Journal of Psychiatric Research 47 (2013) 141–148

Contents lists available at SciVerse ScienceDirect

Journal of Psychiatric Research

journal homepage: www.elsevier.com/locate/psychires

An introduction to Rasch analysis for Psychiatric practice and research

Neusa Sica da Rocha [a,b,*], Eduardo Chachamovich [c], Marcelo Pio de Almeida Fleck [a,d], Alan Tennant [e]

[a] *Hospital de Clinicas de Porto Alegre, Programa de Pós Graduação em Ciências Médicas: Psiquiatria, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil*
[b] *Post-doc Programa de Pós Graduação em Ciências Médicas: Psiquiatria, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil*
[c] *Department of Psychiatry, McGill University, Douglas Mental Health University Institute, Canada*
[d] *Departamento de Psiquiatria e Medicina Legal da Universidade Federal do Rio Grande do Sul (UFRGS), Brazil*
[e] *Department of Rehabilitation Medicine, The University of Leeds, UK*

A R T I C L E I N F O

Article history:
Received 28 February 2012
Received in revised form
17 August 2012
Accepted 20 September 2012

Keywords:
Rasch analysis
Patient reported outcomes
Scales
Psychometrics
Validity

A B S T R A C T

This article aims to present the main characteristics of Rasch analysis in the context of patient reported outcomes in Psychiatry. We present an overview of the main features of the Rasch analysis, using as an example the latent variable of depressive symptoms, with illustrations using the Beck Depression Inventory. We will show that with fitting data to the Rasch model, we can confirm the structural validity of the scale, including key attributes such as invariance, local dependency and unidimensionality. We also illustrate how the approach can inform on the meaning of the numbers attributed to scales, the amount of the latent traits that such numbers represent, and the consequent adequacy of statistical operations used to analyse them. We would argue that fitting data to the Rasch model has become the measurement standard for patient reported outcomes in general and, as a consequence will facilitate a quality improvement of outcome instruments in psychiatry. Recent advances in measurement technologies built upon the calibration of items derived from Rasch analysis in the form of computerized adaptive tests (CAT) open up further opportunities for reducing the burden of testing, and/or expanding the range of information that can be collected during a single session.

1. Introduction

The use of patient reported outcomes in health care in general, and psychiatry in particular, has seen a rapid expansion over recent years. The ascertainment of latent constructs such as anxiety, depression and self harm has seen a steady increase in the number of instruments designed to measure such attributes (Bowen et al., 2008; Brunner et al., 2007; Fliege et al., 2009; Gamez et al., 2007; Garlow et al., 2008; Honarmand and Feinstein, 2009; King et al., 2008; Klonsky et al., 2003; Latimer et al., 2009; Parker et al., 2005; Pedersen, 2006; Pomerleau et al., 2003; Terluin et al., 2006; Tuisku et al., 2009). While some instruments are administered by professionals, the majority are self completed 'patient reported outcomes' and are widely used in both clinical practice and research (Bech, 2008; Chan et al., 2010; Chandler et al., 2010; Counts et al., 2010; Hawton et al., 2002; Norris and Aroian, 2008; Steinhausen et al., 2009). The obvious value of such instruments is that they can minimize the burden of assessment upon patients, and can be applied to large numbers, which may be more restricted, or not feasible in the case of structured clinical interviews.

However, the use of such scales has been the subject of some debate. Marshall et al. (2000), examining a number of controlled trials in schizophrenia, found that the intervention was more likely to be effective when unpublished scales were used, in opposite to validated ones. Another issue, which has been rarely considered, is that the majority of instruments derive ordinal scores, which indicate rank relationships (Stevens, 1946). Such scores are not capable of supporting mathematical calculations such as change scores, or parametric effect sizes (Smith, 2001). Consequently using ordinal scores in sophisticated parametric analyses could lead to misinference of the findings (Merbitz et al., 1989). However, ordinal scales, which provide a magnitude of the trait under consideration, are perfectly acceptable when the object is to identify a cut point, or magnitude of the trait, such as found in many instruments, for example, to ascertain depression. This application just relies on a specific magnitude, which is available from an ordinal scale. Thus, the problem is not necessarily the scale themselves (although it may be), but rather the way in which they are analysed.

In the formation of patient reported outcomes, the usual procedure has been to generate a scale with a certain number of

* Corresponding author. Hospital de Clinicas de Porto Alegre, Serviço de Psiquiatria, Avenida Ramiro Barcelos, 2350 4° Andar, CEP 90035-003 Porto Alegre, RS, Brazil. Tel.: +5551 33598413.
E-mail address: neusa-rocha@via-rs.net (N.S. da Rocha).

0022-3956/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jpsychires.2012.09.014

items that intend to assess some observable behaviours related to the construct of interest (Tesio, 2003). Therefore, when setting out to measure such a construct we look for indicators (items) which are related to the construct, preferably in a way to be specified by an underlying theory. When someone responds to a certain question or item, the probability of the subject to endorse the item should depend on their level of the latent trait or ability (Baker, 2001). For example, it is expected that a more depressed subject will endorse an item regarding hopelessness more frequently than a non-depressed one. While this particular item does not directly measure depression (it addresses hopelessness), it helps in the construction of the depression score, together with other related items, which are designed to measure the latent variable (depression in this case).

In order to put together a set of items with the expectation that they measure the target construct, a set of psychometric requirements must be satisfied, and these requirements can be grouped into those associated with Classical Test Theory (CTT), and Modern Test Theory (MTT) (although in practice there is considerable overlap between the two). The present article aims to briefly review the former, and then go on to describe the potential contributions of the latter, in particular Rasch analysis, with respect to the development and testing of instruments. The Beck Depression Inventory (BDI) will be used as a practical example of this purpose.

2. Classical Test Theory

The measurement properties of most patient reported outcomes to-date have been evaluated from the CTT perspective. This has entailed publication of evidence concerning the reliability and the validity of the instrument. Reliability concerns whether or not the instrument has consistency, both internally (Cronbach's alpha) and over time (test–retest). Validity is often reported to comprise three central aspects, namely construct validity, criterion and content validity. These represent appropriate targeting, its relationship with a gold standard (e.g. a structured clinical interview), and whether the items appear to be consistent with expectations of an underlying theory (Nunnally, 1978). In practice, validity falls into two primary components, internal and external (Loevinger, 1957). The former concerns whether or not it is valid to add together the set of items and, within the framework of CTT, is primarily concerned with factorial validity. The latter is concerned with whether or not the instrument measures what is intended, and would include criterion validity. Reliability sits between these two, as in order to test reliability the summed score must be valid (i.e. internal validity). In order to test external validity, both the summed score and reliability must be shown to be adequate. Thus, the focus of CTT lies on the summed score, and its decomposition into true score and measurement error, the estimation of reliability, and the correlation between that summed score and other comparator measures, whether they are judged to be a gold standard, or not.

The Beck Depression Inventory – second edition (BDI-II) is one such example of a well-known instrument used to quantify depression (Beck et al., 1996) which has been developed using CTT. When a patient completes the BDI-II, a set of 21 items (scored 0–3) indicate the level of depression of this patient on a score which ranges from 0 to 63. A score of 29 and above is indicative of severe depression. A considerable body of evidence exists with regard the reliability and validity of this instrument (and the original version) (Beck et al., 1996; Hayden et al., 2010; Helm and Boward, 2003; Levin et al., 1988; Osma et al., 2004; Siegert et al., 2010). However, some concern has been expressed about the unidimensionality of the scale, and whether or not it is valid to add together all the items (Storch et al., 2004). Concerns have also been expressed (with regard the earlier version) about the reliability (test–retest) of the instrument (Ahava et al., 1998). While there is a myriad of adaptations of the BDI into different languages, and for different diagnoses, some have raised issues about the absence of relevant scales in certain diagnoses or with particular groups, such as older people with cancer (Nelson et al., 2010). Nevertheless, such group/diagnosis-specific reliability and validity is fundamental, and has been recognized as a requirement for some time (Loevinger, 1957). Scales should have evidence of reliability and validity in every group for which their use is intended.

Although there are a few exceptions, one interesting aspect of the CCT approach is that every item is given an equal weight with respect to their contribution to the summed score. For example, an item that assesses suicidal ideation is given the same weight (raw score) as one that assesses inattention. Nevertheless, it is known that clinically a depressive syndrome with suicidal ideation is more severe and that this item alone indicates higher intensity of depression (Alexandrino-Silva et al., 2009; Clark et al., 1983; Pompili et al., 2008; Selvi et al., 2010; Van Gastel et al., 1997). Yet surprisingly, there are circumstances when the simple raw score is a sufficient statistic for the estimate of the persons underlying level of the trait. This notion of 'sufficiency' has also been around for a long time (Fisher, 1921) and implies that the raw score contains all the information required to estimate the persons level of, in our example, depression. It is also equivalent to a stochastically consistent ordering of all item pairs (Fischer and Molenaar, 1995). To ascertain whether or not this is the case, we can invoke Modern Test Theory and, specifically, the Rasch measurement model.

3. Modern Test Theory (MTT) and the Rasch model

The first MTT models (under the generic label of Item Response Theory –IRT) appeared in the 1950s in the education area based on the need to build tests that would be at the same time simple, valid and with high discrimination power (Embretson and Reise, 2000). IRT represents a group of several distinct models, which share in common an assumption that the response to any particular item is a function of the difference between the ability of the person (or in our example their level of depression) and the characteristics of the item which, in the Rasch model, is the difficulty of the item (or in our case, the level of depression implied by the item). Other IRT models have additional characteristics of items, but lose the key characteristics of sufficiency in doing so.

The Rasch Model is a one-parameter IRT approach that has been increasingly utilized in the health field (Reise and Waller, 2009; Rocha et al., 2012; Tennant et al., 2004a,b). In this model, the parameter of discrimination is fixed in the value of 1 for all the items, and then only the parameter of *difficulty* varies. As a consequence, the Rasch model is frequently considered a model of 1 parameter (*difficulty*) (Baker, 2001; Rasch, 1960). The main strength of this model is that it allows for testing if the simple summed raw score is a sufficient statistic (which cannot be done with other models) and also tests whether or not the data are consistent with the axioms of conjoint measurement, so providing a transformation to interval scaling, which also cannot be done with other models (Karabatos, 2001; Michell, 2003). By fitting data to the Rasch model, we can assume that the estimated latent measure, when generated by an instrument that fits Rasch Measurement Model requirements, is interval scaled. As such, given appropriate distributional properties, this estimate may be suitable for parametric operations, including basic aspects such as the calculation of change scores, and group comparisons using a *t*-test, as well as more complex models (such as structure equation modelling), given other requirements are also met (O'Connor and Tennant, 2008).

IRT in general, including Rasch analysis, explores the performance of each individual item rather than the total test score as in

CTT. All explorations are based on the assumption that the probability of someone endorsing an item (in the Rasch dichotomous case) depends *only* on the difficulty of the item and on the subject's ability. This probabilistic relationship is tested by a series of fit statistics, which examine the comparison between the theoretical item performance (i.e., subjects with more ability should get right answers, and more difficult items should be correctly answered by those who have higher ability) and the observed data (Andrich, 1988). Results are reported as a series of chi-square statistics and fit residuals. All are concerned with the amount of discrepancy between expected and observed data for that particular item. For example, where an item fits the Rasch Model, a chi-square probability should exceed 0.05 (that is no significant deviation), and a fit residual should be within a specified range (e.g. ±2.5) (Pallant and Tennant, 2007).

4. An example using the BDI

To illustrate how data are fitted to the Rasch model, data were collected from a sample composed of 122 chronic patients, of whom 66 (54.1%) were male, and 56 (45.9%) were female. The most frequently reported health problems were hypertension (18%), heart diseases (15.6%), neoplasm (13.1%), diabetes (13.1%), emphysema/asthma/bronchitis (11.5%), autoimmune diseases (8.2%), and kidney diseases (8.2%). They were recruited in a tertiary hospital in Porto Alegre-RS-Brazil, in the different clinical and surgical inpatient units and outpatient clinics. The Ethics Research Committee of Hospital de Clínicas de Porto Alegre approved this investigation.

Table 1 shows the results of fitting the data from the BDI to the Rasch model. Thus, used the RUMM2020 software package (Andrich et al., 2004). With a Bonferroni correction to the Chi-Square item probability, all items are shown to fit the model, except the item 19 "Weight loss", which was excluded because of misfit (fit residual = 3.08; chi-square = 10.3; $P = 0.016$). Furthermore, all fit residuals are within the (99%) range of ±2.5. Note also the location of the items. These indicate the severity of depression associated with the item. Thus, a disturbed sleep pattern and loss of energy are associated even with low levels of depression, whereas suicidal thoughts would be affirmed by those only with very high levels of depression.

Table 1
Measures of fit and location (SE) of BDI items.

	BDI items[a]	Location	SE	FitResid	ChiSq	Prob
16	Sleep pattern	−1.186	0.18	1.226	1.742	0.628
15	Loss of energy	−1.163	0.134	−0.118	0.536	0.911
20	Excessive worrying about health	−1.124	0.192	0.035	2.893	0.408
17	Tiredness or fatigue	−1.077	0.193	−1.122	5.237	0.155
21	Loss of interest in sex	−0.85	0.173	−0.761	5.168	0.160
6	Punishment feelings	−0.816	0.25	0.516	0.385	0.943
14	Appearance	−0.72	0.179	1.06	4.751	0.191
8	Self-criticalness	−0.704	0.143	2.063	8.685	0.034
10	Crying	−0.589	0.187	0.519	2.888	0.409
4	Loss of pleasure	−0.522	0.191	−2.322	9.848	0.020
13	Indecisiveness	−0.101	0.203	−0.238	3.109	0.375
2	Pessimism	−0.03	0.215	−1.388	5.178	0.159
18	Changes in appetite	0.048	0.18	1.486	10.542	0.014
11	Irritability	0.098	0.215	0.471	7.954	0.047
5	Guilty feelings	0.1	0.226	−0.568	3.339	0.342
1	Sadness	0.18	0.212	−2.084	10.214	0.017
12	Loss of interest	0.979	0.257	−1.381	4.607	0.203
7	Self-dislike	1.797	0.226	−0.82	1.738	0.628
3	Past failure	2.748	0.277	−1.5	5.403	0.145
9	Suicidal thoughts or wishes	2.931	0.423	0.805	2.372	0.499

[a] Collapsing categories and excluding item 19 "Weight loss" because of misfit: fit residual >2.5.

Item fit can also be evaluated graphically by the Item Characteristic Curve (ICC), sometimes called the Item Response Function. It is based on the fact that individuals with more ability (latent trait) have more chance of succeeding the item. As we can observe the slope is a sigmoid and reminds us of an "S" (Baker, 2001). Fig. 1 shows the ICC of the item "Indecisiveness" of the BDI. Axis *X* indicates the latent depression estimate on an interval 'logit' scale and the *Y* axis represents the expected response value of the item. The sigmoid is the relationship expected by the model, and the dots on the line, represent the average response for groups at different ability (depression) levels (Andrich, 1978; Rasch, 1960). It can be seen that in this instance, the dots closely follow the expected curve, and so the item represents a good fit to the model expectations. Consequently the fit statistics, as reported at the top of the graph in the form of the fit residual, and the chi-square probability, each indicate a good fit to the model. These fit statistics are also reported in Table 1.

The BDI has a polytomous response structure, which is why the expected response ranges from 0 to 3 (the Rasch programmes always set the first category to 0). It is important to note that, in such circumstances, the distances between response options are not equal with respect to the underlying trait. For example, in the case of the BDI these distances vary considerably (Fig. 2) which is consistent with the partial credit parameterization of the Rasch model.

Also, sometimes the categories may not be ordered properly, and this may contribute to misfit. This is where the transition point between categories (threshold) does not follow an increasing level of the underlying trait. Such 'disordered thresholds' may arise because of ambiguity in response option wording, or by respondents having difficulty discriminating between options (perhaps when the category semantics are too close to one another). This can be accommodated within the Rasch model framework by collapsing categories until they are all ordered. Fig. 2 shows the item set after collapsing categories, where necessary. The way in which categories are collapsed is shown in Fig. 3 which illustrates an item with a disordered threshold together with its new Rasch-generated well-performing format. Thus, the Rasch model allows for testing potential alternative response formats, as well as for checking if these alterations improve the overall scale (Chachamovich et al., 2009; Chachamovich et al., 2008). In our case, originally only 6 out of the 21 BDI items showed ordered categories. Mostly, in 15 out of 21 BDI items, categories "1" and "2" which represents less severe symptoms, had to be collapsed. The resulting variability in scoring range across items is another reason to use the Partial Credit Model in the present analysis.

In addition, summary fit statistics indicate how well the scale, as a whole satisfies Rasch model expectations. Initially, in this sample,

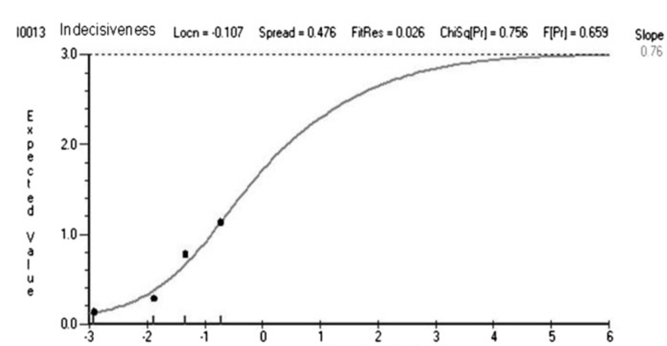

Fig. 1. Item characteristic curve of item 13 (Indecisiveness) of the Beck Depression Inventory (BDI).

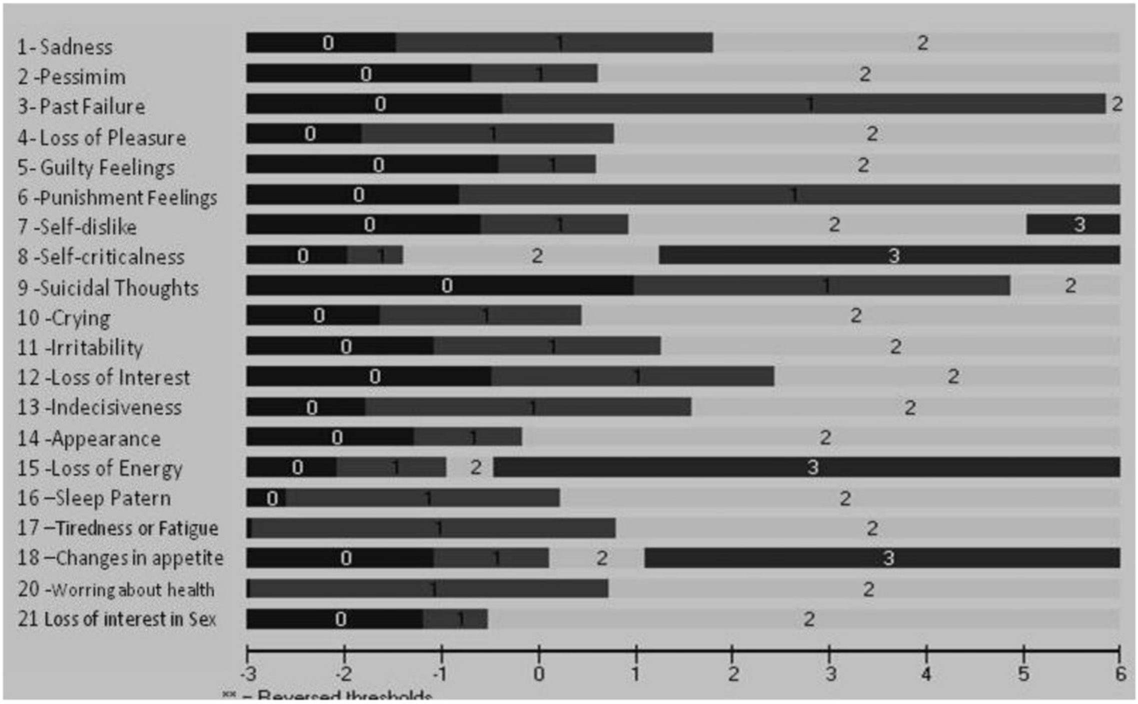

Fig. 2. Threshold map for BDI items.

before rescoring disordered thresholds, the BDI shows poor fit to model expectations, as indicated by a summary Chi-Square significance value that was very low, and a summary item residual standard deviation of 1.39 (Table 2). After rescoring, fit improved, but item 19 ‘Weight loss’ showed significant misfit to and so it was deleted. This improved the summary fit statistics and no individual item showed misfit to the model (Bonferroni corrected).

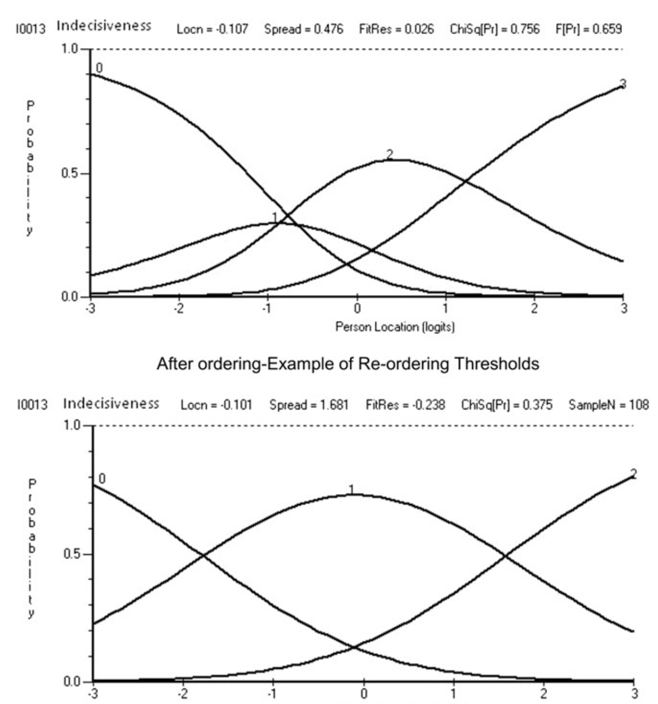

Fig. 3. Probability categories curves of the item 13 (Indecisiveness) of the BDI-example of disordered thresholds.

Where comparisons are desired, it is also required that an item is invariant across different subjects, such male/female, older/younger, ill/health, etc. As a consequence, a well-performing item should not show differential item functioning (DIF) (Tennant et al., 2004a,b). For example, when DIF is present, the probability of a subject endorsing an item (or category) when they have the same amount of depression differs according to group membership (e.g. gender). Thus, the estimation of depression level will be biased (e.g. by the gender of the subject) (McKenna et al., 2007). The process of Rasch Analysis also allows for a test for the presence of DIF, and provides information regarding item invariance and indicates which items require alterations or deletion in order to generate a DIF-free scale. In practice, the BDI items showed no DIF for gender, as is shown in Fig. 4 which plots the ICC for both males and females, as well as the model expectation.

Depending upon the application, the targeting of persons to items is an important feature of validity, and another additional feature of Rasch analysis. As the process plots both persons' ability and items' difficulty on one metric logit scale, a comparison (both visual and statistical) of the location of persons and items is possible, including the magnitude of difference between the mean person and item location, giving the overall targeting of the scale (Fig. 5).

Table 2
Summary of measures of Rasch model fit for BDI items.

Measures of fit	Basal model	Adjusted model[a]	Subtest model[b]
Item fit residual (SD)	−0.13 (1.39)	−0.21 (1.22)	−0.17 (1.27)
Person fit residual (SD)	−0.13 (0.90)	−0.25 (0.99)	−0.22 (0.92)
Total item × 2	122.15	96.59	90.74
Chi-square *P*	0.000012	0.002	0.002[c]
PSI	0.82	0.86	0.85
t-test P (IC 95%)	7.02% (6%–12%)	6.5% (2%–11%)	7.5% (3%–12%)

[a] Collapsing categories and excluding item 19 “Weight loss”.
[b] Subtest analysis:subtest1 items 5&7 5 ‘Guilty Feelings’ & 7 ‘Self-dislike’; subtest 2 13&16 13 ‘Indecisiveness’ & 16 ‘Changes in sleep pattern’; subtest3 6&20 6 ‘Punishment feelings’ & 20 ‘Excessive worrying about health’.
[c] Bonferroni adjusted Chi_Square 0.002.

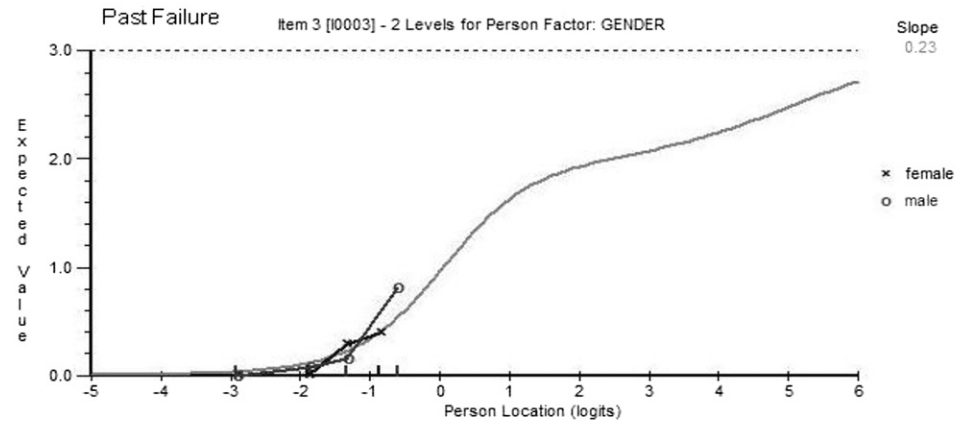

Fig. 4. DIF-free item from the BDI "Past failure".

In Fig. 5 the offset of persons (upper half) to items (lower half) may suggest a poorly targeted scale. Unless a substantial increase in depression is expected, this skewed distribution (with a floor effect) would not serve well as an outcome measure as considerable deterioration would be needed before any points were added to the scale score. The patients are a long way below the operational range of the scale (although from a probabilistic perspective they may nevertheless gain some points). Thus when a task is to measure the spectrum of depressive symptoms, new items around the low- and high extremes would improve the measurement range. By doing this, the amount on information around the clinical cut-point may decrease, but the knowledge about the whole spectrum will increase significantly (Bond and Fox, 2007). If, on the other hand, the scale is used as a screening instrument for depression among those not expected to have the condition (or a general population) then this distribution may be expected. What is a concern under these circumstances is that the clinical cut points are associated with the maximum degree of precision of the scale (usually at zero logits).

Dropping of one item and the collapsing of categories of 15 items have, as a consequence, changed the operational range of the scale, rendering the existing cut points invalid. However, within the framework of the Rasch analysis it is possible to equate tests. Thus, the calibrations of those items which were unchanged in the revised scale (that is not rescored) were used to anchor the revised metric to the original metric. In that way, the logit value of the original cut point (e.g. 29) can be used in the revised scale to determine what the equivalent raw score would be. Thus, the original cut points of 19 and 29 would become 15 and 26 on the revised scale (Fig. 6).

Other requirements of the Rasch model are unidimensionality and local independence of items. Although some items (5 'Guilty Feelings' & 7 'Self-dislike'; 6 'Punishment feelings' & 20 'Excessive worrying about health'; 13 'Indecisiveness' &16 'Changes in sleep pattern') were shown to be highly correlated in the residuals, suggesting potential redundancy (correlations >0.3 in residual correlation matrix), when these items were grouped (Subtest analysis) this procedure did not improve fit. Unidimensionality was confirmed by a post hoc *t*-test (% outside range 7.5%); binomial IC% (3–18%) (Table 2).

As such, by assessing these requirements, it is assured that, for example, the BDI measures exclusively depression (unidimesionality), and that the value attributed to each question (item) of the scale can be adequately added to the value of the other; that each item is measuring a relevant aspect and, given the level of depression of the person, does not depend on another item to have this information (local independence), and even if this item is administrated to other respondent belonging to a different group it will continue measuring the same ability (invariance) (Tesio, 2003).

5. Discussion: Rasch applications in clinical research

This paper is an introductory paper to stress the potentialities of Rasch analysis for Psychiatric practice and research. The BDI was

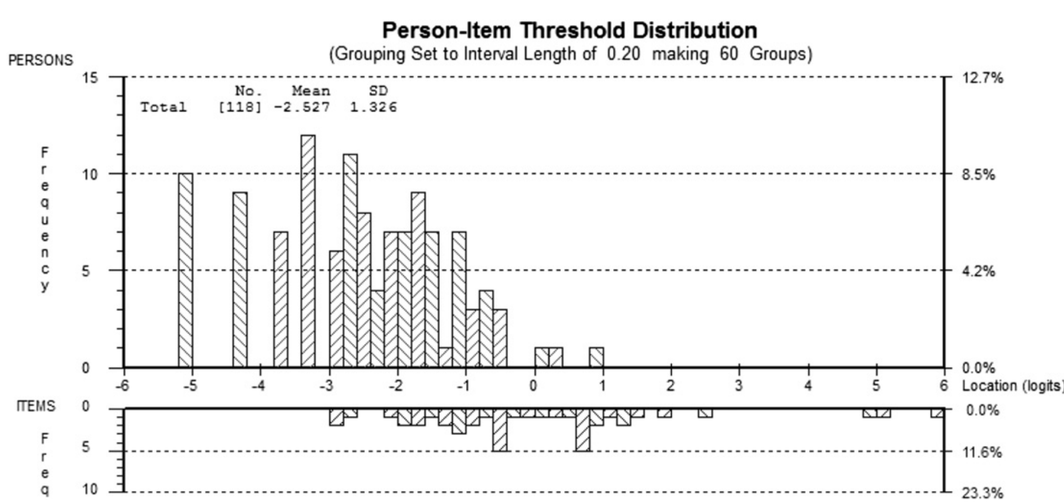

Fig. 5. Person–item distribution map.

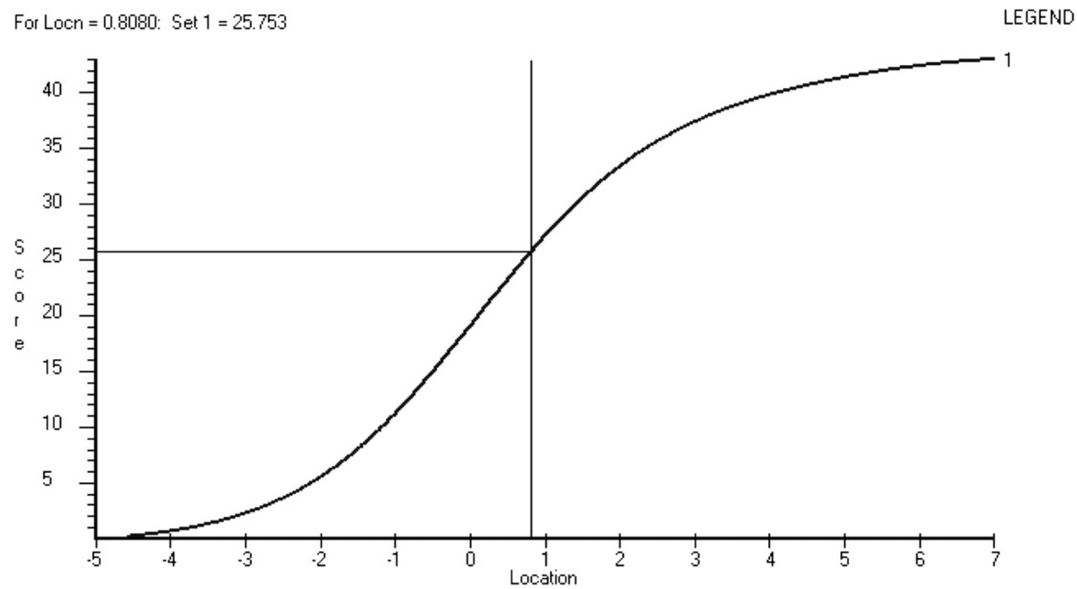

Fig. 6. Equating the original cut point (raw score 29) of logit value 0.808 to give the new raw score cut point for the revised scale.

used here merely as an example. The BDI has been shown to satisfy Rasch model expectations after some adjustments, in a mixed diagnostic sample of a tertiary hospital. Designed to be used in a clinical sample of depressed patients to ascertain the severity of that depression, the distribution of thresholds across the continuum of depression is consistent with that purpose. The removal of item 19 "weight loss" is consistent with the confounding of co-morbidity that may be expected when applied to other diagnostic groups, and this type of confounding has been found in other depression scales (Gibbons et al., 2011). In the present example, it reflects the fact that weight loss does not share a probabilistic structure with the other items in the scale. A purpose beyond our paper is to make definitive conclusions about the psychometric properties of BDI.

Rasch analysis represents the current quality standards in measuring outcomes (Sloan and Mandrekar, 2005; Tennant et al., 2004a,b). It complements Classical Test Theory by providing detailed analysis of how items work within scales, and whether or not their summed score is valid (Chachamovich et al., 2008). Ultimately, it examines scales and items in depth, and statistically tests the theoretical requirements. Where data are shown to fit the Rasch model, a transformation to interval scaling is available through exporting the latent estimate from the Rasch analysis programme. Consequently, Rasch analysis has been applied to several distinct areas and specialties, such as quality of life, pain, rheumatology, rehabilitation, neurology and ophthalmology (Hagell et al., 2003; Lamoureux et al., 2008; Pesudovs et al., 2010; Sloan and Mandrekar, 2005; Tennant and Conaghan, 2007; von Steinbüchel et al., 2010; Wolfe, 2003).

Increasingly, the development (Adler and Brodin, 2011; Cinnamon et al., 2011) and reviews of existing measures widely used in psychiatry are being published in appropriate journals (Castro-Costa et al., 2008; Kendel et al., 2010; Kørner et al., 2012; Licht et al., 2005; Pallant et al., 2006; Shea et al., 2009; Smith et al., 2006.)

There are several distinct advantages of applying the Rasch model to outcome scales in Psychiatry. Besides ensuring that the best quality standards for measurement are attained for any outcome scale, the process adds a layer of diagnostic information which is not available in CTT, and which may have clinical relevance. For example, just as much as items may misfit model expectations, so may persons. Consequently, persons whose responses differ from model expectations may indicate some unknown pathology or co-morbidity which affects those responses. Where different scales are used in the same diagnostic groups, clinical caseness may vary solely because different scales result in different prevalence, for example, of depression (Covic et al., 2009). Rasch analysis allows for direct comparison of scale cut points under a common person equating study (same people fill out different scales at same time), so adding to the knowledge of the true variability of depression, and other conditions, as opposed to the potentially spurious variability derived from different scale-specific case ascertainment. Furthermore, an interesting application of this method would be its use for the definition of more homogeneous syndromes (Bouman and Kok, 1987), since the heterogeneity of depression construct for several types of settings: clinical, psychiatric or general population samples.

When items from different scales are calibrated on the same metric, an 'item bank' is formed (Forkmann et al., 2009). This opens up the possibility of Computer Adaptive Testing (CAT). CAT makes use of the calibrated items to provide 'tailored testing' for the individual. Often starting at the item representing an average level of depression, response to that item will determine the next item to be administered, and so on. In this way, relatively few items need to be administered, so providing a useful way to screen patients in, for example, a busy out-patient clinic.

Rasch analysis can also help adjust for cross-cultural differences where data is pooled, for example, in international clinical trials (Tennant et al., 2004a,b). It has frequently been used to explore cross-cultural properties of several well-known instruments and, again, whether the format demands adaptations for certain cultural contexts (Rocha et al., 2012; Ravens-Sieberer et al., 2007; Tennant et al., 2004a,b).

In summary, the present article briefly reviews the Rasch Measurement Model, its practical applications and potential for psychiatry. It is now widely adopted in many specialties, and has the potential to provide high quality measurement for everyday practice, and for research.

Role of the funding source

This study was partially funded by FIPE-HCPA, CAPES and the University of Edinburgh.

Contributors

All authors managed the literature searches. Neusa Rocha and Alan Tennant undertook the statistical analysis, and Neusa Rocha, Eduardo Chachamovich and Marcelo Fleck wrote the first draft of the manuscript. All authors contributed to and have approved the final manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

None.

References

Adler M, Brodin U. An IRT validation of the Affective Self Rating Scale. Nordic Journal of Psychiatry 2011;65:396–402.
Ahava GW, Iannone C, Grebstein L, Schirling J. Is the Beck Depression Inventory reliable over time? an evaluation of multiple test–retest reliability in a nonclinical college student sample. Journal of Personality Assessment 1998; 70(2):222–31.
Alexandrino-Silva C, Pereira ML, Bustamante C, Ferraz AC, Baldassin S, Andrade AG, et al. Suicidal ideation among students enrolled in healthcare training programs: a cross-sectional study. Revista Brasileira de Psiquiatria 2009;31(4):338–44.
Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;43:561–73.
Andrich D. Rasch models for measurements. Beverly Hills: SAGE; 1988.
Andrich D, Sheridan B, Luo G. RUMM: a Windows program for analysing item response data according to Rasch Unidimensional Measurement Models. Perth, Western Australia: RUMM Laboratory; 2004.
Baker FB. The basics of item response theory. Madison: ERIC Clearinghouse on Assessment and Evaluation; 2001.
Bech P. Pichot – a tribute to the European psychopharmacologist on his 90th birthday. European Psychiatric Review 2008;1:76–80.
Beck AT, Steer RA, Brown GK. BDI-II manual. San Antonio: The Psychological Corporation; 1996.
Bond TG, Fox CM. Applying the Rasch model-fundamental measurement in the human sciences. Mahwah: Lawrence Erlbaum Associates; 2007.
Bouman TK, Kok AR. Homogeneity of Beck's Depression Inventory (BDI): applying Rasch analysis in conceptual exploration. Acta Psychiatrica Scandinavica 1987;76: 568–73.
Bowen A, Bowen R, Maslany G, Muhajarine N. Anxiety in a socially high-risk sample of pregnant women in Canada. Canadian Journal of Psychiatry 2008;53(7):435–40.
Brunner R, Parzer P, Haffner J, Steen R, Roos J, Klett M, et al. Prevalence and psychological correlates of occasional and repetitive deliberate self-harm in adolescents. Archives of Pediatrics & Adolescent Medicine 2007;161(7):641–9.
Castro-Costa E, Dewey M, Stewart R, Banerjee S, Huppert F, Mendonca-Lima C, et al. Ascertaining late-life depressive symptoms in Europe: an evaluation of the survey version of the EURO-D scale in 10 nations. The SHARE project. International Journal of Methods in Psychiatric Research 2008;17:12–29.
Cinnamon JS, Finch L, Miller S, Higgins J, Mayo N. Preliminary evidence for the development of a stroke specific geriatric depression scale. International Journal of Geriatric Psychiatry 2011;26:188–98.
Chachamovich E, Fleck MP, Power M. Literacy affected ability to adequately discriminate among categories in multipoint Likert Scales. Journal of Clinical Epidemiology 2009;62(1):37–46.
Chachamovich E, Fleck MP, Trentini CM, Laidlaw K, Power MJ. Development and validation of the Brazilian version of the Attitudes to Aging Questionnaire (AAQ): an example of merging classical psychometric theory and the Rasch measurement model. Health and Quality of Life Outcomes 2008;6:5.
Chan YF, Leung DY, Fong DY, Leung CM, Lee AM. Psychometric evaluation of the Hospital Anxiety and Depression Scale in a large community sample of adolescents in Hong Kong. Quality of Life Research 2010;19(6):865–73.
Chandler GM, Iosifescu DV, Pollack MH, Targum SD, Fava M. RESEARCH: validation of the Massachusetts general hospital Antidepressant Treatment History Questionnaire (ATRQ). CNS Neuroscience & Therapeutics 2010;16(5):322–5.
Clark DC, vonAmmon Cavanaugh S, Gibbons RD. The core symptoms of depression in medical and psychiatric patients. Journal of Nervous and Mental Disease 1983;171(12):705–13.
Counts JM, Buffington ES, Chang-Rios K, Rasmussen HN, Preacher KJ. The development and validation of the protective factors survey: a self-report measure of protective factors against child maltreatment. Child Abuse & Neglect 2010; 34(10):762–72.
Covic T, Pallant JF, Tennant A, Cox S, Emery P, Conaghan PG. Variability in depression prevalence in early rheumatoid arthritis: a comparison of the CES-D and HAD-D Scales. BMC Musculoskeletal Disorders 2009;10:18.
Embretson SE, Reise SP. Item response theory for psychologists. Mahwan: Lawrence Erlbaum Associates, Inc.; 2000.
Fischer G, Molenaar IW. Rasch models. Foundations, recent developments and applications. New York: Springer-Verlag; 1995.
Fisher RA. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society 1921:222.
Fliege H, Lee JR, Grimm A, Fydrich T, Klapp BF. Axis I comorbidity and psychopathologic correlates of autodestructive syndromes. Comprehensive Psychiatry 2009;50(4):327–34.
Forkmann T, Boecker M, Norra C, Eberle N, Kirche T, Schauerte P, et al. Development of an item bank for the assessment of depression in persons with mental illnesses and physical diseases using Rasch analysis. Rehabilitation Psychology 2009;54(2):186–97.
Gamez W, Watson D, Doebbeling BN. Abnormal personality and the mood and anxiety disorders: implications for structural models of anxiety and depression. Journal of Anxiety Disorders 2007;21(4):526–39.
Garlow SJ, Rosenberg J, Moore JD, Haas AP, Koestner B, Hendin H, et al. Depression, desperation, and suicidal ideation in college students: results from the American Foundation for Suicide Prevention College Screening Project at Emory University. Depression and Anxiety 2008;25(6):482–8.
Gibbons CJ, Mills RJ, Thornton EW, Ealing J, Mitchell JD, Shaw PJ, et al. Rasch analysis of the hospital anxiety and depression scale (HADS) for use in motor neurone disease. Health and Quality of Life Outcomes 2011;9:82.
Hagell P, Whalley D, McKenna SP, Lindvall O. Health status measurement in Parkinson's disease: validity of the PDQ-39 and Nottingham Health Profile. Movement Disorders 2003;18:773–83.
Hawton K, Rodham K, Evans E, Weatherall R. Deliberate self harm in adolescents: self report survey in schools in England. British Medical Journal 2002; 325(7374):1207–11.
Hayden MJ, Dixon JB, Dixon ME, O'Brien PE. Confirmatory factor analysis of the Beck Depression Inventory in obese individuals seeking surgery. Obesity Surgery 2010; 20(4):432–9.
Helm Jr HW, Boward MD. Factor structure of the Beck Depression Inventory in a university sample. Psychological Reports 2003;92(1):53–61.
Honarmand K, Feinstein A. Validation of the Hospital Anxiety and Depression Scale for use with multiple sclerosis patients. Multiple Sclerosis 2009;15(12):1518–24.
Karabatos G. The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. Journal of Applied Measurement 2001;2: 389–423.
Kendel F, Wirtz M, Dunkel A, Lehmkuhl E, Hetzer R, Regitz-Zagrosek V. Screening for depression: Rasch analysis of the dimensional structure of the PHQ-9 and the HADS-D. Journal of Affective Disorders 2010;122:241–6.
King M, Semlyen J, Tai SS, Killaspy H, Osborn D, Popelyuk D, et al. A systematic review of mental disorder, suicide, and deliberate self harm in lesbian, gay and bisexual people. BMC Psychiatry 2008;8:70.
Klonsky ED, Oltmanns TF, Turkheimer E. Deliberate self-harm in a nonclinical population: prevalence and psychological correlates. American Journal of Psychiatry 2003;160(8):1501–8.
Kørner A, Brogaard A, Wissum I, Petersen U. The Danish version of the Baylor Profound Mental State Examination. Nordic Journal of Psychiatry 2012;66(3): 198–202.
Lamoureux EL, Pallant JF, Pesudovs K, Tennant A, Rees G, O'Connor PM, et al. Assessing participation in daily living and the effectiveness of rehabilitation in age related macular degeneration patients using the impact of vision impairment scale. Ophthalmic Epidemiology 2008;15(2):105–13.
Latimer S, Covic T, Cumming SR, Tennant A. Psychometric analysis of the self-harm inventory using Rasch modelling. BMC Psychiatry 2009;9:53.
Levin BE, Llabre MM, Weiner WJ. Parkinson's disease and depression: psychometric properties of the Beck Depression Inventory. Journal of Neurology Neurosurgery and Psychiatry 1988;51(11):1401–4.
Licht RW, Qvitzau S, Allerup P, Bech P. Validation of the Bech-Rafaelsen Melancholia Scale and the Hamilton Depression Scale in patients with major depression; is the total score a valid measure of illness severity? Acta Psychiatrica Scandinavica 2005;111:144–9.
Loevinger J. Objective tests as instruments of psychological theory. Psychological Reports 1957;3:635–94.
Marshal M, Lockwood A, Bradley C, Adams C, Joy C, Fenton M. Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. British Journal of Psychiatry 2000;176:249–52.
McKenna SP, Doward LC, Meads DM, Tennant A, Lawton G, Grueger J. Quality of life in infants and children with atopic dermatitis: addressing issues of differential item functioning across countries in multinational clinical trials. Health and Quality of Life Outcomes 2007;5:45.
Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Archives of Physical Medicine and Rehabilitation 1989;70(4):308–12.
Michell J. Measurement: a beginner's guide. Journal of Applied Measurement 2003; 4(4):298–308.
Nelson CJ, Cho C, Berk AR, Holland J, Rot AJ. Are gold standard depression measures appropriate for use in geriatric cancer patients? A systematic evaluation of self-report depression instruments used with geriatric, cancer, and geriatric cancer samples. Journal of Clinical Oncology 2010;28(2):348–56.
Norris AE, Aroian KJ. Assessing reliability and validity of the arabic language version of the Post-traumatic Diagnostic Scale (PDS) symptom items. Psychiatry Research 2008;160(3):327–34.
Nunnally JC. Psychometric theory. New York: McGraw-Hill; 1978.

O'Connor RJ, Tennant A. Measuring pain: issues of interpretation. The Lancet 2008; 372(9637):443.

Osma A, Kopper BA, Barrios F, Gutierrez PM, Bagge CL. Reliability and validity of the Beck Depression Inventory—II with adolescent psychiatric inpatients. Psychological Assessment 2004;16(2):120–32.

Pallant JF, Miller RL, Tennant A. Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis. BMC Psychiatry 2006;6:28.

Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology 2007;46(Pt 1):1–18.

Parker G, Malhi G, Mitchell P, Kotze B, Wilhelm K, Parker K. Self-harming in depressed patients: pattern analysis. Australian and New Zealand Journal of Psychiatry 2005;39(10):899–906.

Pedersen AG. Citalopram and suicidality in adult major depression and anxiety disorders. Nordic Journal of Psychiatry 2006;60(5):392–9.

Pesudovs K, Gothwal VK, Wright T, Lamoureux EL. Remediating serious flaws in the National Eye Institute Visual Function Questionnaire. Journal of Cataract & Refractive Surgery 2010;36:718–32.

Pomerleau OF, Fagerstrom KO, Marks JL, Tate JC, Pomerleau CS. Development and validation of a self-rating scale for positive- and negative-reinforcement smoking: the Michigan Nicotine Reinforcement Questionnaire. Nicotine and Tobacco Research 2003;5(5):711–8.

Pompili M, Rihmer Z, Akiskal HS, Innamorati M, Iliceto P, Akiskal, et al. Temperament and personality dimensions in suicidal and nonsuicidal psychiatric inpatients. Psychopathology 2008;41(5):313–21.

Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research; 1960.

Ravens-Sieberer U, Schmid S, Gosch A, Erhart M, Petersen C, Bullinger M. Measuring subjective health in children and adolescents: results of the European KIDSCREEN/DISABKIDS Project. Psychosocial Medicine 2007;4(Doc08):1–13.

Reise SP, Waller NG. Item response theory and clinical measurement. Annual Review of Clinical Psychology 2009;5:27–48.

Rocha NS, Power MJ, Bushnell DM, Fleck MP. Cross-cultural evaluation of the WHOQOL-BREF domains in primary care depressed patients using Rasch analysis. Medical Decision Making 2012;32(1):41–55.

Selvi Y, Aydin A, Boysan M, Atli A, Agargun MY, Besiroglu L. Associations between chronotype, sleep quality, suicidality, and depressive symptoms in patients with major depression and healthy controls. Chronobiology International 2010; 27(9–10):1813–28.

Shea TL, Tennant A, Pallant JF. Rasch model analysis of the Depression, Anxiety and Stress Scales (DASS). BMC Psychiatry 2009;9:21.

Siegert RJ, Tennant A, Turner-Stokes L. Rasch analysis of the Beck Depression Inventory-II in a neurological rehabilitation sample. Disability and Rehabilitation 2010;32(1):8–17.

Sloan J, Mandrekar S. Item response theory: when is it useful and where does classical test theory fit in? Quality of Life Research 2005;14(9):1982.

Smith AB, Wright EP, Rush R, Stark DP, Velikova G, Selby PJ. Rasch analysis of the dimensional structure of the Hospital Anxiety and Depression Scale. Psychooncology 2006;15:817–27.

Smith Jr EV. Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. Journal of Applied Measurement 2001;2(3):281–311.

Steinhausen HC, Gundelfinger R, Winkler Metzke C. Prevalence of self-reported seasonal affective disorders and the validity of the seasonal pattern assessment questionnaire in young adults findings from a Swiss community study. Journal of Affective Disorders 2009;115(3):347–54.

Stevens SS. On theory of scales of measurement. Science 1946;103(2684):677–80.

Storch EA, Roberti JW, Roth DA. Factor structure, concurrent validity, and internal consistency of the Beck Depression Inventory-Second Edition in a sample of college students. Depression and Anxiety 2004;19(3):187–9.

Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? when should it be applied, and what should one look for in a Rasch paper? Arthritis and Rheumatism 2007;57(8):1358–62.

Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value in Health 2004a;7(Suppl. 1): S22–6.

Tennant A, Pent M, Tesio L, Grimby G, Thonnard JL, Slade A, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. Medical Care 2004b;42(1 Suppl.): I37–48.

Terluin B, van Marwijk HW, Ader HJ, de Vet HC, Penninx BW, Hermens ML, et al. The Four-Dimensional Symptom Questionnaire (4DSQ): a validation study of a multidimensional self-report questionnaire to assess distress, depression, anxiety and somatization. BMC Psychiatry 2006;6:34.

Tesio L. Measuring behaviors and perceptions: Rasch analysis as a tool for rehabilitation research. Journal of Rehabilitation Medicine 2003;35:105–15.

Tuisku V, Pelkonen M, Kiviruusu O, Karlsson L, Ruuttu T, Marttunen M. Factors associated with deliberate self-harm behaviour among depressed adolescent outpatients. Journal of Adolescence 2009;32(5):1125–36.

Van Gastel A, Schotte C, Maes M. The prediction of suicidal intent in depressed patients. Acta Psychiatrica Scandinavica 1997;96(4):254–9.

von Steinbüchel N, Wilson L, Gibbons H, Hawthorne G, Höfer S, Schmidt S, et al. Quality of Life after Brain Injury (QOLIBRI): scale development and metric properties. Journal of Neurotrauma 2010;27:1167–85.

Wolfe F. Pain extent and diagnosis: development and validation of the regional pain scale in 12,799 patients with rheumatic disease. Journal of Rheumatology 2003; 30:369–78.