report

The opportunities Item Response Theory (IRT) offers to health psychologists

Methods in Health Psychology Symposium IV

Alexandra Dima
University of Amsterdam
Chris Gibbons
University of Manchester
Mieke Kleppe
Philips Research and Eindhoven University of Technology
Katarzyna Byrka
University of Social Sciences and Humanities of Wroclaw
Marijn de Bruin
University of Aberdeen
Marie Johnston
University of Aberdeen

The 4th methods in Health Psychology Symposium

(Marijn de Bruin)

The 4th Methods in Health Psychology symposium at the EHPS in Innsbruck was organised by Alexandra Dima (University of Amsterdam), who together with Chris Gibbons (University of Manchester), Mieke Kleppe (Phillips Research and Eindhoven University of Technology), and Katarzyna Byrka (University of Social Sciences and Humanities of Wroclaw) organised an excellent and inspiring symposium on using Item Response Theory for developing and validating questionnaires and theory testing. The presentations were diverse: from an introduction to IRT to examples of Rasch models and Mokken scale analyses. Prof. Marie Johnston, who has extensive expertise on measurement issues, closed the session as a discussant. A summary of this symposium is presented below. To give you an impression of the relevance of the issues raised: in the well-filled Kaiser-Leopold-Saal, about 90% of the audience indicated to know little about IRT prior to the symposium; after the symposium, about 90% said they would attend an IRT workshop if one would be organised at a future EHPS conference. No doubt this article will be similarly compelling.

Background to Item Response Theory

(Alexandra Dima, Chris Gibbons, Mieke Kleppe, Katarzyna Byrka)

Researchers in health psychology rely on questionnaires to measure abstract constructs such as people's illness perceptions, treatment beliefs, relationship styles, depression, medication adherence and quality of life. It is therefore essential that reliable and valid questionnaire measures are available to allow health psychology researchers to produce high-quality evidence. Questionnaires are often developed and validated using a number of techniques including exploratory or confirmatory factor analyses and internal consistency (Cronbach's alpha), which are usually referred to as Classical Test Theory (CTT). CTT methods are considerably enhanced when used alongside additional psychometric techniques such as item response theory (IRT) and Rasch analyses.

The IRT approach to psychometric analysis includes a number of different but related techniques, including Mokken Analysis (Mokken, 1971), Samejima's Graded Response Model analysis (Samejima, 1969) and Rasch Analysis (Rasch, 1960). These methods have been shown to result in refined, accurate and concept-relevant questionnaires that are often shorter than measures developed with CTT. These IRT-derived measures also allow researchers to perform hypothesis testing from a different and often more theoretically-appropriate angle.

The added value that IRT approaches can bring to health psychology lies mainly in their different assumptions regarding measurement, compared to

more familiar CTT methods. CTT techniques are often based solely on correlations between questionnaire items. This statistical approach assumes that items are interchangeable and have a uniform relationship with the phenomena they represent. This assumption often does not hold in practice, as many concepts in health psychology consist of attributes that fall on a continuum which may be ordered from lowest to highest (e.g., impairment, quality of life or strength of beliefs). For such concepts, questionnaire development should take into account the way in which items are ordered to represent the underlying latent phenomenon. IRT and related methods allow the researcher to account for item ordering, and thus better understand the content of the construct and estimate individual scores more precisely. This, in turn, reduces measurement error and increases the statistical power to test the substantive relationships of interest (Fries, Krishnan, Rose, Lingala, & Bruce, 2011).

Due to the many advantages in questionnaire construction and theory testing (see Box 1) IRT offers, it is considered the gold standard approach in many academic disciplines and is quickly becoming the preferred methodology for large-scale projects in health outcomes measurement (e.g., Fries, Bruce, & Cella, 2005; Power, Quinn, Schmidt, & the WHOQOL-Old Group, 2005; Ravens-Sieberer et al., 2008). The purpose of the 4th EHPS Methods in Health Symposium was to promote IRT approaches by summarising the opportunities that they offer for both theory and questionnaire development and encourage a wider adoption of IRT techniques to improve the rigour of quantitative research in health psychology.

BOX 1: IRT for Health Psychology - Key advantages

1. *IRT improves measurement properties:* reliable and valid construct measurement leads to more accurate substantive findings, and thus to better evidence-based real-life decisions.
2. *IRT offers more sensitive detection of change due to interventions:* tools with good measurement properties at all levels of the latent continuum give better chances of detecting change when it happens.
3. *IRT enables a different, additional approach to theory testing:* IRT can be used for theory development and testing (not only for tool development).
4. *IRT helps find parsimonious measurement models:* obtains simple structure for data that within standard models (e.g., factor analysis) appear complex.
5. *IRT reduces measurement costs for funders and respondents:* lower measurement error leads to lower variability in scores, and thus to more precision in estimates and ability to identify significant differences with fewer subjects or less restricted inclusion criteria.
6. *IRT software is readily available:* software now available with good documentation, for example several IRT packages are free to use in the R environment.
7. *IRT is increasingly used:* researchers in health psychology have started to use IRT more for both measure development and theory testing.
8. *Using IRT is easier than expected:* the most important thing is to understand the principles, the method itself is easy to learn and one can also collaborate with IRT-trained researchers.

How can item-response theories improve questionnaire research in health psychology?

(Chris Gibbons)

In the first presentation of the symposium, Chris Gibbons (University of Manchester) discussed some of the scientific and practical advantages of IRT by introducing the IRT concepts of category threshold ordering, interval scaling, scale targeting, dimensionality, differential item functioning, and computer adaptive testing (Gibbons et al., 2013; Revicki & Cella, 1997).

Category threshold ordering

Most questionnaires used in health psychology have items with multiple response options, often using Likert scaling, and respondents rate their agreement to each item along a series of ordered responses. Category threshold ordering allows researchers to ensure that these responses are properly ordered by analysing how people respond to items. If certain response categories are systematically ignored the scoring structure for the item becomes inaccurate and unreliable. Category thresholds may become disordered when respondents do not distinguish between all of the possible response categories, indicating the need to rescore the item or to choose a more suitable response format. Such analyses help researchers choose the optimal response format for their questionnaires that strikes a perfect balance between item information and participant burden.

Interval scaling

For questionnaires to achieve the highest standard of measurement, it is crucial that resulting scores are intervally scaled. A questionnaire is intervally scaled when the scores are properly ordered and the difference between scores (the intervals) are uniform throughout the scale (e.g., the difference between scores of 1 and 2 is the same as the difference between scores of 11 and 12). Scales must provide interval-level measurement if individual item scores are to be added together and used in arithmetic operations and parametric statistics (Karabatsos, 2001). Classical test theories are only capable of creating ordinal-level measurement, which does not guarantee consistent intervals between scale scores and therefore does not meet the additivity requirement of fundamental measurement (Wright, 1992). To assist researchers in gaining interval-level estimates from questionnaire research, IRT models can provide a 'conversion rate' to transfer raw scale scores into interval-level estimates to fully satisfy the conditional requirements of statistical tests (Gibbons et al., 2013).

Scale targeting

Scale targeting lets a researcher know how well matched a certain set of items is to the target population. Such analyses can also indicate if questionnaires are not providing information about certain members of a population. If the scale information is not closely matched to the population we are unable to gain much information about the population we are interested in. Developing and selecting items on IRT principles and item diagnostics leads to constructing questionnaires that can obtain maximum information on the construct of interest in the target population (see the medication adherence study below).

Unidimensionality

Unidimensionality is another important characteristic of questionnaire measures. A scale is unidimensional when all items correspond to the same underlying construct. The stricter tests of dimensionality provided in IRT models are better placed to give an accurate view of dimensionality than factor analyses, and thus more appropriate for testing structural validity of questionnaires (see the study on the dimensionality of health performance below; for another example see Tennant and Pallant

(2006)

Differential item functioning

Differential item functioning (DIF; Holland & Weiner, 1995) occurs when different groups respond differently to certain items for reasons other than differences in their level of the underlying trait. For example, on a fatigue scale for patients with motor neurone disease men were more likely than women to agree with the item "I wake up in the night on most nights", irrespective of their underlying level of fatigue. This suggested that there was some reason other than fatigue that men frequently woke during the night, possibly due to nocturia which is present in more than half of elderly men (Gibbons et al., 2011; Jackson, 1999). Failure to identify DIF may result in erroneous disparities in scale scores between demographic groups that may be wrongly attributed to the trait being measured. Identifying and excluding items with DIF allows unbiased comparisons between different subgroups (e.g., based on gender, age or nationality).

Computer adaptive testing

Of special interest are the close links that IRT and Rasch analysis have with computer adaptive testing (CAT), an exciting method for administering questionnaires that has a number of significant advantages over pencil-and-paper questionnaires. CAT is a technique for the electronic administration of questionnaires that significantly reduces questionnaire length and response burden (Haley, Raczek, Coster, Dumas, & Fragala-Pinkham, 2005; Wainer, 2000). This is achieved through selective item administration based on each individual's previous responses, and omitting irrelevant items based on individual characteristics (e.g., patient's disease group or other demographic factor) (Ware, Bjorner, & Kosinski, 2000; Weiss, 1985). CAT questionnaires take a fraction of the time to complete and can be just as reliable, valid and sensitive to change when compared to their paper-based counterparts (Haley et al., 2005). Instantaneous calculation of questionnaire scores, including comparison with previous scores and graphical feedback, is also achievable using CAT platforms, thus increasing their simplicity for use in time-pressured clinical environments.

There are numerous other practical and methodological advantages to IRT and Rasch analysis that are described more comprehensively elsewhere (Pallant & Tennant, 2007).

Using the Rasch model to compare medication adherence questionnaires

(Mieke Kleppe)

One construct that has proven difficult to measure using self-report questionnaires is medication adherence (i.e., the extent to which medication is taken as prescribed by a physician). Recent research on measuring medication adherence using self-reported measures provides a perfect illustration of how 'scale targeting' using IRT can enable researchers to develop a better assessment tool. Commonly-used self-report measures often provide heavily skewed results with limited variance, suggesting that most participants are highly adherent to their prescribed medication. This finding contrasts with results of objective adherence measures which indicate that many people are non-adherent (Nguyen, Caze, & Cottrell, 2014; Reach et al., 2011; Vermeire, Hearnshaw, Van Royen, & Denekens, 2001). In the second presentation, Mieke Kleppe (Philips Research and Eindhoven University of Technology) argued that a possible explanation for these results is that these adherence questionnaires cover a restricted range of adherence behaviours. That is, the items do not match the non-adherence behaviours people perform (i.e., they are too easy for the sample). In developing these questionnaires researchers implicitly assumed that item difficulties are similar for all items and did not take into account that for example forgetting a pill might occur more often (and it is thus more difficult to report being adherent on this item) than stop taking pills for a whole week. To resolve this

issue, we developed a new item set (the ProMAS), aiming specifically to cover a broader range of difficulties (Kleppe, Lacroix, Ham & Midden, 2014). Winsteps software (Linacre, 2007) was used for all Rasch analyses, including the calculation of adherence estimates, item and person fit statistics and dimensionality analyses.

A study was conducted among elderly taking medication for chronic conditions (N = 370). A selection of the items was made to shorten the scale based on fit statistics and item difficulties, and 18 items remained in the final scale. The final item set of the ProMAS was compared to the Medication Adherence Report Scale (MARS), one of the most frequently used current adherence measures. The ProMAS adherence estimates were less skewed and provided more variance than the MARS adherence scores. To test whether the ProMAS item difficulties covered a wider range of non-adherence behaviours than the MARS, items from both scales were entered into one Rasch analysis. Results indicated that the ProMAS items cover a wider range of item difficulties that are better matched to participants' behaviours. While the MARS only provided one item to distinguish between the 50% most adherent patients, the ProMAS provided six items. These items are most relevant for distinguishing between participants with higher adherence scores. The wider item difficulty range resulted in adherence scores that better accord with those obtained with objective adherence measures in previous studies. This study showed that using the IRT concept of scale targeting, questionnaires can be developed that are better capable of discriminating participants on the variable of interest. In this case, the Rasch model provided the statistical tools to obtain an improved measure of medication adherence.

Health performance within the Campbell paradigm: IRT models for testing new approaches in health psychology

(Katarzyna Byrka)

Beyond its psychometric value, IRT offers unique solutions for testing novel theories. In the third presentation, Katarzyna Byrka (University of Social Sciences and Humanities of Wroclaw) described how examining dimensionality with IRT models provokes a paradigm shift in thinking about interdependence of health behaviors.

In health psychology, it is believed that behaviors such as screening for cancer, calorie counting or fastening seatbelts do not belong to a single, general behavioral class (Stroebe & Stroebe, 1995). The independence of health behaviors, however, has been judged on the basis of correlations and related methods such as factor analysis. The problem is that correlations between behaviors are likely to be artificially deflated when examined items differ significantly in their 'difficulties' (i.e., the percentage of people that perform a certain behavior). Consequently, meaningful psychological interpretation of the data structure using correlations is only possible if examined items are homogeneous with respect to their difficulty (Ferguson, 1941). Obviously, health behaviors differ in 'difficulty', as the costs (both figurative and literal) of performing some is higher than of others; for example, light exercise 15 minutes per day bears far less behavioural cost than jogging (Kaiser, Byrka, & Hartig, 2010). In such situations when items are not homogenous the complex and the multidimensional structure of the data obtained with factor analyses likely stems from a statistical artifact.

Contrary to common findings in health psychology, a recently developed approach, the Campbell paradigm, assumes that all specific health behaviors

are interdependent and belong to a single behavioral class (Byrka & Kaiser, 2013). In this approach, the interdependence of behaviors is conceptualized as steps of variable difficulty undertaken by people to achieve a particular goal (i.e., being healthy). Such an assumption can be only tested within a model that takes into account the 'difficulty' of behaviors.

In a cross-sectional study with a sample of Dutch adults (N = 396) a one-parameter logistic Rasch model was applied to corroborate the assumptions of the Campbell paradigm (Byrka & Kaiser, 2013). Specifically, unidimensionality of a comprehensive health performance measure composed of behavioral self-reports was tested. It was found that health behaviors associated with different domains such as sustenance, hygiene, and physical exercise formed a homogenous class. A more complex five-dimensional model, a multidimensional extension of a one-parameter Rasch model (Adams, Wilson & Wang, 1997), did not predict the data meaningfully better than a parsimonious one-dimensional version (the models were compared within the Conquest software). Additionally, the same data were explored using factor analysis. Out of fifty items ten dimensions were generated based on eigenvalues above 1. As a result, some dimensions appeared of rather poor psychometric quality as they were composed of only one or two items. In sum, these findings speak of the unity of health performance when explored with IRT and of multidimensional complexity when explored with factor analysis.

In sum, applying psychometric models stemming from IRT is the best solution to find unbiased relations between behaviors that are heterogeneous in difficulty (Embretson & Reise, 2000). Moreover, to test certain theoretical assumptions, such as the Campbell paradigm, IRT models are the only conceptually-appropriate methods. IRT allows the researchers to find parsimonious models and simple structures for data that within standard CTT models appear complex, and thus help minimize the ongoing segmentation of the field of health psychology (Schwarzer, 2008) and lead researchers to derive more meaningful models.

Mokken Scaling Analysis: scale development the NIRT way

(Alexandra Dima)

Developing questionnaires that achieve fundamental measurement and thus fully meet the requirements of parametric statistical tests is one of the main advantages of using IRT approaches, and is best achieved by parametric methods such as Rasch modeling. However for some psychological concepts and measures this may not always be an attainable goal. One reason is that some concepts may by definition only refer to differences in degree between cases (be it people, groups, or events); for these concepts, no matter how well the items have been developed, the data does not fit a parametric IRT model. Moreover, in some research settings data can only be collected for a few items (e.g., surveys with numerous scales) and from fewer cases (e.g., small population, low resources); for these datasets, no matter how well the measurement was performed, the study will be underpowered for a parametric IRT analysis. In the fourth presentation of the symposium, Alexandra Dima (University of Amsterdam) described how non-parametric IRT (NIRT) allows the researcher to account for item ordering using less restrictive models, and how NIRT can also provide complementary information for questionnaires that do fit parametric models (for a more detailed introduction see Sijtsma, 1998).

The best developed and most accessible NIRT method is Mokken Scaling Analysis (MSA; Schuur, 2003); a well-documented software package (*mokken*) is available for the R environment (Ark, 2007) and produces several quite intuitive and easily interpretable outputs. Homogeneity (H) indicates to what degree an item, a pair of items or a scale can be considered as reflecting a single latent dimension.

When a more exploratory approach to dimensionality testing is needed, an automatic item selection procedure (aisp) can cluster items to optimize scale homogeneity given increasing homogeneity thresholds (c), and gives an informative overview on the dimensionality of item sets (Hemker, Sijtsma, & Molenaar, 1995). As its parametric equivalents, MSA also allows examination of various item properties (e.g., via Item Response Function graphs). These outputs make MSA a useful and flexible research tool for many health psychology topics, such as exploring health communication processes, assessing patient preferences, and understanding the structure and quality of medical or socio-behavioral care.

A health communication process that lends itself to NIRT analyses is diagnosis disclosure, an incremental process of shifting from no disclosure to being completely open about one's diagnosis. This may happen as a single process, or there might be several distinct processes that include specific groups or single individuals. A recent study used MSA to examine the dimensionality of HIV status disclosure in people living with HIV in Tanzania and identified several distinct voluntary disclosure processes (to spouse, children, family members, and larger community), each showing different patterns of association with relevant concepts such as stigma and social support; these differences would be overlooked if sum scores were used (Dima, Stutterheim, Lyimo, & de Bruin, 2014). This new way of examining disclosure processes can inform more targeted disclosure counseling and may prove informative for studying disclosure in other contexts.

Another phenomenon that can be conceptualized as an ordered item set refers to patients' treatment beliefs; these may range from being strongly against to strongly supportive of a treatment. In a recent study of low back pain treatments, the aisp analysis allowed a comparative examination of patients' beliefs about four treatments recommended in UK primary care: medication, exercise, manual therapy and acupuncture (Dima et al., 2013). We examined four themes hypothesized as distinct dimensions (concerns, credibility, effectiveness and individual fit) and found that the distinction between concerns and the other three (closely-related) themes is more salient regarding medication, but applies less to acupuncture, exercise and manual therapy. The findings suggest that the cost-benefit dichotomy in treatment decision-making may not apply to a broader range of treatments beyond medication, and highlights the usefulness of NIRT in investigating dimensionality of patients' beliefs.

Quality of care is yet another phenomenon intuitively described as a set of activities of increasing difficulty, from basic to more advanced care, for which NIRT can prove useful. Within a large ongoing observational cohort study on asthma treatment, reports of medical care and adherence support activities by French general practitioners were examined via MSA. Preliminary results showed that, while medical care activities do not form a single dimension, several key adherence support activities can be ordered from basic to comprehensive support and form a scale showing significant associations with relevant determinants of adherence support (Dima, van Ganse, Le Cloarec, de Bruin, & the ASTRO-LAB group, 2014). This encourages the development and use of NIRT-based questionnaires to assess quality of care.

These are just three situations in which MSA can offer relevant insights into the data and lead to novel interpretations. MSA has been used in health research for several decades, and many other examples are available (Watson et al., 2012).

How can the use of IRT methods be enhanced in health psychology?

(Marie Johnston)

Papers using IRT have been presented intermittently at EHPS for at least ten years. The papers presented in this symposium provide very persuasive arguments about the potential gains for

health psychology of using these methods to improve measurement, provide additional methods of testing theory, and reduce measurement burden. IRT methods can be used alongside the psychometric methods of CTT to achieve more sensitive, accurate measurement, and to reject redundant, insensitive and irrelevant items. All of this is clearly of considerable value not only in health psychology but in areas of routine professional practice where reducing respondent burden while retaining sensitive accurate measurement may be particularly valuable.

Given the immense potential value of IRT, why has the use of these methods been so sporadic in health psychology? The symposium audience was almost unanimous in agreeing that IRT methods held considerable potential and were likely to result in good research outcomes, but had low confidence in using the methods. Using a Social Cognitive Theory analysis of our behavior, outcome expectancies were high but self-efficacy was low with the result that the methods are not frequently used. Considering Bandura's four methods of enhancing self-efficacy, the symposium presentations had included persuasive messages and had modelled successful vicarious experiences of using IRT. However perhaps we need more mastery experiences and opportunities to reduce our emotional responses to these sophisticated analytic methods.

It was therefore proposed that a workshop on IRT methods will be organized prior to a future EHPS conference. We encourage readers interested in increasing their IRT self-efficacy and skills to look out for announcements.

References

Ark, L. A. van der. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1–19.

Byrka, K., & Kaiser, F. G. (2013). Health performance of individuals within the Campbell paradigm. *International Journal of Psychology: Journal International De Psychologie, 48*(5), 986–999. doi:10.1080/00207594.2012.702215

Dima, A. L., Lewith, G., Little, P., Moss-Morris, R., Foster, N. E., Hankins, M., & Bishop, F. L. (2013). Beyond medication beliefs: A comparative NIRT analysis of patients' beliefs on four back pain treatments. *Psychology & Health, 28*(sup1), 87. doi:10.1080/08870446.2013.810851

Dima, A. L., Stutterheim, S. E., Lyimo, R., & de Bruin, M. (2014). Advancing methodology in the study of HIV status disclosure: The importance of considering disclosure target and intent. *Social Science & Medicine, 108*, 166–174. doi:10.1016/j.socscimed.2014.02.045

Dima, A. L., van Ganse, E., Le Cloarec, H., de Bruin, M., and the ASTRO-LAB group (2014, November). *Adherence support in routine asthma care: development and validation of a clinician-report tool.* Poster presented at the 17th annual meeting of the European Society for Patient Adherence, Compliance and Persistence, Lausanne, Switzerland.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum Publishers.

Ferguson, G. A. (1941). The factorial interpretation of test difficulty. *Psychometrika, 6*(5), 323–329. doi:10.1007/BF02288588

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, Suppl 39), S53–57.

Fries, J. F., Krishnan, E., Rose, M., Lingala, B., & Bruce, B. (2011). Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. *Arthritis Research & Therapy, 13*(5), R147. doi:10.1186/ar3461

Gibbons, C. J., Kenning, C., Coventry, P. A., Bee, P., Bundy, C., Fisher, L., & Bower, P. (2013). Development of a Multimorbidity Illness Perceptions Scale (MULTIPleS). *PLoS ONE, 8*(12), e81852. doi:10.1371/journal.pone.0081852

Gibbons, C. J., Mills, R. J., Thornton, E. W., Ealing, J., Mitchell, J. D., Shaw, P. J., … Young, C. A. (2011). Development of a patient reported outcome measure for fatigue in motor neurone disease: the Neurological Fatigue Index (NFI-MND). *Health and Quality of Life Outcomes, 9*(1), 101. doi:10.1186/1477-7525-9-101

Haley, S. M., Raczek, A. E., Coster, W. J., Dumas, H. M., & Fragala-Pinkham, M. A. (2005). Assessing mobility in children using a computer adaptive testing version of the pediatric evaluation of disability inventory. *Archives of Physical Medicine and Rehabilitation, 86*(5), 932–939. doi:10.1016/j.apmr.2004.10.032

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of Unidimensional Scales From a Multidimensional Item Bank in the Polytomous Mokken I RT Model. *Applied Psychological Measurement, 19*(4), 337–352. doi:10.1177/014662169501900404

Holland, P., & Weiner, H. (1995). *Differential Item Functioning*. Philadelphia, PA: Lawrence Erlbaum.

Jackson, S. (1999). Lower urinary tract symptoms and nocturia in men and women: prevalence, aetiology and diagnosis. *BJU International, 84*(S1), 5–8. doi:10.1046/j.1464-410X.84.s1.6.x

Kaiser, F. G., Byrka, K., & Hartig, T. (2010). Reviving Campbell's Paradigm for Attitude Research. *Personality and Social Psychology Review, 14*(4), 351–367. doi:10.1177/1088868310366452

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement, 2*(4), 389–423.

Kleppe, M., Lacroix, J. P. W., Ham, J. R. C. & Midden, C. J. H. (2014). *The development of the ProMAS: A Probabilistic Medication Adherence Scale*. Manuscript submitted for publication.

Linacre, J. M. (2007). *Winsteps* (version 3). Chicago. Retrieved from www.winsteps.com

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in Political Research*. New York: de Gruyter (Mouton).

Nguyen, T.-M.-U., Caze, A. L., & Cottrell, N. (2014). What are validated self-report adherence scales really measuring?: a systematic review. *British Journal of Clinical Pharmacology, 77*(3), 427–445. doi:10.1111/bcp.12194

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*(1), 1–18. doi:10.1348/014466506X96931

Power, M., Quinn, K., Schmidt, S., & the WHOQOL-Old Group. (2005). Development of the WHOQOL-Old Module. *Quality of Life Research, 14*(10), 2197–2214. doi:10.1007/s11136-005-7380-9

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Ravens-Sieberer, U., Gosch, A., Rajmil, L., Erhart, M., Bruil, J., Power, M., … Kilroe, J. (2008). The KIDSCREEN-52 Quality of Life Measure for Children and Adolescents: Psychometric Results from a Cross-Cultural Survey in 13 European Countries. *Value in Health, 11*(4), 645–658. doi:10.1111/j.1524-4733.2007.00291.x

Reach, G., Michault, A., Bihan, H., Paulino, C., Cohen, R., & Le Clésiau, H. (2011). Patients' impatience is an independent determinant of poor diabetes control. *Diabetes & Metabolism, 37*(6), 497–504. doi:10.1016/j.diabet.2011.03.004

Revicki, D. A., & Cella, D. F. (1997). Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *Quality of Life Research, 6*(6), 595–600. doi:10.1023/A:1018420418455

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100.

Schuur, W. H. van. (2003). Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory. *Political Analysis, 11*(2), 139–163. doi:10.1093/pan/mpg002

Schwarzer, R. (2008). Modeling health behavior change: How to predict and modify the adoption and maintenance of health behaviors. *Applied*

Psychology, 57(1), 1–29. doi:10.1111/j.1464-0597.2007.00325.x

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22*(1), 3–31. doi:10.1177/01466216980221001

Stroebe, W., & Stroebe, M. S. (1995). *Social psychology and health*. Buckingham: Open University Press.

Tennant, A., & Pallant, J. F. (2006). Unidimensionality matters! (a tale of two Smiths?). *Rasch Measurement Transactions, 20*(1), 1048–1051.

Vermeire, E., Hearnshaw, H., Van Royen, P., & Denekens, J. (2001). Patient adherence to treatment: three decades of research. A comprehensive review. *Journal of Clinical Pharmacy and Therapeutics, 26*(5), 331–342. doi:10.1046/j.1365-2710.2001.00363.x

Wainer, H. (2000). *Computer adaptive testing: A primer.* Hillsdale, NJ: Earlbaum Associates.

Ware, J. E., Bjorner, J. B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. *Medical Care, 38*(9 Suppl), II73–82.

Watson, R., van der Ark, L. A., Lin, L.-C., Fieo, R., Deary, I. J., & Meijer, R. R. (2012). Item response theory: How Mokken scaling can be used in clinical practice. *Journal of Clinical Nursing, 21*(19-20), 2736–2746. doi:10.1111/j.1365-2702.2011.03893.x

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*(6), 774–789. doi:10.1037/0022-006X.53.6.774

Wright, B. D. (1992). Raw Scores Are Not Linear Measures: Rasch vs. Classical Test Theory CTT Comparison. *Rasch Measurement Transactions, 6*(1), 208.

Alexandra Dima
Department of Communication Science, University of Amsterdam, The Netherlands
a.dima@uva.nl

Chris Gibbons
NIHR Collaboration for Applied Health Research and Care, Centre for Primary Care, Institution of Population Health, University of Manchester, UK
chris.gibbons@manchester.ac.uk

Mieke Kleppe
Philips Research, Eindhoven, the Netherlands; Eindhoven University of Technology, Department of Human-Technology Interaction, School of Innovation Sciences, Eindhoven, The Netherlands
m.kleppe@tue.nl

Katarzyna Byrka
University of Social Sciences and Humanities of Wroclaw, Poland
kbyrka@swps.edu.pl

Marijn de Bruin
Aberdeen Health Psychology Group,
Institute of Applied Health Sciences,
College of Life Sciences and
Medicine, Aberdeen, UK
m.debruin@abdn.ac.uk

Marie Johnston
Aberdeen Health Psychology Group,
Institute of Applied Health Sciences,
College of Life Sciences and
Medicine, Aberdeen, UK
m.johnston@abdn.ac.uk