Quality of Life Research (2018) 27:1671–1672
https://doi.org/10.1007/s11136-018-1886-4

EDITORIAL

Introduction to special section: test construction

Muirne C. S. Paap[1] · Jan R. Böhnke[2] · Carolyn E. Schwartz[3,4] · Frans J. Oort[5]

Published online: 25 May 2018

In the last few decades, it has been recognized that it is imperative to include quality of life (QoL) as an outcome measure in the evaluation of treatment effects. Typically, QoL—due to its subjective nature—is measured using patient-reported outcomes (PROs). PROs allow the clinician to gain insight into the way patients perceive their own health, and facilitate the evaluation of treatment effectiveness beyond the physical domain measured by clinical outcomes. Since QoL has become increasingly recognized as a key outcome, the development and application of PROs in clinical settings has increased tremendously. PRO development has been accompanied by an increased interest in test theory and psychometrics, including more advanced latent variable modeling techniques and test administration procedures. PRO developers have become increasingly concerned with selecting the most appropriate techniques to ensure test quality.

In response to these developments, we set out to publish a special section on test construction. Four of the papers included in this special section concern validity. Two of these papers draw attention to the trade off between reliability and validity. Both Smits et al. [1] and Choi and Van der Linden [2] observe that the focus in PRO construction is typically on reliability rather than validity. Whether one wants to optimize one over the other, or preferably both, depends on the goal of measurement [3], something about which QOL and PRO communities should be much clearer in test development and validation. Where Smits et al. used an example of a more traditional static questionnaire within the context of the classical test theory framework, Choi and Van der Linden focus on computerized adaptive testing (CAT) based on item response theory (IRT) [e.g., 4]. The intended use of test scores also plays an important role in the papers by Hawkins et al. [5] and Edwards et al. [6]. Both papers consider contemporary validity theory, where developing arguments regarding the proposed use of test scores plays a pivotal role. Hawkins et al. illustrate how contemporary validity theory can be applied to PRO measures. Edwards et al. [6] compare the psychometric approach to validity to that used by the US Food and Drug Administration (FDA), where the focus is on determining whether an instrument is "fit for purpose."

The second set of papers in this special section highlight advanced modeling techniques that can be employed to obtain reliable and unbiased test scores. One of the challenges in applying IRT to PROs is that IRT estimation methods require large sample sizes in order to produce stable and precise parameter estimates. Houts et al. [7] present and evaluate a potential solution: using longitudinal IRT modeling to boost measurement precision. Establishing whether items show measurement invariance across subpopulations is essential if meaningful interpretations of observed mean differences are to be obtained [4]. Edwards et al. [8] show the impact it can have on scores and ensuing conclusions based on those scores, if violations of measurement invariance are not taken into account. Sawatzky et al. [9] illustrate how latent variable mixture models can be used to select items that show measurement invariance in a situation where it is not known a priori which population characteristic might result in a breach of measurement invariance.

As Houts et al. [7] point out, the fields of psychology and educational measurement have a long-standing reputation and experience in the application of modern test theory and methods of test construction. The field of health measurement still has ground to make up, which is becoming

✉ Frans J. Oort
oort.qolr@uva.nl

1 Department of Special Needs, Education, and Youth Care, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

2 Dundee Centre for Health and Related Research, School of Nursing and Health Sciences, University of Dundee, Dundee, UK

3 DeltaQuest Foundation, Inc., Concord, MA, USA

4 Departments of Medicine and Orthopaedic Surgery, Tufts University School of Medicine, Boston, MA, USA

5 Research Institute of Child Development and Education, Faculty of Social and Behavioural Sciences, University of Amsterdam, Amsterdam, The Netherlands

ever more relevant given the increase in development and use of PROs. The papers in this special section show how modern approaches and innovative solutions can be implemented in our field. As editors, we would like to appeal to our readership to follow their lead. Innovation enriches our field and it encourages multidisciplinary development. CAT is a prime example of this. CAT was initially developed to support educational testing and intelligence testing in the military. Consequently, CAT literature focuses on models and algorithms that are relevant for these contexts. In developing CATs for health assessment, researchers have been inspired by and successfully taken advantage of the existing CAT research from other fields. However, some researchers have drawn attention to the fact that a transition to the IRT/CAT framework also poses a number of challenges that are specific to the health measurement field [e.g., 10, 11]. In a recent study, the consequences of these different design factors were evaluated and discussed [12], and the resulting findings can be used to design CATs that are optimally fine-tuned to the health measurement setting.

QoL/PRO research has always been a multidisciplinary endeavor, guided by the different disciplines involved in the clinical care for particular disorders and diseases and social science research methods to investigate the subjective views on health conditions by patients and those involved in their care and recovery. Other disciplines with a strong tradition in test construction and psychometrics can contribute to our conceptualisation and understanding of QoL. We hope that this special section inspires our readers to embrace methodological developments from the past decade. We would be delighted if the section were used as a launching pad for research on test construction in the context of health measurement.

References

1. Smits, N., Van der Ark, L. A., & Conijn, J. M. (2017). Measurement versus prediction in the construction of patient-reported outcome questionnaires: Can we have our cake and eat it? *Quality of Life Research*. https://doi.org/10.1007/s11136-017-1720-4.
2. Choi, S. W., & Van der Linden, W. J. (2017). Ensuring content validity of patient-reported outcomes: A shadow-test approach to their adaptive measurement. *Quality of Life Research*. https://doi.org/10.1007/s11136-017-1650-1.
3. Cronbach, L. J. (1954). Report on a psychometric mission to Clinicia. *Psychometrika, 19*, 263–270. https://doi.org/10.1007/BF02289226.
4. Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Lawrence Erlbaum Associates.
5. Hawkins, M., Elsworth, G. R., & Osborne, R. H. (2018). Application of validity theory and methodology to patient-reported outcome measures(PROMs): building an argument for validity. *Quality of Life Research*. https://doi.org/10.1007/s11136-018-1815-6.
6. Edwards, M. C., Slagle, A., Rubright, J. D., & Wirth, R. J. (2017). Fit for purpose and modern validity theory in clinical outcomes assessment. *Quality of Life Research*. https://doi.org/10.1007/s11136-017-1644-z.
7. Houts, C. R., Morlock, R., Blum, S. I., Edwards, M. C., & Wirth, R. J. (2018). Scale development with small samples: A new application of longitudinal item response theory. *Quality of Life Research*. https://doi.org/10.1007/s11136-018-1801-z.
8. Edwards, M. C., Houts, C. R., & Wirth, R. J. (2017). Measurement invariance, the lack thereof, and modeling change. *Quality of Life Research*. https://doi.org/10.1007/s11136-017-1673-7.
9. Sawatzky, R., Russell, L. B., Sajobi, T. T., Lix, L. M., Kopec, J., & Zumbo, B. D. (2017). The use of latent variable mixture models to identify invariant items in test construction. *Quality of Life Research*. https://doi.org/10.1007/s11136-017-1680-8.
10. Fayers, P. M. (2007). Applying item response theory and computer adaptive testing: The challenges for health outcomes assessment. *Quality of Life Research, 16 Suppl 1*, 187–194. https://doi.org/10.1007/s11136-007-9197-1.
11. Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*(1), 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553.
12. Paap, M. C. S., Born, S., & Braeken, J. (2018). Measurement efficiency for fixed-precision multidimensional computerized adaptive tests: Comparing health measurement and educational testing using example banks. *Applied Psychological Measurement*. https://doi.org/10.1177/0146621618765719.