determine which of the quality aspects is most important. In most cases, the validity of a test score interpretation will be seen as most important. In some cases, the reliability of a test score interpretation might be particularly important for individual assessment whereas cross-cultural fairness might be particularly important for international studies such as PIAAC or the ESS. The way we address challenges in scale construction and the way we generate and select items will facilitate reliability, validity, and fairness in different ways. Subsequently, we will discuss response styles, appropriate reliability estimation, and measurement invariance as three key challenges in scale construction.

## Current Challenges Response Styles

Response styles such as acquiescence (Paulhus, 1991), extreme responding (Baumgartner & Steenkamp, 2001), or faking (Ziegler, MacCann, & Roberts, 2011) are answering tendencies that manifest in the items but are independent of the construct to be measured. Thus, response styles can threaten the validity of measures and bias mean scores, correlations, and the factor structure of personality inventories (e.g., Danner, Aichholzer, & Rammstedt, 2015; Rammstedt & Farmer, 2013; Wetzel, Böhnke, Carstensen, Ziegler, & Ostendorf, 2013; Ziegler, 2015). Response styles can be triggered by the ways items are generated. For example, generating items based on a scientific theory can lead to complex formulations which, in turn, can facilitate acquiescent responding (McBride & Moran, 1967; Trott & Jackson, 1967), especially in samples with heterogeneous educational background (Rammstedt & Farmer, 2013) or heterogeneous age (Soto, John, Gosling, & Potter, 2008). Likewise, ambiguous response categories can trigger extreme or midpoint responding (Moors, Kieruj, & Vermunt, 2014; Weijters, Cabooter, & Schillewaert, 2010). Hence, controlling response styles by optimizing the way we generate items or by changing the way we design response scales is one challenge for future scale construction. We will discuss cognitive pretests and anchoring vignettes as two promising avenues that can help to address these issues.

## Finding the Appropriate Reliability Estimation Method

Reliable interpretations of scores reflect systematic individual differences and allow describing individuals accurately

and investigating the association with other variables. There are multiple ways to estimate the reliability of a measure like the items' internal consistency, the splithalf method, or the test-retest method. These different methods make different assumptions about the underlying measurement model (Graham, 2006) and the way we generate and select items influences which method is best to use. For instance, items that have been selected based on item-total correlations or factor loadings are typically homogenous and suitable to measure unidimensional, narrow constructs while maintaining the assumption of $\tau$-equivalence. In such instances, estimates for the items' internal consistency, like Cronbach's $\alpha$, McDonald's $\Omega$ (McDonald, 1999), or Raykov's $\rho$ (Raykov, 1997), can be suitable estimation methods. However, items that have been selected based on expert ratings or correlations with external criteria tend to be more heterogeneous or even multidimensional and in such instances methods like split-half correlations or test-retest correlations can be more appropriate. The same is true for scores representing heterogeneous constructs. The underlying items need to be as heterogeneous as the construct itself. Here, a large number of items are needed in order to obtain a satisfactory internal consistency. Other methods might yield much better estimates with fewer items needed. However, no method is a silver bullet: using split-half correlations requires that the test halves are parallel and using test-retest correlations requires that the measured construct is stable over time. The challenge is in finding the reliability estimation method that fits best with the underlying measurement model and the intended use of a measure. For example, measures used to make prognoses should provide a stability estimate like test-retest correlations. We will discuss that incorporating the fit of the underlying measurement model in the item selection process can help to master that challenge and find the appropriate reliability estimation method.

**Ensuring Measurement Invariance**

Meaningful comparisons between groups can only be made if measured scores have the same meaning across groups (Chen, 2008). Especially in studies such as PIAAC or the ESS, between-country comparisons can only be made if the measured construct and the items have the same meaning across countries. The foundation for measurement invariance (Vandenberg & Lance, 2000) is laid when items are generated. For example, items that are generated based on prototypes (Buss & Craik, 1980) can only be as invariant as the prototype itself. The rationale of using prototypes is that persons can be grouped into different categories and there are persons that are perceived to be typical for a certain category. However, even prototypes such as Donald Trump who may be seen as a prototypical narcissist by most people may be perceived differently by others and hence, items that are generated based on such a prototype may not be measurement invariant across different groups. The same argument applies for other item generation approaches such as theory-based approaches where items are generated according to a scientific model or lexical approaches where items are generated based on extensive literature reviews. The resulting measures will only be as invariant across groups or cultures as the theory, the model, or the literature review itself (e.g., Thalmayer & Saucier, 2014). Hence, an additional challenge for future scale construction is in incorporating steps to ensure measurement invariance of items in an early stage of item generation or item selection. Selecting items based on cognitive pretests (see below) can be one way to address this issue.

In sum, traditional approaches for generating and selecting items do not explicitly address the problem of response styles, the question of how the reliability can be estimated best, or how to generate and select measurement invariant items. However, there have been some innovative developments in scale construction that can help to close these gaps.

## New Developments

### Improving Items With Cognitive Pretests

As discussed before, item responding might be subject to several methodological problems such as response styles or lacking measurement invariance. Next, academic language used in item formulations might be particularly prone to such problems. Methodological and language issues of item responding can be addressed by means of cognitive pretests (e.g., Beatty & Willis, 2007; Willis, 2004). Cognitive pretests have been conducted in the social science research to identify respondents' cognitive burden while understanding the question, retrieving relevant information, obtaining a response, and providing the response onto the response options. The same technique has also been applied to investigate faking (Robie, Brown, & Beaty, 2007; Ziegler, 2011). Cognitive interviewing is a qualitative method, where according to the aim of the study a small number of persons (usually between 10 and 30) representing the target population are interviewed. During an interview, the interviewees are asked to reflect upon their response with the help of further questions or tasks. Probing, for example, is a cognitive pretesting technique directly investigating the interviewees' understanding of a specific term (e.g., citizen of the world, education system in Germany, a certain type of work hours) by asking what they were taking into consideration when hearing or reading this term. Another probing method concentrates on the appropriateness of response formats by asking the respondent why s/he endorsed a specific response option (see Lenzner, Neuert, & Otto, 2016 for a practical guideline). Identifying how items are understood by the respondents helps to increase the item quality and to reduce invariances across different subpopulations. For example, in the item „To be housewife is as fulfilling as to professionally work“ the term „fulfilling“ was found to be

understood by some respondents as „time demanding“ instead of the intended meaning of „satisfactory“ or „fair.“ In another study, respondents were asked to evaluate their ability to calculate percentages. Some interviewees thought of performing these calculations only in the mind and thus stated having serious problems while others thought of writing figures down or using calculators. Without a revision, the presented questions would have measured the target construct differently in different groups of respondents providing biased results with respect to the true value.

**Developing Items With Anchoring Vignettes**

Response styles or different frames of references can threaten the comparability of measures and validity of test score interpretations. Anchoring vignettes offer an innovative approach to control these biases (e.g., Chevalier & Fielding, 2011; Crane, Rissel, Greaves, & Gebel, 2016; King, Murray, Salomon, & Tandon, 2004). In a nutshell, individual responses are adjusted by means of an anchor generated within the context of the study. Typically, an anchoring vignette encompasses a hypothetical scenario, for example,

> „Daisy is not capable of dealing with one thing for a long time. She has started to learn an instrument several times, but after a few weeks of practicing she has quit. This has been the case also with many language courses. In the morning it is difficult for Daisy to wake up and therefore she is often late for work.“ (Mottus et al., 2012, p. 317).

Subsequently, the respondents are asked to judge the scenarios on a rating scale, for example, from „inept“ to „competent.“ The same rating scale is used for conventional self-report items and the responses to the anchoring vignettes are used to rescale the subjects' responses to conventional self-report items. Both nonparametric and parametric methods have been proposed for rescaling the subjects' responses (Bolt, Lu, & Kim, 2014). The approach is, however, explicitly or implicitly based on two assumptions. First, all respondents must perceive the concept of interest in the anchoring vignettes similarly (vignette equivalence; Jürges & Winter, 2013). Second, the answers respondents give when using a conventional measurement instrument must be comparable to the answers they give when using anchoring vignettes (response consistency; Au & Lorgelly, 2014). An additional burden is that the anchoring vignettes themselves have to be developed, which requires expertise, time, and resources. However, if these hurdles are overcome, anchoring vignettes can help to increase the reliability and validity of measures (e.g., Primi, Zanon, Santos, De Fruit, & John, 2016).

**Selecting Items Using Iterative Structural Equation Modeling**

Relatively new developments are iterative structural equation modeling approaches such as ant colony optimization (e.g., Janssen, Schultze, & Grötsch, 2015; Leite, Huang, & Marcoulides, 2008; Marcoulides & Drenzer, 2003; Olaru, Witthöft, & Wilhelm, 2015). The core of these methods is (a) specifying a measurement model, (b) evaluating the fit of a selection of items with that model, and (c) iteratively changing the selection of items while reevaluating model fit. This procedure is repeated until an optimal selection of items is found. The number of possible item combinations can be tremendous (e.g., there are $\frac{100!}{10!\times 90!}$ possible combinations if we want to select 10 out of 1OO items) and testing all combination could take very long. Thus ant colony optimization does not evaluate all possible combinations of items but favors items that have been shown to fit in the measurement model. In particular, ant colony optimization starts with a random selection of items and marks items with „pheromones“ (Leite et al., 2OO8, p. 415) if they fit with the measurement model. This procedure is repeated and in subsequent iterations items that received many pheromones are preferably selected.

The particular charm of such iterative approaches is that they are flexible and – in principle – can be used to explicitly address challenges like response styles, fit with measurement models, and measurement invariance. For example, items could not only be rewarded for the overall fit with the measurement model but also for their acquiescence specificity (Danner et al., 2015) or their measurement invariance across educational groups or cultures.

**Estimating Reliability With Structural Equation Models**

Probably still the most widely used estimator for reliability is Cronbach's α (Cronbach, 1951). This remains so despite the tremendous amount of criticism this coefficient had to endure (Gu, Little, & Kingston, 2013; Revelle & Zinbarg, 2009; Sijtsma, 2009; Yang & Green, 2011). Even Cronbach himself was doubtful regarding the usefulness of α (Cronbach & Shavelson, 2004). One of the major issues troubling Cronbach's α is the underlying assumption of τ-equivalent indicators (Graham, 2006). Violations of this assumption yield reliability estimates that are too small. However, Cronbach's α is not without an alternative. Ω (McDonald, 1999) and a variety of different Ω versions (Padilla & Divers, 2013; Raykov & Pohl, 2013; Revelle & Zinbarg, 2009; Rodriguez, Reise, & Haviland, 2016; Zhang & Yuan, 2016; Ziegler & Brunner, 2016) have been suggested. The big advantage of Ω is that it does not need τ-equivalent indicators for the latent variable. In fact, there are even versions of Ω which can be used for indicators with more than one underlying latent variable (Rodriguez et al., 2016). Thus, using Ω ensures that the reliability estimate for a specific test score representing a specific trait contains only the variance explained by this trait (Brunner & Süß, 2005). Moreover, the necessary information needed to estimate Ω is derived from the results of structural equation models. Thus, the assumed theoretical model explaining the data is actually tested before reliability is estimated (Ziegler & Hagemann, 2015). This way, Ω and its versions in a way combine evidence for the reliability and factorial validity of a score

interpretation.

**Conclusions**

Developing valid and reliable measurement instruments can be challenging, especially facing challenges such as response styles, complex measurement models, or required measurement invariance. New approaches such as anchoring vignettes, cognitive pretests, or iterative structural equation models address these challenges and can improve the quality of measurement instruments and psychological and social research in general. Moreover, changes in the way we estimate coefficients representing the psychometric properties of a measure such as Ω should not be regarded as short-lived eccentricities but rather as potential lifelines out of psychometric dead ends. At the same time, such new developments should also raise new questions. For example, whether anchoring vignettes are understood equivalently by all respondents (vignette equivalence), to what extent changing items based on cognitive pretests improves the psychometric properties of these items, or whether iterative approaches such as ant colony optimization can also be used to optimize test criterion validity or measurement invariance across countries. We hope this editorial stimulates the discussion, the use, and the advancement of further developments in scale construction.

**Acknowledgments**

This Editorial is based on the meeting „New Developments in Scale Construction“ that took place at GESIS – Leibniz Institute for the Social Sciences in October 2015.

**References**

American Educational Research Association. (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Au, N., & Lorgelly, P. (2014). Anchoring vignettes for health comparisons: an analysis of response consistency. Quality of Life Research, 23, 1721-1731.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. Journal of Marketing Research, 38, 143-156.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. Public Opinion Quarterly, 71(2), 287-311.

Brunner, M., & Süß, H. M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. Educational and Psychological Measurement, 65, 227-240.

Bolt, D. M., Lu, Y., & Kim, J. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. Psychological Methods, 19, 528-541.

Buss, D. M., & Craik, K. H. (1980). The frequency concept of disposition: Dominance and prototypically dominant acts. Journal of Personality, 48, 379-392. doi: 10.1111/j.1467-6494.1980.tb00840.x

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. Journal of Personality and Social Psychology, 95, 1005-1018.

Chevalier, A., & Fielding, A. (2011). An Introduction to Anchoring Vignettes. Journal of the Royal Statistical Society: Series A (Statistics in Society), 174, 569-574.

Crane, M., Rissel, C., Greaves, S., & Gebel, K. (2016). Correcting bias in self-rated quality of life: An application of anchoring vignettes and ordinal regression models to better understand QoL differences across commuting modes. Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 25, 257-266.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334. Retrieved from http://download.springer.com/static/pdf/417/art%3A10.1007%2FBF02310555.pdf?auth66=1407595415_4605-f078b28cec2fcecedbb9e76e4c47&ext=.pdf

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. Educational and Psychological Measurement, 64, 391-418. doi: 10.1177/0013164404266386

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. Journal of Research in Personality, 57, 119-130.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. Educational and Psychological Measurement, 66, 930-944.

Gu, F., Little, T. D., & Kingston, N. M. (2013). Misestimation of reliability using coefficient alpha and structural equation modeling when assumptions of tau-equivalence and uncorrelated errors are violated. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 9, 30-40. doi: 10.1027/1614-2241/a000052

Janssen, A. B., Schultze, M., & Grötsch, A. (2015). Following the Ants. In European Journal of Psychological Assessment. Advance online publication. doi: 10.1027/1015-5759/a000299

Jürges, H., & Winter, J. (2013). Are anchoring vignettes ratings sensitive to vignette age and sex? Health Economics, 22, 1-13.

King, G., Murray, C. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross – cultural comparability of measurement in survey research. American Political Science Review, 97, 567-583.

Leite, W., Huang, I., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. Multivariate Behavioral Research, 43, 411-431. doi: 10.1080/00273170802285743

Lenzner, T., Neuert, C., & Otto, W. (2016). Cognitive pretesting. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences (GESIS Survey Guidelines).

Marcoulides, G. A., & Drenzer, Z. (2003). Model specifications searchers using ant colony optimization algorithms. Structural

Equation Modeling, 10, 154-164.
McBride, L., & Moran, G. (1967). Double agreement as a function of item ambiguity and susceptibility to demand implications of the psychological situation. Journal of Personality and Social Psychology, 6, 115-118.
McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Erlbaum.
Moors, G., Kieruj, N. D., & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. Sociological Methodology, 44, 369-399. doi: 10.1177/0081175013516114
Mottus, R., Allik, J., Realo, A., Pullmann, H., Rossier, G., Zecca, G., ... Tseung, C. N. (2012). Comparability of self-reported conscientiousness across 21 countries. European Journal of Personality, 26, 303-317. doi: 10.1002/per.840
Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. Journal of Research in Personality, 59, 56-68. doi: 10.1016/j.jrp.2015.09.001
Padilla, M. A., & Divers, J. (2013). Coefficient omega bootstrap confidence intervals: Nonnormal distributions. Educational and Psychological Measurement, 73, 956-972. doi: 10.1177/0013164413492765
Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), Measures of personality and social psychological attitudes (pp. 17-59). San Diego, CA: Academic Press.
Primi, R., Zanon, C., Santos, D., De Fruit, F., & John, O. P. (2016). Can they make adolescent self-reports of social-emotional skills more reliable, discriminant, and criterion-valid? European Journal of Psychological Assessment, 32, 39-51.
Rammstedt, B., Beierlein, C., Brähler, E., Eid, M., Hartig, J., Kersting, M., ... Weichselgartner, E. (2015). Quality standards for the development, application, and evaluation of measurement instruments in social science survey research. RatSWD Working Papers 245, http://www.ratswd.de/dl/RatSWD_WP_245.pdf
Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. Psychological Assessment, 25, 1137-1145. doi: 10.1037/a0033323
Raykov, T. (1997). Estimation of composite reliability for congeneric measures. Applied Psychological Measurement, 21, 173-184.
Raykov, T., & Pohl, S. (2013). On studying common factor variance in multiple-component measuring instruments. Educational and Psychological Measurement, 73, 191-209. doi: 10.1177/0013164412458673
Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. Psychometrika, 74, 145-154.
Robie, C., Brown, D. J., & Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. Journal of Business and Psychology, 21, 489-509. doi: 10.1007/s10869-007-9038-9
Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. Psychological Methods, 21, 137-150. doi: 10.1037/met0000045
Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74, 107-120. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2792363/pdf/11336_2008_Article_9101.pdf
Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. Journal of Personality and Social Psychology, 94, 718-737. doi: 10.1037/0022-3514.94.4.718
Thalmayer, A. G., & Saucier, G. (2014). The questionnaire big six in 26 nations: Developing cross-culturally applicable big six, big five and big two inventories. European Journal of Personality, 28, 482-496. doi: 10.1002/per.1969
Trott, D. M., & Jackson, D. N. (1967). An experimental analysis of acquiescence. Journal of Experimental Research in Personality, 2, 278-288.
Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 3, 4-70. doi: 10.1177/109442810031002
Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. International Journal of Research in Marketing, 27, 236-247. doi: 10.1016/j.ijresmar.2010.02.004
Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (2013). Do individual response styles matter? Journal of Individual Differences, 34, 69-81. doi: 10.1027/1614-0001/a000102
Willis, G. (2004). Cognitive Interviewing Revisited: A useful technique, in theory? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. E. Singer (Eds.), Methods for testing and evaluating survey questionnaires (pp. 299-317). Hoboken, NJ: Wiley.
Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? Journal of Psychoeducational Assessment, 29, 377-392. doi: 10.1177/0734282911406668
Zhang, Z., & Yuan, K.-H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. Educational and Psychological Measurement, 76, 387-411. doi: 10.1177/0013164415594658
Ziegler, M. (2011). Applicant faking: A look into the black box. The Industrial and Organizational Psychologist, 49, 29-36.
Ziegler, M. (2014). Stop and state your intentions! Let's not forget the ABC of test construction. European Journal of Psychological Assessment, 30, 239-242. doi: 10.1027/1015- 5759/a000228

Ziegler, M. (2015). „F*** you, I won't do what you told me!“ Response biases as threats to psychological assessment. European Journal of Psychological Assessment, 31, 153-158. doi: 10.1027/1015-5759/a000292

Ziegler, M., & Brunner, M. (2016). Test standards and psychometric modeling. In A. A. Lipnevich, F. Preckel, & R. Roberts (Eds.), Psychosocial skills and school systems in the 21st century (pp. 29-55). New York, NY: Springer.

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items: Pitfalls and loopholes. European Journal of Psychological Assessment, 31, 231-237. doi: 10.1027/1015-5759/a000309

Ziegler, M., MacCann, C., & Roberts, R. D. (2011). Faking: Knowns, unknowns, and points of contention. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), New perspectives on faking in personality assessment (pp. 3-16). New York, NY: Oxford University Press.

**Daniel Danner**
GESIS – Leibniz Institute for the Social Sciences
P.O. Box 122155
68072 Mannheim
Germany
E-mail daniel.danner@gesis.org

**Matthias Ziegler**
Institut für Psychologie
Humboldt Universität zu Berlin
Rudower Chaussee 18
12489 Berlin
Germany
E-mail zieglema@hu-berlin.de