Psychological Test and Assessment Modeling, Volume 57, 2015 (3), 338-341

Advances in Rasch modeling: New applications and directions
Guest Editorial

Steffen Brandt[1], *Mark Moulton*[2] & *Brent Duckor*[3]

In 1960 Georg Rasch helped open the field of Item Response Theory by the model that bears his name, distinguished by the use of a single parameter to model the relationship between item difficulty and person ability. Various extensions of this relatively simple model have been proposed since then and are regularly applied in assessments. By including additional parameters in order, for example, to model variation in item discriminations (2-PL) or variation in guessing probabilities (3-PL) (Birnbaum, 1968), these extensions model the observed data more exactly and in principle improve the fit to the data of the response probabilities used to calculate test scores. However, the gain in model fit (and arguably reliability for particular item types) has a cost: not only are these models more complex but the resulting test scores are also more difficult to interpret.

In the U.S., various stakeholders including courts and states have adopted the Rasch model, in part, because it leads to logical and transparent results. With the Rasch model all items are weighted equally in order to define the ability that is to be measured, whereas in the 2-PL- and 3-PL-model the weighting of the items is defined recursively within the estimation process. Wright and Masters (1981) distinguish the Rasch approach as one that meets the requirements of measurement science: fitting the data to the model is a principled method for testing the hypothesis that the variable is, in fact, stable and meaningfully structured. This point can be stated another way: when one tries to fit the model to the data, we lose the property of "specific objectivity" (Rasch, 1977). Wilson (2005) reminds us that the Rasch model framework is fundamentally important to the sorts of interpretations one can make in measurement science. When choosing and evaluating a measurement model, we should think of the geographic map: The idea of "location" of an item response with respect to the location of another item response only makes sense if that relative meaning is independent of the location of the respondent involved – i.e.,

[1] *Correspondence concerning this article should be addressed to:* Steffen Brandt, PhD, Art of Reduction, Wissenschaftszentrum Kiel, Fraunhoferstraße 13, 24118 Kiel, Germany; email: steffen.brandt@artofreduction.com

[2] Educational Data Systems

[3] San Jose State University

the interpretation of relative *locations needs to be uniform* no matter where the respondent is. This *invariance* requirement corresponds to the idea that an "inch represents a mile" or a “meter represents a kilometer” *wherever you are* on a geographical map. The Rasch model upholds these principles of measurement science and meets more squarely with common sense notions of fairness and order.

While the Rasch model requires that all items be equally discriminating in order to define the ability that is to be measured, the 2-PL- and 3-PL-models allow discrimination to vary across items and calculates it recursively as part of the estimation process. However, this also impacts the estimation of the item difficulty parameter, creating a sharp discontinuity between how Rasch and 2-PL/3-PL item difficulties can be interpreted. Because item discrimination is at least in part a property of how a particular sample of examinees interacts with an item and is not exclusively a property of the item, and because examinees vary across tests, the inescapable consequence is that the scores calculated using the 2-PL and 3-PL models are not guaranteed to be as generalizable across tests as scores calculated from data that is constrained to fit the requirements of the Rasch model, especially when the scores are based on the highly sample-dependent 3-PL model. This is in part the reason why the Rasch model is still applied and remains an indispensable tool in the psychometrician’s toolkit. In high-stakes tests such as licensure exams, judges and policymakers work to ensure that all items employed to make a consequential decision embody the same construct, i.e., that all items are equally discriminating. These stakeholders rely on the measurement scientists’ efforts to remove items that discriminate along different and possibly unknown dimensions. The convenience and parsimony offered by the less exacting 2-PL and 3-PL models for test construction is a high price to pay for losing the ability to claim comparability of scores across tests or to know where one is on the variable map and the distance from one location to another.

The four papers we gathered for the special topic “Advances in Rasch Modeling: New Applications and Directions” consider the Rasch model from very different angles.

The first paper from Hagquist and Andrich (2015) shows how the interpretation of parameter estimates is not always as clear as one might assume. The case of artificial differential item functioning (DIF) proves that sometimes parameter estimates vary as a result of purely statistical artifacts. In this case the observed effects depend, for example, on the distribution of the item difficulties rather than actually existing DIF due to gender. We think a general method for identifying such statistical artifacts is essential in order to avoid possible misinterpretations of DIF and possibly other item-related parameters such as difficulty, discrimination, and “guessing.”

The second paper from Salzberger (2015) also considers the estimation of the difficulty parameters, here however for rating scale items. Salzberger describes different approaches to analyze the ordering of the thresholds of a rating scale item (in which the thresholds represent the difficulties of the different categories of a rating scale item) and suggests an additional approach in order to verify that the observed ordering is in accordance with the theory of the construct that is to be measured. In our opinion the presented approaches are an important means to check the construct validity of items and therefore of the test. An increasingly important issue in the last few years, validity theory still offers few

practical guidelines on how to check test validity. Salzberger's approach might prove a valuable contribution in this regard.

In the third paper, Torres, Diakow, Freund, & Wilson (2015) propose a new model, the Latent Class Level Partial Credit Model (Latent Class L-PCM). The presented model supports identifying and interpreting latent classes of respondents according to empirically estimated performance levels. In educational assessment, there is an increasing desire to document a student's performance not only quantitatively, as a scale score or ranking, but qualitatively, as a description corresponding to a specific level of performance. We think that the Latent Class L-PCM can be quite useful in helping experts to identify actually existing performance levels and to interpret them properly, using the actual performances of the students at each performance level.

The final paper from Wind (2015) presents a method based on Mokken scaling that supports examining data in terms of the basic requirements for invariant measurement, which is assumed in the Rasch model. We included this paper into this special topic for two reasons: (1) we think it is essential for any application of a model to validate its compliance with the model's basic assumptions, "invariance" in the case of the Rasch model; and (2) we feel that this paper illustrates the importance of looking "outside of the box" – in this case the box of parametric measurement. As in the real world, it is sometimes only possible to judge certain characteristics from an outside perspective.

In accordance with the traditions established by the predecessor series *Objective Measurement: Theory into Practice* (Vols. 1-5) and *Advances in Rasch Measurement* (Vols. 1-2), we are pleased to offer both the theoretical and practical applications of Rasch measurement models in this journal. All papers were originally presented at the International Objective Measurement Workshop (IOMW) 2014 in Philadelphia, PA, USA and were solicited for thematic coherence and fit. Each manuscript was blind reviewed by at least two experts.

The IOMW is a bi-annual conference which takes place before the conference of the American Educational Research Association (AERA). We look forward to readers joining the next IOMW conference to be held in Washington D.C. in spring 2016.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 397-479). Reading, MA: Addison-Wesley.

Hagquist, C., & Andrich, D. (2015). Determinants of artificial DIF – a study based on simulated polytomous data. *Psychological Test and Assessment Modeling, 3*, 342-376.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rasch, G. (1977). On Specific Objectivity: An attempt at formalizing the request for generality and validity of scientific statements. The Danish Yearbook of Philosophy, 14, 58-93.

Salzberger, T. (2015). The validity of polytomous items in the Rasch model – The role of statistical evidence of the threshold order. *Psychological Test and Assessment Modeling, 3*, 377-395.

Torres, D., Diakow, R., Freund, R., & Wilson, M. (2015). modeling for directly setting theory-based performance levels. *Psychological Test and Assessment Modeling, 3*, 396-422.

Wind, S. (2015). Evaluating the quality of analytic ratings with Mokken scaling. *Psychological Test and Assessment Modeling, 3*, 423-444.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Routledge.

Wright, B. D., & Masters, G. N. (1981). *The measurement of knowledge and attitude*. Statistical Laboratory, Department of Education, University of Chicago.