Special Issue on Item Response Theory in Medical Studies

Item Response Theory (IRT) comprises a class of latent variable models and associated statistical procedures that connects observed item responses to an underlying construct[1,2]. IRT models are used primarily to measure a continuous latent variable from observed categorical item observations. The latent variable represents an underlying construct that cannot be measured directly, but it is estimated from the observed item responses. IRT provides a way to construct a common measurement scale on which the latent variable scores can be represented. Applications of IRT can be found in educational, social and health sciences, and include the measurement of quality-of-life[3] and the assessment of health status[4].

New technologies have stimulated further the usefulness of IRT. Large-scale assessment programs have been developed using digital technology (e.g., tablets, smartphones) to collect response data. This development created opportunities to implement computer adaptive testing[5,6] and to improve assessments in terms of item and test design, data analysis, and reporting of test results. For instance, the Patient-Reported Outcome Measurement Information System (PROMIS) comprehends a collection of freely available evaluation assessments for physical, mental and social domains. The PROMIS instruments provide a way to obtain standardized outcome measures and to monitor patient outcomes[7] using IRT. They are operationalized on a tablet computer, which is simple and convenient in a healthcare setting. With the ever-increasing availability of computer devices (smart devices), computer-assisted assessments are becoming the standard. Together with the increase in data storage opportunities, computer-assisted test administration becomes applicable for clinical trials and longitudinal observational studies.

These developments also created new computational, mathematical and statistical challenges. Due to the innovative types of data collection, more and more data are collected with complex dependence structures. To make reliable and accurate decisions from observed data, more general IRT models are needed to account for dependencies that go beyond the well-known data correlations caused by the clustering of item responses within persons and items. This special issue on IRT in medical studies highlights new developments in this research area.

In the first of six contributions, a latent growth model (LGM) is discussed in which the IRT model serves as a building block. Longitudinal questionnaire data is modeled with an IRT-based LGM to measure health-related constructs and to estimate individual growth trajectories. In this study of Gorter and co-authors, a comparison is made between the IRT-based LGM and an LGM with a classical test theory (CTT) measurement model. A short introduction is given to both measurement models. The most popular and frequently used sum scores under the CTT model is compared to IRT scores in a latent growth analysis to stress the importance of the measurement model choice and to show its influence on the LGM parameter estimates. A novel plausible value procedure is proposed to accommodate scaling differences between measured health constructs under IRT and CTT. Through a simulation study they show bias in latent growth parameter estimates, when using the CTT model with sum scores.

An interesting and relevant discussion is given by van Breukelen, who commented on the paper. In the second contribution, he claims that the explanations of Gorter and co-authors cannot be correct. They argue that differences in response patterns leading to the same sum score are ignored under CTT, which is the main reason for biased

LGM estimates. Van Breukelen replies that the sum score is the sufficient statistic for the latent (construct) variable. Therefore, the sum score contains all data information regarding the latent variable, which makes the remaining response pattern information irrelevant. In a theoretical exposition, he argues that the non-linear relationship between the item responses and the latent variable in the data-generating IRT model leads to bias, when analyzing the data with the linear CTT model. This discussion provides a nice overview of differences between the IRT and CTT-based LGM analysis, when using the sum score as a sufficient statistic.

In the third contribution, co-author Fox replies to this discussion by explaining that the sum score is often used as the sufficient statistic for the latent variable. However, in a repeated measurements analysis, other measurement occasions provide information about each occasion-specific measurement. In that case, the sum score representing the data information of one measurement occasion is not the sufficient statistic for the occasion-specific latent variable. A theoretical decomposition of the total variance for continuous item responses is presented without using the sum score as the sufficient statistic for the latent variable. This leads to improved insights in the bias detected by Gorter, Fox and co-authors.

In paper four, Gorter and co-authors extend the comparison between IRT and CTT by including effects of missing item responses. The popular approach is to use CTT-based scores for the latent variable in a longitudinal data analysis, but this approach is more complicated with missing item responses. For longitudinal patient-reported outcomes, estimation results can differ between CTT-based and IRT-based approaches in the presence of missing response data[8]. To improve the comparison between IRT and CTT, a novel CTT-based imputation model is proposed to deal with missing item responses, while accounting for the longitudinal growth of the latent variable. This CTT-based imputation method is compared to the more conventional method of predictive mean matching and to the IRT-based imputation method to deal with missing item responses in a latent growth analysis. Gorter and co-authors developed a method to assess effects of missingness on IRT- and CTT-based LGM estimates on a common scale through a multiple imputation method, which is integrated in their plausible value procedure. They report about differences in estimated trend effects between CTT-based and IRT-based imputations for different missing data conditions.

Assessing change in patient-reported outcomes is much more difficult, when the measurement instrument is not measurement invariant and its characteristics change over time. Statistical methods to examine the measurement invariance property of a measurement instrument have received much attention and developed methods are still under much debate[9]. In the fifth paper of this issue, Blanchin and co-authors consider the response shift problem. This response shift can occur when a patient's self-evaluation of the target construct changes over time for instance due to a change in internal standards of measurement of the patient. A response shift leads to a violation of measurement invariance of the measurement model and can lead to bias in the measurement of longitudinal change. Blanchin and co-authors compare IRT-based methods with a structural equation model (SEM) method to detect response shift at the level of item responses. The performance of the methods is compared and evaluated to draw conclusions about optimal strategies to deal with response shift at the item level.

In IRT, a distribution for the latent variable can be assumed to facilitate parameter estimation and/or to model the population distribution of respondents. Usually this distribution is assumed to be normal but recently IRT models have been proposed with more flexible latent variable distributions[10]. In the sixth paper, Smits and co-authors give an introduction to scenarios that can lead to a non-normal latent variable distribution, which includes observing a surplus of zero-patterns. They discuss two IRT models with a non-normal latent variable distribution to account for zero-inflation in observed data or more general for a skewed latent variable distribution. In a simulation study, item and person parameter estimates of the graded response model and the extended IRT models are examined under different response processes leading to a non-normal latent variable distribution. Their results and conclusions are applicable to the situation of an item bank development for health outcomes.

The six papers represent a cross-section of current research in IRT for medical studies but is certainly not a complete overview of this research field. Further developments in this area are expected due to the continual advancement in technology. The debate on the use of IRT in health research and practices is highlighted, and advanced comparisons between IRT-based and CTT-based approaches in longitudinal data analysis are shown. We hope that this special issue will be of use to interested researchers both in psychometric/statistical methods and in relevant applications. This issue might also serve as a point of reference for advanced IRT modeling, for instance when the response process is more complicated leading to complex response behavior.

Jean-Paul Fox
Department of Research Methodology, Measurement and Data Analysis, Faculty of Behavioral, Management and Social Sciences, University of Twente, The Netherlands.

References

1. van der Linden, W. J., & Hambleton, R. K. (Eds.). (2013). *Handbook of modern item response theory.* Springer Science & Business Media.
2. Embretson, S. E., & Reise, S. P. (2013). *Item response theory.* Psychology Press.
3. Titman, A. C., Lancaster, G. A., & Colver, A. F. (2016). Item response theory and structural equation modelling for ordinal data: Describing the relationship between KIDSCREEN and Life-H. *Statistical Methods in Medical Research, 25(5)*, 1892–1924. https://doi.org/10.1177/0962280213504177
4. Fries, J. F., Bruce, B., & Cella, D. (2005). The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. *Clinical and experimental rheumatology*, *23*(5), S53.
5. Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of Life Research, 16(1)*, 95-108.
6. Van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). Computerized adaptive testing: Theory and practice. Dordrecht, The Netherlands: Kluwer Academic.
7. Papuga, M. O., Dasilva, C., McIntyre, A., Mitten, D., Kates, S., & Baumhauer, J. F. (2018). Large-scale clinical implementation of PROMIS computer adaptive testing with direct incorporation into the electronic medical record. *Health Systems, 7(1)*, 1-12.

8. de Bock, E., Hardouin, J. B., Blanchin, M., Le Neel, T., Kubis, G., Bonnaud-Antignac, A., ... & Sebille, V. (2016). Rasch-family models are more valuable than score-based approaches for analysing longitudinal patient-reported outcomes with missing data. *Statistical Methods in Medical Research, 25(5)*, 2067-2087.
9. Van de Vijver, F., et al. (2019). Invariance analyses in large-scale studies. OECD Education Working Papers, No. 201, OECD Publishing, Paris. https://doi.org/10.1787/254738dd-en.
10. Azevedo, C. L., Bolfarine, H., & Andrade, D. F. (2011). Bayesian inference for a skew-normal IRT model under the centred parameterization. *Computational Statistics & Data Analysis, 55(1)*, 353-365.