Journal of Affective Disorders 208 (2017) 191–197

Contents lists available at ScienceDirect

Journal of Affective Disorders

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

Research paper

The 52 symptoms of major depression: Lack of content overlap among seven common depression scales

Eiko I. Fried

University of Amsterdam, Department of Psychology, Nieuwe Achtergracht 129-B, room G0.28, 1001NK Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Content analysis
Major depression
Measurement
Scales
Symptom overlap

ABSTRACT

Background: Depression severity is assessed in numerous research disciplines, ranging from the social sciences to genetics, and used as a dependent variable, predictor, covariate, or to enroll participants. The routine practice is to assess depression severity with one particular depression scale, and draw conclusions about depression in general, relying on the assumption that scales are interchangeable measures of depression. The present paper investigates to which degree 7 common depression scales differ in their item content and generalizability.

Methods: A content analysis is carried out to determine symptom overlap among the 7 scales via the Jaccard index (0=no overlap, 1=full overlap). Per scale, rates of idiosyncratic symptoms, and rates of specific vs. compound symptoms, are computed.

Results: The 7 instruments encompass 52 disparate symptoms. Mean overlap among all scales is low (0.36), mean overlap of each scale with all others ranges from 0.27 to 0.40, overlap among individual scales from 0.26 to 0.61. Symptoms feature across a mean of 3 scales, 40% of the symptoms appear in only a single scale, 12% across all instruments. Scales differ regarding their rates of idiosyncratic symptoms (0–33%) and compound symptoms (22–90%).

Limitations: Future studies analyzing more and different scales will be required to obtain a better estimate of the number of depression symptoms; the present content analysis was carried out conservatively and likely underestimates heterogeneity across the 7 scales.

Conclusion: The substantial heterogeneity of the depressive syndrome and low overlap among scales may lead to research results idiosyncratic to particular scales used, posing a threat to the replicability and generalizability of depression research. Implications and future research opportunities are discussed.

1. Introduction

> *"The appearance of yet another rating scale for measuring symptoms of mental disorder may seem unnecessary, since there are so many already in existence and many of them have been extensively used."* (Hamilton, 1960).

Major Depressive Disorder (MDD) is among the most common mental disorders (Kessler et al., 2003), and studied in various disciplines ranging from the social sciences to genetics. Depression severity is studied so pervasively – to enroll study participants or track treatment efficacy, as a dependent variable, predictor, covariate, or moderator – that 3 rating scales are among the 100 most cited papers in science (van Noorden et al., 2014): the Hamilton Rating Scale for Depression (HRSD; rank 51) (Hamilton, 1960), the Beck Depression Inventory (BDI; rank 53) (Beck et al., 1961), and the Center of Epidemiological Scales (CES-D; rank 54) (Radloff, 1977).

Interestingly, a great variety of rating scales are used to assess depression severity; Santor et al. (2006) identified 280 different instruments developed in the last century, of which many are still in use. The routine practice is to conduct research based on *one* particular scale that is chosen for variable reasons: the scale may be available as a tool in the library of the University, it may be the gold standard in the particular subfield of depression research (such as the HRSD for antidepressant trials), or it may be the local custom of the department or hospital. The rationale for using specific scales – say, the HRSD instead of the CES-D or BDI – is rarely provided in scientific publications, and conclusions are drawn about depression in general, not about depression measured by a particular scale.

The tacit – and untested – assumption underlying this practice is that various depression instruments can be used as interchangeable measurements of depression severity. If this assumption does not hold, results of depression studies may be idiosyncratic to the particular scale used, posing a major challenge to the replicability and generalizability of depression research (Santor et al., 2006; Snaith, 1993). For

E-mail address: eiko.fried@gmail.com.

http://dx.doi.org/10.1016/j.jad.2016.10.019
Received 29 July 2016; Received in revised form 3 September 2016; Accepted 21 October 2016
Available online 21 October 2016

example, a large clinical trial may establish the efficacy of an antidepressant drug in a particular scale – which could have real implications for patients – although participants may show no clinical improvement on a range of other scales.

A number of reasons speak towards the possibility that rating scales are not interchangeable measures of depression severity. First, studies using multiple depression scales have identified differential scale performance. For instance, common instruments differ markedly in their classification of depressed patients into severity categories (Zimmerman et al., 2012). Second, psychometric analyses have documented that most scales are multidimensional, meaning they assess several constructs (Fried et al., 2016b); these factor structures, however, do not generalize across scales (Shafer, 2006; van Loo et al., 2012). Since scales measure different constructs, using different instruments may lead to different results; this is more likely to be problematic the more severe the heterogeneity of depression symptoms across different rating scales is. Finally, depression is a highly heterogeneous syndrome with many clinical presentations (e.g., Fried and Nesse, 2015a; Olbert et al., 2014) and numerous biological and neuroimaging correlates (e.g., Cassano and Fava, 2002), and individual depression symptoms such as sadness, insomnia, concentration problems or suicidal ideation differ in important properties such as biological markers, risk factors, and impact on impairment of functioning (for a review, see Fried and Nesse, 2015b). Symptoms also seem to respond differentially to antidepressant treatment (Hieronymus et al., 2016, 2015). Overall, this implies that rating scales may only be interchangeable indicators of depression severity inasmuch as their item content overlaps.

If overlap of symptom content among scales is high, interchangeable use of depression instruments may not pose a severe challenge. If overlap is low, however, the routine practice of using one particular scale in depression research may lead to idiosyncratic results and threaten the validity of a very large and important field of research. Given the pronounced heterogeneity of the depressive syndrome that may well be reflected in clinical instruments, the concern that depression instruments vary widely in symptom content is not far-fetched.

The main goal of the present report is thus to quantify the overlap of items among widely used depression rating scales.

2. Methods

2.1. Depression rating scales

To estimate the extent to which common rating scales of depression differ in terms of item content, 7 common rating scales for depression were examined: the 21-item BDI-II (Beck et al., 1996; from here on referred to as BDI), the 17-item HRSD, the 20-item CES-D, the 30-item Inventory of Depressive Symptoms (IDS) (Rush et al., 1996), the 16-item Quick Inventory of Depressive Symptoms (QIDS) (Rush et al., 2003), the 10-item Montgomery-Åsberg Depression Rating Scale (MADRS) (Montgomery and Asberg, 1979), and the 20-item Zung Self-Rating Depression Scale (SDS) (Zung, 1965). IDS and QIDS symptoms were collapsed consistent with their respective manuals, resulting in 28 IDS and 9 QIDS symptoms. For instance, the QIDS has 4 different questions on sleep problems, but only the highest one is used to score the domain 'sleep problems'. Of note, the nine QIDS items correspond to the nine DSM-5 (APA, 2013) MDD criterion symptoms.

The 7 scales were selected based on their frequency in the literature, inclusion in recent reviews, appearance in studies comparing multiple scales, and citation count (Gullion and Rush, 1998; Santor et al., 2006; Shafer, 2006; Snaith, 1993; van Noorden et al., 2014). The limitations section entails a discussion on whether analyzing different scales, or following a different procedure than the one described below to compare overlap, may have impacted on the results.

2.2. Content analysis

All scales together encompass 125 items. A content analysis was carried out to determine content overlap among scales. First, similarly worded items were combined *within* questionnaires to avoid biasing further analyses: 'apparent sadness' and 'reported sadness' that are both featured in the MADRS were collapsed into one item, as well as 'sad', 'depressed', and 'blue' in the CES-D. This reduces the number of MADRS items from 10 to 9, the number of CES-D items from 20 to 18 items, and the overall number of items to 122 that were used in subsequent analyses.

The primary objective of the present study was to determine the degree to which scales feature similar content. Therefore, in a second step, each potential item pair *across* scales was examined to determine symptom overlap (i.e. does any item in any scale overlap with any item of any other scale, for all possible combinations). It is impossible to carry out these comparisons objectively because there is no way to clearly determine whether two similarly worded symptoms are meant to measure the same problem or not. I therefore used a highly conservative approach and only differentiated between symptoms if they clearly differ from each other. Items were considered as equal (i.e. as the same item content across scales) as long as they were (a) roughly similarly worded, such as 'feeling sad' (IDS), 'feeling depressed' (HRSD), and 'feeling blue' (SDS), or (b) roughly oppositely worded, such as 'pessimism' (IDS, BDI, MADRS) and 'being hopeful about the future' (SDS, CES-D). Note that this is likely overly conservative, considering plenty of research showing that positive and negative emotions (such as being pessimistic and being hopeful) are only moderately negatively correlated and often form different dimensions. A less conservative approach would have considered all these to be different symptoms, and yielded a much higher number of total symptoms across all scales. Nonetheless, expecting a very large number of distinct depression symptoms, I would much rather err on the side of caution in this analysis.

Third, contrasting prior investigations of symptoms and scale overlap (Santor et al., 2006; Snaith, 1993), I differentiated between specific symptoms such as 'hypersomnia' and different types of 'insomnia', or between 'weight gain' and 'weight loss'. This is important because recent work has shown that these specific symptoms differ regarding important properties and should not be combined into compound items (Fried and Nesse, 2015b). To remain conservative in estimating when items are disparate from each other, however, specific (e.g., 'weight loss' in the HRSD) and compound (e.g., 'weight change' in the IDS) symptoms were considered to be overlapping, seeing that one is sufficient for fulfilling the other. A less conservative approach – not considering specific and compound symptoms as overlapping – would have increased the heterogeneity of depression and idiosyncrasy of scales markedly.

The content analysis described above resulted in a number of distinct symptoms, and information on whether these symptoms were (a) not featured in a scale, (b) featured as a part of a compound symptom, or (c) as a specific symptom. The results of the content analysis are attached in the form of a large table in the Supplementary Materials.

2.3. Statistical analyses

Content overlap was estimated using the Jaccard Index, a commonly used similarity coefficient for binary data that ranges from 0 (no overlap among scales) to 1 (complete overlap). The Jaccard Index or Jaccard similarity coefficient is calculated by $s/(u1 + u2 + s)$, where s is the number of items two questionnaires share, and $u1$ and $u2$ the number of items that are unique to each of the two scales. In the absence of a well-cited guideline on what a weak or strong Jaccard similarity coefficient is, I will use the rule from Evans (1996) for the correlation coefficient: very weak 0.00–0.19, weak 0.20–0.39, moder-