Personality and Individual Differences 54 (2013) 530–535

Contents lists available at SciVerse ScienceDirect

Personality and Individual Differences

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

Rasch analysis of a simplified Beck Depression Inventory

Sebastian Sauer [a,*], Matthias Ziegler [b], Manfred Schmitt [c]

[a] *Generation Research Program, Human Science Center, Ludwig-Maximilians-Universität, München, Germany*
[b] *Department of Psychology, Humboldt Universität, Berlin, Germany*
[c] *Department of Psychology, University of Koblenz-Landau, Germany*

ARTICLE INFO

Article history:
Received 5 July 2012
Received in revised form 21 October 2012
Accepted 29 October 2012
Available online 30 November 2012

Keywords:
Depression
Rasch
Item response theory
Beck's Depression Inventory
BDI

ABSTRACT

Depression is one of the most clinically relevant mood disorders, and many assessment instruments have been developed to measure it. Probably the most frequently used instrument is Beck's Depression Inventory (BDI). The simplified BDI (BDI-S) is a more efficient version of the BDI that has been shown to be no less reliable or valid. As the BDI-S has not yet been subjected to rigorous tests of Item Response Theory, it is the aim of the present paper to conduct such an analysis using the Rasch model. This study subjected a simplified version of the BDI consisting of 20 items (BDI-S20) to a Rasch analysis in a sample of $N = 5{,}035$ participants. The scale, minus one misfitting item (BDI-S19), yielded a good approximation to Rasch assumptions. Moderate differential item functioning (DIF) was present. It is concluded that the BDI-S19 is an internally valid instrument for assessing depression, although some room for improvement exists.

1. Introduction

The Beck Depression Inventory (BDI) is one of the most commonly used measures of depressive symptoms in normal and clinical samples. A large number of studies have attested that the BDI has high levels of reliability and validity (Beck, Steer, & Carbin, 1988; Richter, Werner, Heerlein, Kraus, & Sauer, 2000). All versions of the BDI have been translated into many languages and have been used around the world. However, several authors have criticized the BDI for being lengthy, resulting in reduced clinical efficiency (Schmitt & Maes, 2000b; Valenstein, Vijan, Zeber, Boehm, & Buttar, 2001; Zimmerman, Chelminski, McGlinchey, & Posternak, 2008). This limitation is a result of the BDI's makeup. The BDI includes 21 depressive symptoms, each of which is represented by four items with increasing severity. Thus, a total of more than 80 statements have to be processed. This can pose an excessive demand for severely depressed patients. Two strategies have been employed for increasing the efficiency of the BDI. First, some authors have proposed maintaining the format of the BDI but reducing the number of symptoms. This strategy was first put forward by Beck (1978) and employed, for instance, by Steer, Cavalieri, Leonard, and Beck (1999).

A second strategy was advanced by Schmitt and Maes (2000a). These authors suggested replacing the four items that represent a symptom by a single item and applying a 6-point frequency scale. They applied this strategy to the German version of the BDI-IA and coined the resulting version the BDI-S with the "S" standing for "simplified." In addition to reducing the number of statements per symptom from 4 to 1, Schmitt and Maes (2000a) dropped the weight-loss item because this symptom has tended to have the lowest item-total correlation in studies using the German BDI-IA (e.g., Hautzinger, Bailer, Worall, & Keller, 1994). Consequently, the BDI-S consists of 20 items in the German language, and its sum score ranges from 0 to 100. Schmitt and Maes (2000a) applied their simplification strategy to the BDI-IA instead of to the BDI-II because the BDI-II (Hautzinger, Keller, & Kühner, 2006) was not yet available at the time they put forward and tested their strategy. The wording of the items (translated ad hoc from German into English) is given in Appendix A. The BDI-S was submitted to investigations of its reliability and validity (Schmitt & Altstötter-Gleich, 2010; Schmitt, Altstötter-Gleich, Hinz, Maes, & Brähler, 2006; Schmitt et al., 2003; Schmitt, Hübner, & Maes, 2010). The BDI-S has been shown to have good psychometric properties according to the standards of Classical Test Theory and has been shown to have good generalizability according to Latent-State-Trait-Theory (Steyer, Schmitt, & Eid, 1999). However, a more rigorous psychometric analysis of the BDI-S according to Item Response Theory (IRT; Molenaar & Fischer, 1995) is needed, even though IRT analyses have been performed for other versions of the BDI (e.g., Castro, Trentini, & Riboldi, 2010; Nuevo et al., 2009; Siegert, Tennant, & Turner-Stokes, 2010). The goal of the present article is to provide such an analysis for the BDI-S based on the Rasch model (Rasch, 1960).

* Corresponding author. Address: Prof.-Max-Lange-Platz 11, 83646 Bad Tölz, Germany. Tel.: +49 8041 79929 15; fax: +49 8041 7992911.

E-mail address: sauer@grp.hwz.uni-muenchen.de (S. Sauer).

0191-8869/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.paid.2012.10.025

1.1. IRT and Rasch modeling

One straightforward way to test the psychometric quality of a given item is to test whether a higher trait level is associated with a higher probability that a person will endorse the item (if the item is coded positively). As straightforward as it seems, this test is not explicitly undertaken in classical test theory (CTT), but is the central idea of IRT; hence, the name IRT. Thus, in short, IRT modeling can be understood as a test of whether item responses adhere sufficiently to a certain item response function (see Figure S1). Rasch models are a subset of IRT models (Rasch, 1960). They possess a number of desirable statistical properties. In particular, if the data fit the Rasch model, then (and only then) can the sum score be taken as a reasonable estimator of a person's trait. Thus, one central benefit of Rasch modeling is to determine whether the use of the sum score is justified. An important property of Rasch models is that the item curves do not intersect, which underscores the idea that a higher item endorsement is associated with a higher trait standing. Whereas the initial Rasch model was suitable for dichotomous items, alternative models for items with more answer options have been proposed. Two widely employed models are Masters' Partial Credit Model (PCM; Masters, 1982) and Andrich's Rating Scale Model (RSM; Andrich, 1978). Details are provided in the Supplementary text.

2. Method

2.1. Participants

To create a representative sample for the present analysis, we combined samples from studies that had previous used the BDI-S. In all subsamples, paper–pencil sampling methods were employed, except for Subsample 2 (interview sampling) and Subsample 5 (online sampling). Subsample 1 (Schmitt & Maes, 2000a) includes $n = 2285$ participants from more than 100 German regions. Subsample 2 (Schmitt et al., 2006) is a representative sample of $n = 2066$. Subsample 3 (Schmitt et al., 2010) is a demographically heterogeneous convenience sample of $n = 232$ German citizens. Subsample 4 (Schmitt, Baumert, Gollwitzer, & Maes, 2010) is a demographically heterogeneous convenience sample of $n = 248$ German citizens. Subsample 5 ($n = 229$) was previously recruited to conduct a study on the association between health and mindfulness (Kohls, Sauer, & Walach, 2009). The total sample size was 5035 with less than 1% missing values. In total, 54% of the participants were male, and 46% were female. The mean age was 46.6 ± 17.2 years (median = 46 years). Ethical treatment of the participants including informed consent in the studies from which we obtained the data for the present secondary analyses was approved by the ethics committee of the respective institution.

2.2. Invariance screening

Invariance is a crucial assumption of measurement. In short, person invariance entails that person parameters be identical (i.e., invariant) across item subsets of a scale (within reasonable error intervals). In the same vein, item invariance necessitates that item difficulties be invariant across subsamples. If the latter property does not hold, it is common to speak of *differential item functioning* (DIF). The idea that underlies person invariance is that if a test is homogenous (i.e., if all items measure the same variable), then person parameters should not differ when different subsets of items from that test are used to estimate them. A straightforward and easy test for person invariance is to split the test into halves and to compare the Rasch person parameters that are estimated from these halves (Bond & Fox, 2007). For testing item invariance, the same logic holds: The sample is split into parts (usually halves), and then the Rasch item parameters are calculated for each of the subsamples. If the sample is homogenous with regard to the variable being tested, then item parameters should not differ (Bond & Fox, 2007).

2.3. Analysis

We chose the RSM from the family of Rasch models as it is more parsimonious than the PCM and can be more appropriately applied to Likert-type rating scales as used in the BDI and the BDI-S (Andrich, 1978; Linacre, 2000). Within the framework of the RSM, items are specified by a location parameter and by category threshold parameters (Fischer & Molenaar, 1995). The location parameter characterizes the mean of the category parameters and can be referred to as the item's *difficulty*; a category threshold parameter quantifies the location on the latent variable at which two adjacent categories are equally probable. We used *Winsteps* 3.72 (Linacre, 2012) for the IRT analyses and *SPSS V18* for data preparation.

According to Cohen, Cohen, West, and Aiken (2002), the Expectation Maximization (EM) algorithm yields reliable estimates when the number of missing values is small. We used the RMV SPSS procedure for EM estimation with default settings to replace missing values as some of the computational methods relied on complete data matrices. In short, the EM algorithm is an iterative procedure that calculates regression weights based on a maximum likelihood computation for estimating missing values (Schafer & Olsen, 1998).

2.4. Rasch model fit

Substantial differences between expected and observed parameter values indicate poor model fit. The software program employed provides infit and outfit mean square (mnsq) coefficients, which are basically averaged squared standardized residuals, for both item and person parameters. Infit and outfit statistics reflect slightly different approaches to assessing the fit of an item or person: The outfit statistic is more sensitive to the influence of more extreme responses, whereas the infit statistic is weighted to reduce the influence of outliers. Both the infit and outfit statistics reflect the ratio of observed variance (i.e., variance attributable to the data) to expected variance (i.e., variance expected under the Rasch model). A mnsq value of 1 indicates ideal fit; mnsq values > 1.5, are indicative of poor fit; overfit (mnsq < 1) was considered to be nonproblematic (Bond & Fox, 2007). Due to the large sample size, we did not report Z statistics.

It is reasonable to assume for polytomous item formats that higher trait levels are associated with the selection of "higher" answer options (and vice versa)—given that the latent variable and the item under scrutiny point in the same direction. However, this hypothesis can and should be empirically tested. Rasch analysis provides category mnsq fit statistics for each answer category, thereby indicating whether the stipulated order of answer options is empirically confirmed. Similar to the item fit statistics, an expected value of 1.0 indicates perfect fit for mnsq statistics for categories. Additionally, category threshold parameter values should be ordered along their respective answer categories.

A central assumption of the Rasch model is unidimensionality, which holds when one single underlying trait is represented by the items. One way to test this assumption is by using Principal Component Analysis (PCA) of the standardized Rasch residuals; PCA is routinely employed for that purpose (Mavranezouli, Brazier, Young, & Barkham, 2010). The test rationale is explained in detail by Linacre (1998). Results were interpreted according to Linacre's (2012) recommendations as we assumed unidimensionality for practical purposes if no strong additional factor was present.