Editorial

Stop and State Your Intentions!

Let's Not Forget the ABC of Test Construction

Matthias Ziegler

Humboldt-Universität zu Berlin, Germany

I have used recent editorials to point out some of the recurring criticisms brought forward by our reviewers that have lead to papers being rejected (Ziegler, 2014; Ziegler & Bensch, 2013; Ziegler, Booth, & Bensch, 2013; Ziegler & Vautier, 2014). In this editorial I want to continue with this tradition and focus on the very beginnings of the test construction process. The reason is that more and more papers are using impressive and sophisticated statistical methods (e.g., Breevaart, Bakker, Demerouti, & Hetland, 2012; Huang & Dong, 2012), large data sets or data sets from different countries (e.g., Caballo, Salazar, Irurtia, Arias, & Hofmann, 2010; Consiglio, Alessandri, Borgogni, & Piccolo, 2013), or new technical approaches to measurement (e.g., Schroeders, Bucholtz, Formazin, & Wilhelm, 2013). All of this is important and contributes to advancing our field. However, the solid base for all of these papers needs to be a sound test construction strategy. Unfortunately, there are numerous papers being submitted and eventually rejected that could also be regarded as examples for the positive attributes stated above. What those papers are missing, though, is what I will call here the "ABC of test construction." What I mean by this is that, after having read the paper, the reader should know the answers to three basic questions:

A. What is the construct being measured?
B. What are the intended uses of the measure?
C. What is the targeted population?

The answers to these questions not only determine the construction strategy in many important aspects, but also provide the basis for correct use of the test by practitioners. Finally, the answers to these questions can be used as stepping stones for other researchers building on the published findings.

What Is the Construct Being Measured?

The first step in test construction should be the definition of the construct which is supposed to be measured. However, such a definition should entail more than just a few sentences explaining the core features. Cronbach and Meehl (1955) made the term "nomological net" popular. They said: "Scientifically speaking, to 'make clear what something is' means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws which constitute a theory as a nomological network." (p. 290). They went on by stating several more important features, of which the idea that a nomological net may relate theoretical constructs to observables and different theoretical constructs to one another is of specific importance here. From this it follows that the definition of a construct using the nomological network approach should not only include the explanation by which behaviors a construct manifests itself, it should also explain the relation to other constructs. Later this became known as convergent and discriminant validity (Campbell & Fiske, 1959). Moreover, the nomological net also has bearings on the criteria that are supposed to be predictable from the test score. Thus, a first summary at this point is that the definition of the nomological net does not only help to understand what is to be measured, it also helps to determine a validation strategy with regard to construct and criterion-related validity evidence. Based on the nomological net, a number of specific hypotheses describing relations with other constructs and criteria can be derived which build the first part of the validation strategy. Finally, an exact definition of the nomological net is also important to compare measures claiming to capture the same construct (Miller, Gaughan, Maples, & Price, 2011; Pace & Brannick, 2010; Ziegler et al., 2013).

It should be obvious that the definition of such a nomological net is also essential for item construction. This was further outlined by Loevinger (1957). She differentiated between a substantive, a structural, and an external component of construct validity. These three components are supposed to mirror three stages of test construction, that is, constitution of the item pool, analysis of the internal structure of this pool, and finally item selection and forming of a scoring key, the result of which is used in correlational analyses with other test scores. Clearly, the substantive component, that is, the item pool, is strongly related to the

European Journal of Psychological Assessment 2014; Vol. 30(4):239–242
DOI: 10.1027/1015-5759/a000228

nomological net. In fact it should be possible to allocate each item within the nomological net. Moreover, it should be evident how the items fulfill Cronbach's and Meehl's demand that a nomological net should relate theoretical constructs to observables. Thus, the item pool first considered should reflect the ideas outlined in the nomological net. While this seems to be a given in test construction, Loevinger already stated: "Although this process, the constitution of the pool of items, is a universal step in test construction, it has not been adequately recognized by test theory. But this step in test construction is crucial for the question of whether evidence for the validity of the test will also be evidence for the validity of a construct." (p. 658 f.). Unfortunately, many papers are being rejected for exactly this reason: It remains unclear what is supposed to be measured and how this is reflected in the items.

Another implication of Loevinger's ideas, more specifically the structural component, is testing factorial validity. Given the development in statistical software and the rise of confirmatory factor analysis (Alonso-Arbiol & van de Vijver, 2010), evidence for factorial validity can be provided relatively easily in many cases. However, such evidence, while certainly important, is but one aspect of validity as outlined above.

A good example of a paper that pays attention to all these aspects is the paper by Mussel, Spengler, Litman, and Schuler (2012). Those authors clearly lay out a nomological net which provides the basis for item construction and finally validation.

What Are the Intended Uses of the Measure?

Psychological tests are currently applied to a wide range of tasks (Kaplan & Saccuzzo, 2012). It has also been shown that psychological tests indeed predict a wide range of behaviors (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). However, not every test is suited for every task (e.g., Kruyen, Emons, & Sijtsma, 2012). Thus, each test publication should clearly state the purposes the measure is intended to be used for. This means not only helping practitioners and other researchers to decide whether to use a measure or not; it also has implications for test construction.

Clearly, the intended use of a measure affects the nomological net. A test that is supposed to predict aggressiveness in children will require a different linking between the theoretical construct aggressiveness and observables than an aggressiveness test intended to select managers. Thus, the constitution of the item pool is likewise affected.

The purpose of the measurement also influences item selection. For example, if the test score is intended to be used to differentiate between a wide range of different construct levels, an equally wide range of item difficulties is necessary. Unfortunately, this conflicts with item selection based on factor loadings or item total discriminations (Ziegler, 2014).

The purpose of the measurement has even further reaching implications as well. Decisions derived from interpreting test scores can be regarded as status assessments or prognoses. Each of these decisions requires a different reliability estimate. Both decisions should be based on confidence intervals. However, whereas status assessments can do without test-retest estimates, prognoses should be based on confidence intervals taking test-retest stability into account. Thus, the measurement purpose affects the way in which reliability of the test scores should be estimated.

Last but not least, the purpose of the measurement also affects the validity strategy. For example, test scores which are intended for use in personnel selection should show their ability to predict job success (see, e.g., Blickle, Momm, Liu, Witzki, & Steinmayr, 2011). Likewise, for a clinical screening tool it should be shown that the test score has sufficient sensitivity and specificity (e.g., Mokros, Vohs, & Habermeyer, 2014).

Summing up, reliability estimation, validation strategy, and item formation as well as selection should be strongly influenced by the intended use of a measure.

What Is the Targeted Population?

As was the case for the intended use, the targeted population also affects the item formation process. For example, depending on the educational background of the targeted population, the phrasing of the items might have to differ (Rammstedt & Kemper, 2011). Similar considerations apply when choosing item difficulties.

It should also be obvious that the samples used to construct the test must be derived from the targeted population. Thus, a test, which is supposed to be used in clinical populations should not be evaluated using student samples.

A lot more could be said about the effects on norm group selection, etc. However, for this editorial it suffices to say that there is a link with the intended use of a measure. If the intended use is in any practical setting, for example, personnel selection, clinical screening, or educational assessment, norm values representative of the population and the setting in question should be reported, along with guidelines for practitioners regarding the interpretation of these norm values. Otherwise, the practical use of a measure is questionable.

Recommendations for Authors

The preceding arguments were intended to provide insights into the editorial process, highlighting mistakes which often lead to rejections. Naturally, at this point some recommendations for authors are warranted. The most global recommendation is to provide answers to the three questions stated above. Depending on the nature of the paper, the depths of those answers may of course differ. Clearly, papers presenting new developments should provide detailed answers to all of the questions, outlining the

nomological net, and how items and validation strategy were derived from it. Such papers should also clearly state the intended use and targeted population and report how these issues were considered during test construction. Appropriate a priori hypotheses should be stated and reliability estimates reported that converge with the intended use. By contrast, a paper that provides new validity evidence for an existing measure might suffice by detailing only the part of the nomological net that is related to the new evidence. Cronbach and Meehl (1955) put it this way:

> " 'Learning more about' a theoretical construct is a matter of elaborating the nomological network in which it occurs, or of increasing the definiteness of the components. ... An enrichment of the net such as adding a construct or a relation to theory is justified if it generates nomologicals that are confirmed by observation or if it reduces the number of nomologicals required to predict the same observations. When observations will not fit into the network as it stands, the scientist has a certain freedom in selecting where to modify the network. That is, there may be alternative constructs or ways of organizing the net which for the time being are equally defensible." (p. 290).

Clearly this can be seen as a plea to use the nomological network ideas presented by others and to test them. Based on empirical findings and sound theoretical arguments, such networks may have to be changed. These changes are of importance to all users of the test and would therefore be of interest to many of our readers.

I hope that the ideas presented here, which are mostly borrowed from the masterminds of test construction, will guide authors in their test construction projects and subsequently in their writing of papers. In the end, such papers will, I hope, provide important guidelines for practitioners and points of contact where other researchers may start their projects.

References

Alonso-Arbiol, I., & van de Vijver, F. J. R. (2010). A historical analysis of the *European Journal of Psychological Assessment*. *European Journal of Psychological Assessment, 26*, 238–247. doi: 10.1027/1015-5759/a000032

Blickle, G., Momm, T., Liu, Y., Witzki, A., & Steinmayr, R. (2011). Construct validation of the Test of Emotional Intelligence (TEMINT). *European Journal of Psychological Assessment, 27*, 282–289. doi: 10.1027/1015-5759/a000075

Breevaart, K., Bakker, A. B., Demerouti, E., & Hetland, J. (2012). The measurement of state work engagement: A multilevel factor analytic study. *European Journal of Psychological Assessment, 28*, 305–313. doi: 10.1027/1015-5759/a000111

Caballo, V. E., Salazar, I. C., Irurtia, M. J., Arias, B., & Hofmann, S. G. (2010). Measuring social anxiety in 11 countries. *European Journal of Psychological Assessment, 26*, 95–107. doi: 10.1027/1015-5759/a000014

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Consiglio, C., Alessandri, G., Borgogni, L., & Piccolo, R. F. (2013). Framing work competencies through personality traits: The Big Five Competencies grid. *European Journal of Psychological Assessment, 29*, 162–170. doi: 10.1027/1015-5759/a000139

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Huang, C., & Dong, N. (2012). Factor structures of the Rosenberg Self-Esteem Scale: A meta-analysis of pattern matrices. *European Journal of Psychological Assessment, 28*, 132–138. doi: 10.1027/1015-5759/a000101

Kaplan, R., & Saccuzzo, D. (2012). *Psychological testing: Principles, applications, and issues*. Belmont, CA: Wadsworth.

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.

Loevinger, J. (1957). Objective tests as instruments of psychological theory: Monograph Supplement 9. *Psychological Reports, 3*, 635–694.

Miller, J. D., Gaughan, E. T., Maples, J., & Price, J. (2011). A comparison of agreeableness scores from the Big Five Inventory and the NEO PI-R: Consequences for the Study of Narcissism and Psychopathy. *Assessment, 18*, 335–339. doi: 10.1177/1073191111411671

Mokros, A., Vohs, K., & Habermeyer, E. (2014). Psychopathy and violent reoffending in German-speaking countries: A meta-analysis. *European Journal of Psychological Assessment, 30*, 117–129. doi: 10.1027/1015-5759/a000178

Mussel, P., Spengler, M., Litman, J. A., & Schuler, H. (2012). Development and validation of the German Work-Related Curiosity Scale. *European Journal of Psychological Assessment*,109–117. doi: 10.1027/1015-5759/a000098

Pace, V. L., & Brannick, M. T. (2010). How similar are personality scales of the "same" construct? A meta-analytic investigation. *Personality and Individual Differences, 49*, 669–676. doi: 10.1016/j.paid.2010.06.014

Rammstedt, B., & Kemper, C. J. (2011). Measurement equivalence of the Big Five: Shedding further light on potential causes of the educational bias. *Journal of Research in Personality, 45*, 121–125.

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: 10.1111/j.1745-6916.2007.00047.x

Schroeders, U., Bucholtz, N., Formazin, M., & Wilhelm, O. (2013). Modality specificity of comprehension abilities in the sciences. *European Journal of Psychological Assessment, 29*, 3–11. doi: 10.1027/1015-5759/a000114

Ziegler, M. (2014). Comments on item selection procedures. *European Journal of Psychological Assessment, 30*, 1–2. doi: 10.1027/1015-5759/a000196

Ziegler, M., & Bensch, D. (2013). Lost in translation: Thoughts regarding the translation of existing psychological measures into other languages. *European Journal of Psychological Assessment, 29*, 81–83. doi: 10.1027/1015-5759/a000167

Ziegler, M., Booth, T., & Bensch, D. (2013). Getting entangled in the nomological net. Thoughts on validity and conceptual overlap. *European Journal of Psychological Assessment, 29*, 157–161.

Ziegler, M., & Vautier, S. (2014). A farewell, a welcome, and an unusual exchange. *European Journal of Psychological Assessment, 30*, 81–85. doi: 10.1027/1015-5759/a000203

Matthias Ziegler

Institut für Psychologie
Humboldt Universität zu Berlin
Rudower Chaussee 18
12489 Berlin
Germany
Tel. +49 30 2093-9447
Fax +49 30 2093-9361
E-mail zieglema@hu-berlin.de

