Editorial

Some Thoughts Concerning the Recent Shift from Measures with Many Items to Measures with Few Items

Karl Schweizer

Editor-in-Chief

The typical inventory of the past included several scales composed of a minimum of at least 10 to 12 items each from which the respondent had to select between a very limited number of response alternatives. In many cases there were only "yes" and "no" alternatives. A major concern in evaluating the quality of such scales was its reliability and its accuracy – which is closely related to its reliability; a good reliability was considered the most important property of the scale. The true-scores theory that guided the construction of such scales concentrated on reliability and provided clues as to how to achieve an agreeable degree of reliability (Lord & Novick, 1968). In this framework increasing the number of items was considered the proper means for increasing the reliability. Consequently, many items usually characterized such a scale.

Recent observations suggest that the described characteristics no longer apply. The issues of the *European Journal of Psychological Assessment* over the last year tell a different story: A number of scales published within this time frame include only five or even four items (Baccman, & Carlstedt, 2010; Balzarotti, John, & Gross, 2010; Di Giunta, Eisenberg, Kupfer, Steca, Tramontano, & Caprara, 2010; Isoard-Gautheur, Oger, Guillet, & Martin-Krumm, 2010; Maiano, Morin, Monthuy-Blanc, & Garbarono, 2010; Molinengo, & Testa, 2010; Moustaka, Vlachopoulos, Vazou, Kaperoni, & Markland, 2010; Newman, Limbers, & Varni, 2010; Petermann, Petermann, & Schreyer, 2010; Rivero, Garcia-Lopez, & Hofmann, 2010; van Baardewijk, Andershed, Stegge, Nilsson, Scholte, & Vermeiren, 2010). These scales were published because the authors were able to demonstrate a sufficient degree of reliability, which was actually alpha consistency despite the small number of items. There was even one scale composed only of three items (Vlachopoulos, Letsiou, Palaiologou, Leptokaridou, & Gigoudi, 2010). Obviously, the authors were able to generate a high degree of alpha consistency *although* the number of items was small.

How did they do this? They replaced the binary response format by multiple response formats consisting of a minimum of four ordered categories. The items of the three-item scale even included seven response categories (Vlachopoulos et al., 2010). The effect of the increase in response categories is quite clear: Omitting the problems concerning the metric of the categories, it can be argued that more categories mean an increase in accuracy and a decrease in the error of measurement. A decrease in error in turn is associated with an increase in reliability. So there is a kind of analogy between the increase in number of binary items and the increase in the number of response categories.

In a way the new development is a favorable development. The scales constructed in this way can be expected to show a higher degree of homogeneity than those constructed in the "old" way, and they are more likely to survive an investigation by means of confirmatory factor analysis according to congeneric test theory (Lucke, 2005; McDonald, 1999). It is simply easier to arrive at a small set of homogeneous items than at a large one. So the new development is really favorable for the construction of very homogeneous scales, and it is instrumental for the research concentrating on very homogeneous constructs.

However, there are also disadvantages that must be taken into consideration. One disadvantage lies in the limitation in the representation of a construct. The danger is that such scales with few items – this also includes short versions of scales and scales for the purpose of screening – do not really properly represent the construct of interest. Especially in the case of short scales associated with second- and third-order constructs only a limited scope of the contents characterizing the construct may be truly represented. Furthermore, there is another problem to be considered: The content of the item must allow for more than two responses. In statements that allow for only two responses a response format that includes several response categories does not apply.

European Journal of Psychological Assessment 2011; Vol. 27(2):71–72
DOI: 10.1027/1015-5759/a000056

So the new development that seems to marginalize reliability as a criterion for evaluating the quality of a scale should be accompanied by an increased concern regarding the validity of a scale. Not surprisingly, validity was recently regarded as the most important property of a scale (Cizek, Rosenberg, & Koons, 2008). The described development in test construction even seems to further increase the importance of validity. In the future much more emphasis should be given to the evaluation of validity than is presently the case. There even seems to be a need for a revision of the criteria for demonstrating the validity of a scale. New methods appropriate for assuring the appropriate representation of a construct may have to be developed. The demonstration of trait-specific equivalence (Schweizer & Schreiner, 2010) may be one of them.

References

Baccman, C., & Carlstedt, B. (2010). A construct validation of a profession-focused personality questionnaire (PQ) versus the FFPI and the SIMP. *European Journal of Psychological Assessment, 26*, 136–142.

Balzarotti, S., John, O. P., & Gross, J. J. (2010). An Italian adaptation of the Emotion Regulation Questionnaire. *European Journal of Psychological Assessment, 26*, 61–67.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*, 397–412.

Di Giunta, L., Eisenberg, N., Kupfer, A., Steca, P., Tramontano, C., & Caprara, G. V. (2010). Assessing perceived empathic and social self-efficacy across countries. *European Journal of Psychological Assessment, 26*, 77–86.

Isoard-Gautheur, S., Oger, M., Guillet, E., & Martin-Krumm, C. (2010). Validation of a French version of the Athlete Burnout Questionnaire (ABQ) in competitive sport and physical education context. *European Journal of Psychological Assessment, 26*, 203–211.

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

Lucke, J. F. (2005). The alpha and the omega of congeneric test theory: An extension of reliability and internal consistency of heterogeneous tests. *Applied Psychological Measurement*, *29*, 65–81.

Maiano, C., Morin, A. J. S., Monthuy-Blanc, J., & Garbarono, J.-M. (2010). Construct validity of the Fear of Negative Appearance Evaluation Scale in a community sample of French adolescents. *European Journal of Psychological Assessment, 26*, 19–27.

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

Molinengo, G., & Testa, S. (2010). Analysis of the Psychometric properties of an assessment tool for deviant behavior in adolescence. *European Journal of Psychological Assessment, 26*, 108–115.

Moustaka, F. C., Vlachopoulos, S. P., Vazou, S., Kaperoni, M., & Markland, D. A. (2010). Initial validity evidence for the Behavioral Regulation in Exercise Questionnaire 2 among Greek exercise participants. *European Journal of Psychological Assessment, 26*, 269–276.

Newman, D. A., Limbers, C. A., & Varni, J. W. (2010). Factorial invariance of child self-report across English and Spanish language groups in a Hispanic population utilizing PedsQL TM 4.0 Generic Core Scales. *European Journal of Psychological Assessment, 26*, 194–202.

Petermann, U., Petermann, F., & Schreyer, I. (2010). The German Strengths and Difficulties Questionnaire: Validity of the teacher version for preschoolers. *European Journal of Psychological Assessment, 26*, 256–262.

Rivero, R., Garcia-Lopez, L., & Hofmann, S. G. (2010). The Spanish Version of the Self-Statements During Public Speaking Scale: Validation in adolescents. *European Journal of Psychological Assessment, 26*, 129–135.

Schweizer, K., & Schreiner, M. (2010). Avoiding the effect of item wording by means of bipolar instead of unipolar items: An application to social optimism. *European Journal of Personality, 24*, 137–150.

van Baardewijk, Y., Andershed, H., Stegge, H., Nilsson, K. W., Scholte, E., & Vermeiren, R. (2010). Development and tests of short versions of the Youth Psychopathic Traits Inventory and the Youth Psychopathic Traits Inventory-Child Version. *European Journal of Psychological Assessment, 26*, 122–128.

Vlachopoulos, S. P., Letsiou, M., Palaiologou, A., Leptokaridou, E. T., & Gigoudi, M. A. (2010). Assessing multidimensional exercise amotivation among adults and older individuals: the amotivation toward exercise scale. *European Journal of Psychological Assessment, 26*, 248–255.

Karl Schweizer

Department of Psychology
Goethe University Frankfurt
Mertonstr. 17
D-60054 Frankfurt a.M.
Germany
Tel. +49 69 798-22081
Fax +49 69 798-23847
E-mail k.schweizer@psych.uni-frankfurt.de

