Special Issue of the Journal of Individual Differences:

Measuring psychological constructs with short scales: Positive outlooks and caveats

Short scales – Five misunderstandings and ways to overcome them [Editorial]

Matthias Ziegler[1], Christoph J. Kemper[2], Peter Kruyen[3]

[1] Humboldt-Universität zu Berlin, Berlin, Germany

[2] Institute for Medical and Pharmaceutical Proficiency Assessment, Mainz, Germany

[3] Institute for Management Research, Radboud University, Nijmegen, the Netherlands

What do we know about short psychological scales? The idea for this special issue was born out of a common interest in short scales, defined as scales measuring a specific psychological construct with fewer than, say, ten items. All editors of this special issue were recently either active in short scale construction themselves (Kemper, Beierlein, Kovaleva, & Rammstedt, 2013; Ziegler, Kemper, & Rammstedt, 2013) or evaluated the current status of short scales in the field (Kemper, Brähler, & Zenger, 2013; Kruyen, Emons, & Sijtsma, 2012, 2013a, 2013b, 2014). Based on discussions with other scientists about the challenges and caveats of short scales, they felt a strong need to share contemporary issues regarding short scales with a broader scientific audience.

This special issue brings together a variety of researchers who have approached the topic of short scales from different angles. It is our ambition to resolve some misunderstandings and open questions regarding the construction, psychometric quality, and use of short scales by bringing together methodological, statistical and construction-oriented perspectives . We aim at contributing to an ongoing debate

regarding the utility as well as caveats concerning the use of short scales for the measurement of individual differences. At the same time we hope to dispel some existing misunderstandings.

Misunderstanding 1: Short scales will soon become obsolete

Psychological constructs like personality or intelligence have been shown to be useful predictors of many relevant real-life outcomes (Kuncel, Hezlett, & Ones, 2004; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). Unsurprisingly, this has sparked the interest of other research fields, e.g. economics (Lindqvist & Vestman, 2011), public health (Lämmle, Ziegler, Seidel, Worth, & Bös, 2012), and education (Heckman, 2011) in the measurement of psychological constructs. Given the many practical advantages of making scales as short as possible, there has always been—without a doubt—a high demand for short scales both within psychology as well as within other disciplines.

In fact, the demand for short scales is currently expanding at an accelerating speed. One reason for the increasing need for short scales could be a changing way to approach psychological research in general. With research questions becoming more and more complex, involving more and more constructs and statistical methods to model such data, new ways of discovery open themselves up. Consequently, it has been suggested that some psychological phenomena need to be discovered first before they can be explained (Haig, 2005). This means that based on a wealth of data and the use of sophisticated methods, it is now possible to generate hypotheses regarding complex psychological phenomena. Based on these hypotheses, testable theories can be suggested. Thus, the combination of large-scale data, sophisticated methods and a breadth of psychological constructs (even if assessed with short scales) might offer

numerous and potentially very promising new research avenues. Unsurprisingly, all authors within this special issue acknowledged the added value of short scales.

Misunderstanding 2: Short scales are intended to substitute longer or non-abbreviated scales in individual decision-making

The proliferation of short scales across disciplines comes at a price. Whereas in the past researchers in the field of psychology preferred measures that maximize the construct coverage neglecting efficiency of measurement in order to fulfill the requirements of validity, researchers in other disciplines may give efficiency a higher weight in order to maximize the representativeness of the sample (see Rammstedt & Beierlein in this issue for a more detailed discussion). Unfortunately, these two general aims are hard to reconcile.

The appearance of short scales that were constructed in other disciplines—sometimes without the necessary expertise in test theory and construction—might be the reason for a common misunderstanding which is that short scales are always intended to be used for the same purposes as longer or non-abbreviated scales. When a psychological measure is constructed and published it is expected that the measurement purpose is clearly stated. However, for many short scales this is not the case unfortunately. Consequently, it is incorrectly assumed by many that such short-scales are supposed to be put to the same purposes as long or non-abbreviated scales: an important misunderstanding.

Thus, authors of short scales are not without blame when it comes to the origin of this misunderstanding. Usually, short scales are constructed in order to make statements regarding correlational patterns within large and population representative samples. Yet, within psychology most measures are supposed to not only render valid group statistics but also to render valid individual-level statistics. Maybe because it is

often not explicitly ruled out, it is sometimes concluded that short scales can be used as substitutes of longer or non-abbreviated scales in individual decision-making. However, short scales are often and rightfully criticized for their diminished usefulness in individual-level decisions (Emons, Sijtsma, & Meijer, 2007; Kruyen et al., 2012, 2013a).

Although we also observe a growing demand for short scales in the context of individual decision-making, we have to keep in mind that a majority of short scales were actually designed for research purposes only and not individual decision-making. The authors within this special issue address critical aspects of short scales for individual decision-making and acknowledge the limited use of short scales in this regard.

Nevertheless, applied settings with a high demand for short scales are clinical practice (e.g. screening for mental disorders) and personnel selection. In these settings it is however especially important that scores derived from such scales are sufficiently validated and demonstrate acceptable levels of specificity and sensitivity at the group level (e.g., Antunes, Murtagh, Bausewein, Harding, & Higginson, in press). Of even greater importance are quality indices at the individual level such as measurement precision and classification consistency (Kruyen, 2012). Considering these individual-level quality indices, Kruyen et al. (2012) found in fact that short scales can be used under very specific conditions for individual decision-making. This, however, requires a full understanding of the psychometric mechanisms that are at work when testing individuals and making decisions about them. In general, they recommended using long scales for individual-decision making

At the same time, the rigid idea that short scales cannot be used for individual decision-making is equally wrong and oftentimes based on the wrong quality indices (e.g., internal consistency, see below). In order to avoid such undeserved criticism in the future, authors of short scales are encouraged to clearly state the measurement

purposes of their scale, e.g. research, screening, selection etc., and provide validity evidence accordingly. The paper by Schipolowski, Schroeders, and Wilhelm (this issue) is a recommendable example for such a practice.

Misunderstanding 3: Short scales yield lower test-criterion correlations compared to longer or non-abbreviated scales

Many studies evaluating the validity of scores derived from short scales compare their ability to yield test-criterion correlations comparable with the non-abbreviated scale (Credé, Harms, Niehorster, & Gaye-Valentine, 2012). Whenever such studies take test family into account (for an explanation see Ziegler, Poropat & Mell, this issue), test criterion correlations of short scales are often just as good as their longer counterparts (Thalmayer, Saucier, & Eigenhuis, 2011). In fact, Heene and Bühner (this issue) showed that test criterion correlations for short scales equal those of the corresponding longer scales when population data are used and no correlated residuals are assumed.

This finding by Heene and Bühner strongly supports the usefulness of short scales for large-scale assessments as described above. In fact the work by Carson, Gosling and Koelkebeck (this issue) as well as Oshio, Abe, Cutrone, and Gosling (this issue) support this interpretation. Nevertheless, this positive evaluation of short scales should be considered with caution. Heene and Bühner also state that sampling error might negatively affect the congruence between long and short scales with regard to test-criterion correlations. Moreover, Beauducel (also this issue) elegantly shows that correlated errors are to be expected when tests are shortened. This more critical outlook is supported by findings from Schiplowski, Schroeders, and Wilhelm.

Thus, based on the existing literature, a generally positive conclusion regarding test-criterion correlation convergence between long and short scales could be drawn.

However, there are also doubts and therefore, each published short scale should provide empirical evidence supporting the interpretation of the test score as intended.

Misunderstanding 4: Short scales have to be as internally consistent as longer or non-abbreviated scales

Reliability is one of the key psychometric properties of test scores (American Educational Research Association, 1999). There are many different ways to estimate reliability (e.g., Cronbach & Shavelson, 2004; Zinbarg, Revelle, Yovel, & Li, 2005). Unfortunately, Cronbach alpha is often the only reported estimate of reliability. Indeed, Cronbach alpha is one way to estimate the internal consistency of a test score by taking number of items and their intercorrelations into account. However, many studies show problems regarding alpha (Sijtsma, 2009) as well as the lower contribution to score validity compared to other reliability estimates such as test-retest reliability (McCrae, Kurtz, Yamagata, & Terracciano, 2011). The mathematics behind alpha are what makes it a problematic estimate for short scales. By definition short scales do not have many items. If a short scale assesses a broad construct, interitem-correlations cannot be high resulting in a low internal consistency. A remedy against this is the selection of items with high item total discriminations resulting in a sufficient Cronbach alpha but also a loss of content validity (see Ziegler et al., this issue). Thus, developers of short scales are caught between a rock and a hard place when it comes to reliability when it is only estimated with Cronbach alpha. However, there are solutions to this problem. Other reliability estimates such as McDonald's omega (see Schipolowskie, Schroeders & Wilhelm, this issue) or test retest reliability (Rammstedt & Beierlein, this issue) can be used instead. Another alternative are new ways of constructing short scales which focus less on internal consistency and more on construct representation (see Sandy, Gosling & Koelkebeck; Oshio et al., this issue). At this point it is again important to focus on the

intended use of short scales. If a scale is only intended to be used for group statistics, emphasizing efficiency of measurement over internal consistency can be acceptable. When scales are intended to be used for individual assessments, it is indeed important to have a reliable score interpretation. Omega might be one way out of the dilemma explained above. However, as already pointed out above, it is important to keep in mind that measurement precision (Cronbach & Shavelson, 2004) certainly needs to be considered in individual decision making. This can be done fairly easily by estimating the confidence interval around the observed score which covers the true score with a specified probability (cf. Kruyen, 2012). By doing this, a test score can be classified, e.g. as below average, average or above average. If, however, the confidence interval for a score should happen to range from below average to above average, the test is practically useless for individual decision making (see Schipolowski, Schroeders & Wilhelm).

Misunderstanding 5: Short scales can be developed overnight

At several points in this special issue it is highlighted that short scales are sometimes constructed on the fly. Thus, no measurement purpose is defined, no working definition for the construct to be measured specified and no transparent item selection strategy laid out. Other times, the item selection strategy is purely based on statistical reasons (e.g., maximize internal consistency). However, the examples within this special issue nicely show that developing a short scale is no easy task. In fact, it requires thorough planning, lots of data and the application of different methodological approaches. Many things can go wrong when constructing a short(ened) scale.

Nevertheless, based on rigorous construction strategies (Schipolowski et al.), methods such as structural equation modeling with other measures (Oshio et al.), innovative item selection strategies (Sandy et al.) or by going beyong classical test

theory (Ziegler et al.) better results can be obtained. At the same time, there are still important lessons to be learned about the mechanisms at work when shortening scales. Besides the aforementioned aspects, more simulation studies are needed to shed light onto these mechanisms (see Beauducel or Heene & Bühner).

Conclusions

As editors of this special issue, our goal was to resolve some misunderstandings and open questions regarding short scales for the measurement of individual differences. The growing demand for efficient measures of psychological constructs in psychology as well as in other disciplines is rarely questioned these days. Thus, short scales will be developed either with or without the necessary skills and expertise in test theory, test development, and the constructs to be measured. However, this growing demand may also be seen as an opportunity to guide the test shortening process by giving advice firmly rooted in empirical studies (best practice). The research presented in this special issue is highly relevant in this regard.

Contributing authors carved out several strengths as well as weaknesses of short scales suggesting that the most important question is not whether short scales are good or bad, useful or useless, psychometrically sound or not. The most important question is whether the profile of a short scale's psychometric quality (internal consistency, test-retest reliability, construct reliability, measurement precision, content validity, criterion-related validity, factorial validity, classification accuracy etc.) matches the necessities of the assessment setting in which the short scale is applied, e.g. research based on small and highly selective samples, research based on large representative samples, diagnosis of individuals in clinical practice, screening in primary care, personal selection etc. (cf. also Kemper et al., 2013; Kruyen, 2012). Thus, potential users of a short scale should be made aware of the main purpose of a scale. As a matter of course, the

developer himself should be aware of the purpose before engaging in construction efforts. This may sound trivial, however, the reality of short scale construction suggests otherwise (Smith, McCarthy, & Anderson, 2000; Kruyen et al., 2012). This seemingly basic request is as true for long scale construction as it is for short scale construction. The same holds true for the following issues.

Before engaging in short scale construction, three basic questions should be answered: (1) *What is the construct to be measured?* The answer should include a definition of the construct, its factorial structure, its convergent and discriminant constructs, and its breadth. In short, the answer should be a thorough definition of the construct's theory and nomological net. (2) *What is the main purpose of the short scale?* Depending on the main purpose, different strategies for item selection, reliability estimation, and validation are necessary. (3) *What is the targeted population of the short scale?* Above we have outlined that short scales oftentimes are intended to be used in representative samples of the general population. Other times, specific clinical populations or populations form other practical areas are the focus of a short scale. In both cases, construction of the short scales as well as its evaluation should be done using samples matching the test user's purpose.

These basic questions may serve as an essential guidance in the construction process. By addressing them before engaging in construction, necessary preconditions for the psychometric quality of a short scale are established.

Hopefully this special issue provides insights into uses and misuses of short scales. We thank the contributors for sharing their perspectives on the issue. We appreciated reading the articles which sparked intensive discussions amongst the editors and generated new ideas. We are hopeful that other readers will be enthused as well and encouraged to provide more research on the issue in the future.

References

Antunes, B., Murtagh, F., Bausewein, C., Harding, R., & Higginson, I. J. Screening for Depression in Advanced Disease: Psychometric Properties, Sensitivity and Specificity of Two Items of the Palliative care Outcome Scale (POS). *Journal of Pain and Symptom Management*. doi: http://dx.doi.org/10.1016/j.jpainsymman.2014.06.014

Association, A. E. R. (1999). *Standards for educational and psychological testing*: Amer Educational Research Assn.

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology, 102*, 874-888. doi: 10.1037/a0027403

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

Emons, W. H., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods, 12*, 105.

Haig, B. (2005). An abductive theory of scientific method. *Psychological Methods, 10*, 371-388.

Heckman, J. J. (2011). The Economics of Inequality: The Value of Early Childhood Education. *American Educator, 35*, 31.

Kemper, C. J., Beierlein, C., Kovaleva, A., & Rammstedt, B. (2013). Entwicklung und Validierung einer ultrakurzen Operationalisierung des Konstrukts Optimismus-Pessimismus [Development and validation of an ultrashort measure for the construct of optimism-pessimism-The Scale Optimism-Pessimism-2 (SOP2)]. *Diagnostica, 59*, 119-129.

Kemper, C. J., Brähler, E., & Zenger, M. (2013). Psychologische und sozialwissenschaftliche Kurzskalen für Wissenschaft und Praxis–Eine Einführung. In C. J. Kemper, E. Brähler & I. Zenger (Eds.), *Psychologische und sozialwissenschaftliche Kurzskalen* (pp. 1-7): MWV Medizinisch Wissenschaftliche Verlagsgesellschaft mbH & Co. KG.

Kruyen (2012). Making decisions about individuals using tests and questionnaires: When is short too short? *Dissertation*. Ridderkerk, the Netherlands.

Kruyen, P. M., Emons, W. H., & Sijtsma, K. (2012). Test Length and Decision Quality in Personnel Selection: When Is Short Too Short? *International Journal of Testing, 12*, 321-344.

Kruyen, P. M., Emons, W. H., & Sijtsma, K. (2013a). On the Shortcomings of Shortened Tests: A Literature Review. *International Journal of Testing, 13*, 223-248.

Kruyen, P. M., Emons, W. H., & Sijtsma, K. (2013b). Shortening the S-STAI: Consequences for research and clinical practice. *Journal of Psychosomatic Research, 75*, 167-172. doi: http://dx.doi.org/10.1016/j.jpsychores.2013.03.013

Kruyen, P. M., Emons, W. H., & Sijtsma, K. (2014). Assessing Individual Change Using Short Tests and Questionnaires. *Applied Psychological Measurement, 38*, 201-216. doi: 10.1177/0146621613510061

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*, 148-161.

Lämmle, L., Ziegler, M., Seidel, I., Worth, A., & Bös, K. (2012). Four classes of physical fitness in German children and adolescents: only differences in performance or at-risk groups? *International Journal of Public Health, 58*, 187-196.

Lindqvist, E., & Vestman, R. (2011). The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment. *American Economic Journal: Applied Economics, 3*, 101-128. doi: doi: 10.1257/app.3.1.101

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal Consistency, Retest Reliability, and Their Implications for Personality Scale Validity. *Personality and Social Psychology Review, 15*, 28-50. doi: 10.1177/1088868310366253

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The Power of Personality: The Comparative Validity of Personality Traits, Socioeconomic Status, and Cognitive Ability for Predicting Important Life Outcomes. *Perspectives on Psychological Science, 2*, 313-345. doi: doi:10.1111/j.1745-6916.2007.00047.x

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.

Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative Validity of Brief to Medium-Length Big Five and Big Six Personality Questionnaires. *Psychological Assessment, 23*, 995-1009. doi: 10.1037/a0024165

Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The Vocabulary and Overclaiming Test (VOC-T). *Journal of Individual Differences, 34*, 32-40.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123-133.

-

All papers from this issue