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*

ARTICLE INFO

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

ABSTRACT

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