The Issue of Fuzzy Concepts in Test Construction and Possible Remedies

Matthias Ziegler,[1] Christoph J. Kemper,[2] and Timo Lenzner[3]

[1]Humboldt-Universität zu Berlin, Germany, [2]Institute for Medical and Pharmaceutical Proficiency Assessment, Mainz, Germany, [3]GESIS – Leibniz-Institute for the Social Sciences, Mannheim, Germany

With the first issue of 2015, I want to continue the series of editorials aimed at highlighting specific topics relevant during test construction. More importantly, I focus on issues that repeatedly lead to paper rejections. In this issue, I want to take up the cudgels for a mixed methods approach in test construction. The quantitative aspects of testing a newly constructed assessment tool range prominently within this journal (Alonso-Arbiol & van de Vijver, 2010). However, some issues arising during the test development process cannot be dealt with using quantitative methods alone. Along with the coauthors of this editorial – experts in the application of a mixed methods approach in test construction – I want to explore these issues.

Many test constructions show a lack of awareness concerning the comprehensibility of items, specifically in personality tests. However, this lack of awareness can have implications for psychometric quality. Frequently, it is simply assumed that respondents' understanding of a test item matches the meaning implied by the test developer. It is, however, rarely ever tested whether all respondents of the targeted population actually understand the test items correctly and in a similar way. As pointed out in the last editorial (Ziegler, 2014), item content should be precisely tuned to the needs of every potential respondent in a population targeted by a newly developed test. More specifically, person variables such as age, gender, and education (Rammstedt & Kemper, 2011) have to be taken into account when constructing test items in order to assure that each potential respondent fully understands the meaning and may respond accordingly. If these person variables are neglected in the process of test construction, the psychometric quality of a test may be substantially affected by fuzzy concepts. The goal of this editorial is to raise awareness for the detrimental effects fuzzy concepts can have in test development and possible remedies.

What Is a Fuzzy Concept?

The concept of fuzziness stems from computer sciences and was first introduced by Zadeh (1976). He emphasized the vast difference in standards of precision between the definition of concepts (constructs) in the *soft sciences* such as psychology, sociology, linguistics, literature, etc. and the *hard sciences* such as mathematics, physics, or chemistry and proposed a framework for the definition of *soft* constructs through the use of fuzzy algorithms. In the hard sciences, constructs can be easily defined in quantitative terms (see also Michell, 1997, 2001) whereas constructs in the soft sciences are inherently fuzzy. *Fuzzy concepts* according to Zadeh (1976) are much too complex or too imprecise to allow for an exact definition. Such concepts do not have clear cut demarcation lines – their boundaries are fuzzy. Examples are abundant from various domains of human knowledge – *migraine* and *cancer* in medicine, *democracy* and *state* in political sciences, *intelligence* and *extraversion* in psychology, or *grammaticality* and *meaning* in linguistics. Fuzzy concepts are involved in at least two stages of test development.

(1) Personality constructs such as extraversion, self-efficacy, or optimism may be considered fuzzy concepts. Each construct has a number of indicators with some being more closely related and others being more distant. Especially the indicators in the fringes of the construct may as well be considered indicators of other constructs in the nomological net (Ziegler, Booth, & Bensch, 2013). Thus, boundaries of psychological constructs are inherently fuzzy. The necessity to deal with this kind of fuzziness in the first step of test construction – the definition of the construct to be measured – and proper methods in this

DOI: 10.1027/1015-5759/a000255

regard were already addressed in previous editorials (Ziegler, 2014).

(2) Moreover, fuzzy concepts play a role in the development of test items. Before potential respondents of a newly developed test are able to provide valid responses to test items, they have to infer meaning from the statements containing terms such as honest, impromptu, citizen of the world – all of which are fuzzy concepts. In this editorial, we address the kind of fuzziness relating to item development.

The Role of Fuzzy Concepts in Item Development

In psychological testing based on questionnaires, respondents are usually instructed to read several statements and evaluate these statements according to their behavior, attitude, knowledge, etc. These statements are combinations of words following sets of rules in a given language which enable respondents to infer meaning from the statements. Ideally, respondents infer the same meaning from a statement the test developer had in mind when constructing the item in the first place, assuming the developer constructed a valid indicator for the construct. For the same meaning to be inferred, terms (concepts) used in the statements need to be precise and unequivocal. However, in the soft science of psychology dealing with human behavior concepts are often complex, ambiguous, probabilistic, vague, or imprecise – concepts are *fuzzy.*

The implications of fuzzy concepts for psychometric quality will be demonstrated with two examples of test items. The first example is an item from a personality test measuring the construct of self-monitoring (Snyder, 1974): "I can make impromptu speeches even on topics about which I have almost no information." Considering several rules of thumb given in textbooks of test construction, this is a good test item. However, the psychological significance of the item is fully dependent upon the respondent's interpretation of the term *impromptu speech*. According to the Corpus of Contemporary American English (http://www.collocates.info) the word *impromptu* is not a frequently used word. The combination of *free* and *speech* is about 100 times more frequent than the combination of *impromptu* and *speech.* Thus, a substantial portion of potential respondents will most likely have problems understanding the fuzzy concept. Whereas most educated people, for example, psychology students, may be able to infer the meaning intended by the test developer and respond accordingly, many less-educated people may not be able to correctly infer the meaning of *impromptu speech*. Thus, interindividual variability in the interpretation of the fuzzy concept and thus unwanted variance is increased.

Detrimental effects of fuzzy concepts on psychometric quality are even stronger when fuzzy concepts have to be compared, as demonstrated with the following item: "All in all, it is better to be humble and honest than to be important and dishonest." In this item from a Machiavellianism scale (Christie, Geis, & Berger, 1970), the respondent has to infer the meaning of *humble* and *honest*, identify the semantic overlap of the two concepts, and compare this overlap with the semantic overlap of the concepts *important* and *dishonest.* For many potential respondents, especially those with low cognitive ability, this test item poses a real challenge.

Detrimental Effects of Fuzzy Concepts on Psychometric Quality

As the examples above demonstrate, fuzzy concepts may introduce a substantial amount of interindividual variability to the measurement of psychological constructs unrelated to the construct targeted. This additional variability may have detrimental effects on psychometric quality, for example, measurement error, criterion-related validity, and construct validity of a test score interpretation. When respondents do not understand or misunderstand the meaning of a statement, they respond to the item on some other basis than the meaning implied by the test developer. More specifically, they try to infer the meaning from other sources, for example, remaining items in the test, past experience, or contextual factors. To a higher degree, test responses may be affected by sources of variance not related to the personality construct to be measured, for example, differential item functioning (Holland & Thayer, 1986), careless responding (Meade & Bartholomew, 2012), Satisficing-Optimizing (Krosnick, 1991), or response styles such as faking, acquiescence, or extreme/midpoint responding (Kemper & Hock, 2015; Kemper & Menold, 2014; Ziegler & Kemper, 2013). By introducing or increasing the impact of these sources of variance on item responses, psychometric quality of the test score interpretation is inevitably reduced.

Cognitive Interviewing

To avoid or reduce detrimental effects of fuzzy concepts, it is a reasonable approach to investigate whether respondents infer the meaning intended by the test developer from the items. However, in psychological research this approach is only rarely used to optimize test items: "Test takers are a valuable source of information concerning the improvement of tests but are normally overlooked" (Gregory, 1996). In contrast to psychology, scale developers in the social sciences put a strong emphasis on item comprehensibility as more heterogeneous samples – samples representative for the general population (Rammstedt & Beierlein, 2014) – are usually used. In the social sciences, one of the most prominent methods for testing and evaluating items prior to their use in a survey is a qualitative method – cognitive interviewing (Beatty & Willis, 2007; Presser et al., 2004). The cognitive interview is typically a semi-structured, in-depth interview conducted with paid volunteers. It aims at getting insights into the cognitive processes underlying survey responding, for example, "How do survey respondents interpret the items?," "How do they retrieve relevant information from memory?," "How do they map the cognitive representation to the response categories provided?" This information is then used to determine whether respondents understand the items in the way intended by the developer and to identify potential difficulties respondents face when responding to the items (Miller, 2011; Willis, 2005). By identifying problematic items and providing useful information for revision, cognitive interviewing contributes to decreasing measurement error (Willis, 2005).

The most commonly used techniques for gathering information about respondents' cognitive processes and about potential item problems are *thinking aloud* and *verbal probing* (Willis, 2005). During thinking aloud, participants of the cognitive interview are asked to vocalize their thought processes while they answer an item. Thereby, researchers can determine whether participants' interpretation of the item actually matches his or her intended understanding. An advantage of this technique is that it is a relatively standardized procedure, which makes it less prone to bias introduced by interviewers. On the negative side, most participants find thinking aloud quite difficult and many are not capable of vocalizing the thought processes leading to their answers (Willis, 2005). Thus, when applying the think-aloud technique, it is important to provide participants of the cognitive interview with a detailed instruction that explains what they are supposed to do. Moreover, it is important to remind participants over and over again to report their thoughts in order to keep them thinking aloud (Willis, 2005).

Verbal probing is a technique that uses follow-up questions administered either immediately after the participant provides a response to an item (concurrent probing) or after completing the questionnaire (retrospective probing). The goal of probing is to gather specific information about participants' understanding of terms, items or response categories and about the processes leading to a specific response. For example, the item "I feel more like a citizen of the world than of any other country" could be followed by a probing question asking participants to explain what the term *citizen of the world* means to them. Thereby, researchers can determine whether their participants are familiar with this term and whether they correctly associate it with the concept of cosmopolitanism. Depending on the specific cognitive process targeted by a probing question, several types of probes can be distinguished (Willis, 2005), such as comprehension probes (e.g., "What does the term X mean to you?"), information retrieval probes (e.g., "How did you remember that you went to the doctor X times in the past 12 months?"), elaborative probes (e.g., "Can you tell me more about that?"), and category selection probes (e.g., "Why did you select this response category?"). A benefit of the verbal probing technique is that it generates information that may not come to light unless a cognitive interviewer explicitly asks for it (Beatty, 2004) and that it should not interfere with the actual process of responding, whereas thinking aloud might (Beatty & Willis, 2007). A drawback of this technique is that it is open to interviewer effects introduced by how and when interviewers apply the probing questions. Thus, cognitive interviewers need to be properly trained in how to conduct the interviews.

Regarding the design and implementation of cognitive interviews studies, there is currently no consensus on best practices (Presser et al., 2004). However, practitioners seem to agree that participants in cognitive interviews should resemble the target group of the survey concerning sex, age, education, and other characteristics relevant to the topic of the questionnaire being tested. Usually about 20 interviews are conducted. Sessions are audio- or video-recorded and transcribed afterwards. Durations of the individual sessions usually do not exceed 60–90 minutes. Willis (2005) provides a comprehensive overview of the design and implementation of cognitive interviews.

To sum up, applying such qualitative methods helps to ensure that items are phrased in a way that conveys the meaning intended by the test developer. Moreover, specific problematic words, phrases, or instructions can be found and changed before subjecting the newly developed test to quantitative checks.

Conclusion

In this editorial, we highlight that test construction can gain substantially from a mixed methods approach. By preceding quantitative methods by qualitative methods, it is possible to ensure a deeper understanding of item content compared to applying quantitative methods alone, and avoid potentially negative influence of fuzzy concepts. We would like to emphasize that we do not argue for a substitution of quantitative methods. Instead, we are strongly convinced that qualitative methods are a valuable complement and combining quantitative and qualitative methods may substantially contribute to the psychometric quality of a test (e.g., see Kemper, 2010; Neuert & Lenzner, 2015; Ziegler, 2011).

A final thought is devoted to the necessity of defining the construct to be measured and its nomological net. Unless such definitions are available, it is impossible to judge whether the understanding of a test item matches the intentions of the test developer. Thus, applying cognitive interviews to psychometric tests targeting psychological constructs necessitates measurement intentions following the ABC of test construction (Ziegler, 2014; Ziegler, Kemper, & Kruyen, 2014) to be explicitly stated in the report of a test development: A. What is the construct being measured? B. What are the intended uses of the measure? C. What is the targeted population?

Therefore, the advice to authors would be to ensure that a cognitive pretest is embedded in a test construction strategy based on these principles.

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

Beatty, P. (2004). The dynamics of cognitive interviewing. In S. Presser, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, J. M. Rothgeb, & E. Singer (Eds.), *Methods for testing and evaluating survey questions* (pp. 45–66). New York, NY: Wiley.

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

Christie, R., Geis, F. L., & Berger, D. (1970). *Studies in machiavellianism*. New York, NY: Academic Press.

Gregory, R. J. (1996). *Psychological testing: History, principles and applications*. Boston, MD: Allyn & Bacon.

Holland, P. W., & Thayer, D. T. (1986). Differential Item Functioning and the Mantel-Haenszel Procedure. *ETS Research Report Series, 2*, 1–24.

Kemper, C. J. (2010). *Das Persönlichkeitsmerkmal Angstsensitivität: Taxon oder Dimension? Eine Analyse mit dem Mischverteilungs-Raschmodell* [The personality trait anxiety sensitivity: Taxon or dimension? An analysis with the mixed Rasch model]. Hamburg, Germany: Dr. Kovač.

Kemper, C. J., & Hock, M. (2015). *New evidence on the construct validity of the ASI-3 and the dimensional conceptualization of trait Anxiety Sensitivity from IRT modeling*. Manuscript submitted for publication.

Kemper, C. J., & Menold, N. (2014). Nuisance or remedy? The utility of stylistic responding as an indicator of data fabrication in surveys. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*, 92–99. doi: 10.1027/1614-2241/a000078

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213–236.

Meade, A. W., & Bartholomew, C. S. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355–383.

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist, 36*, 211–218.

Miller, K. (2011). Cognitive interviewing. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question evaluation methods: Contributing to the science of data quality* (pp. 51–75). New York, NY: Wiley.

Neuert, C., & Lenzner, T. (2015). Supplementing cognitive interviewing with eye tracking to pretest survey questions. Manuscript submitted for publication.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*, 109–130.

Rammstedt, B., & Beierlein, C. (2014). Can't we make it any shorter? *Journal of Individual Differences, 35*, 212–220. doi: 10.1027/1614-0001/a000141

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. doi: 10.1016/j.jrp.2010.11.006

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30*, 526–537.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Zadeh, L. A. (1976). A fuzzy-algorithmic approach to the definition of complex or imprecise concepts. *International Journal of Man-Machine Studies, 8*, 249–291.

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., Booth, T., & Bensch, D. (2013). Getting entangled in the nomological net. *European Journal of Psychological Assessment, 29*, 157–161. doi: 10.1027/1015-5759/a000173

Ziegler, M., & Kemper, C. J. (2013). Extreme response style and faking: Two sides of the same coin? In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers deviations in surveys – impact, reasons, detection and prevention* (pp. 217–233). Frankfurt a.M.: Peter Lang.

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales – Five misunderstandings and ways to overcome them. *Journal of Individual Differences, 35*, 185–189. doi: 10.1027/1614-0001/a000148

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

Christoph J. Kemper

Institute for Medical and Pharmaceutical Proficiency Assessment
Große Langgasse 8
55116 Mainz
Germany
E-mail contact@christoph-kemper.net

Timo Lenzner

GESIS – Leibniz-Institute for the Social Sciences
Survey Design and Methodology
Pretest Lab
PO Box 12 21 55
68072 Mannheim
Germany
Tel. +49 621 1246-227
Fax +49 621 1246-100
E-mail timo.lenzner@gesis.org

