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
0165-0327/ © 2016 Elsevier B.V. All rights reserved.

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-

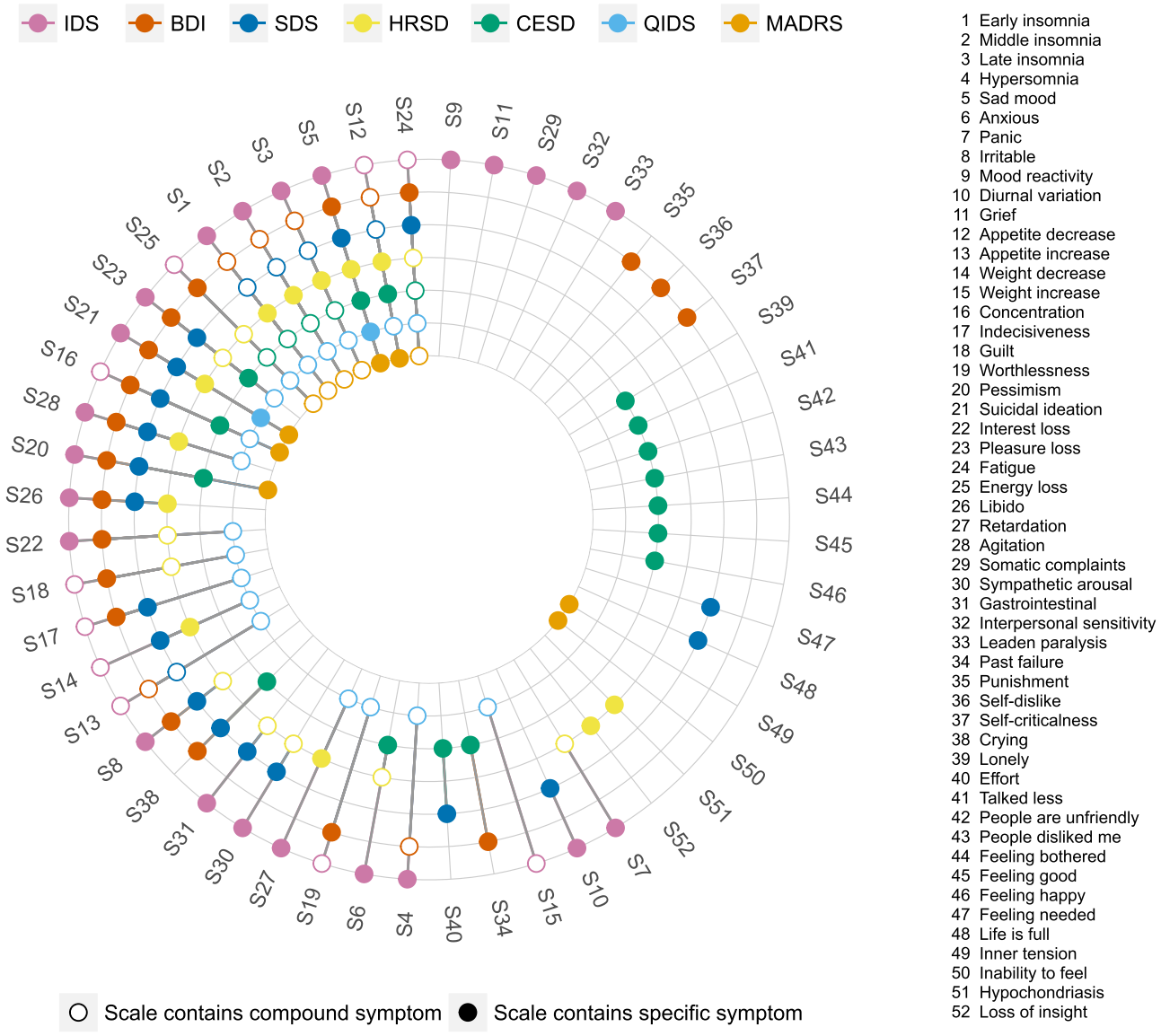

Fig. 1. Co-occurrence of 52 depression symptoms across 7 depression rating scales. Colored circles for a symptom indicate that a scale directly assesses that symptom, while empty circles indicate that a scale only measures a symptom indirectly. For instance, the IDS assesses item 4 hypersomnia directly; the BDI measures item 4 indirectly via a general question on sleep problems; and the SDS does not capture item 4 at all. Note that the 9 QIDS items analyzed correspond exactly to the DSM-5 criterion symptoms for MDD. Please see the online version for colors; in the black and white version, the circles respresent (from outer to inner circle): IDS, BDI, SDS, HRSD, CESD, QIDS, and MADRS.

ate 0.40–0.59, strong 0.60–0.79, and very strong 0.80–1.0.

In addition the Jaccard Index, I calculated the rate of specific (e.g., 'weight loss') vs. compound (e.g., 'weight change') symptoms per scale, and the rate of idiosyncratic symptoms per scale (i.e. symptoms that appear in no other scale). Analyses were conducted using *R*; data and code are available in the Supplementary Materials.

3. Results

The content analysis of 125 items across 7 scales resulted in 52 disparate depression symptoms (Fig. 1).

Symptoms appear in a mean of 3 of the 7 rating scales (mode=1, median=2.5). Of the 52 symptoms, 21 (40%) appear only in one single instrument, whereas 6 (12%) feature across all instruments: *sad mood*, *appetite decrease*, *fatigue*, and the 3 insomnia items *early*, *middle*, and *late insomnia* (cannot fall asleep, wakes up during the night, wakes up in the very early morning). Of these, *sad mood* is the only item captured specifically by all scales, while the others are queried with a mix of specific and unspecific questions. For instance, the HRSD and IDS include the 3 individual insomnia items, while other scales either query participants for insomnia in general (CES-D, MADRS, SDS), or – even broader – sleep disturbances (QIDS, BDI) that include both insomnia and hypersomnia.

The commonest specific symptoms captured by all scales are *sad mood* (featured in 7 scales), *suicidal ideation* (6 scales), and *pessimism* (5 scales). The DSM-5 MDD core symptom anhedonia was disaggregated into *loss of interest* and *loss of pleasure*; these two symptoms are present in 4 (IDS, BDI, HRSD, QIDS) and 6 scales (all but MADRS), respectively.

Table 1 lists in how many scales each of the symptoms are listed; for instance, 7 of the 52 symptoms (13%) appear across a subset of 3 scales.

3.1. Scale properties and performance

Table 2 summarizes to what degree the symptoms in each scale are idiosyncratic (i.e. do not appear in other scales) and specific (i.e. capture an item such as *weight loss* instead of a compound like *weight*

Table 1
Number of symptoms that appear across combinations of scales.

Symptoms	Scales	%
21	1	40
5	2	10
7	3	13
7	4	13
2	5	4
4	6	8
6	7	12

Table 2
Idiosyncratic and specific symptoms per scale..

Scale	Symptoms captured (No.)	Adjusted scale length (No.)	Idiosyncratic items (%)	Specific items (%)	Compound items (%)
IDS	33	28	18	70	30
QIDS	20	9	0	10	90
BDI	25	21	12	76	24
CES-D	21	18	33	76	24
SDS	23	20	9	78	22
MADRS	12	9	17	58	42
HRSD	22	17	9	55	45

Note: Symptoms captured, how many of the 52 specific symptoms does the scale capture; *adjusted scale length*, number of items per scale after combining similar items; *idiosyncratic items*, rate of items that appear in no other scale (e.g., 18% of the 33 symptoms captured by the IDS are idiosyncratic); *specific items*, rate of items that measure specific symptoms such as 'weight loss'; *compound items*, rate of items that measure compound symptoms such as 'weight loss or weight gain'.

changes), along with the number of specific symptoms captured per scale and the adjusted scale length.

The CES-D stands out as scale with the largest number of idiosyncratic items (33%). All other scales except for the QIDS feature between 9% and 18% idiosyncratic symptoms, whereas the QIDS encompasses no single item that does not appear in any other scale; this is not surprising when considering that the QIDS also has the largest amount of unspecific compound symptoms (90%) and captures items only very broadly. In contrast, only 22–45% of the symptoms assessed by the other scales are compound symptoms (Table 2).

3.2. Scale overlap

Finally, overlap among questionnaires was estimated via the Jaccard Index. The mean overlap among all scales is 0.36, which implies a weak similarity of the scales (Evans, 1996); specific overlap among all individual scales, and mean overlap of each scale with all other 6 scales, are presented in Table 3.

Several points in Table 3 are noteworthy.

(1) The CES-D and MADRS have the lowest mean overlap with other instruments of 0.27 and 0.30, respectively; mean Jaccard coefficients of the other scales range from 0.36 to 0.40. Only one coefficient (0.40) implies moderate similarity, all others can be considered weak.
(2) The BDI and QIDS exhibit the largest average overlap with all other scales (both 0.40; moderate overlap).
(3) The highest overlap among *individual* scales is between the IDS and QIDS (0.61; strong overlap). This is not unexpected, seeing that the IDS is a long version of the QIDS that captures additional items beyond the pure DSM criteria. The overlap among QIDS and BDI is also 0.61.
(4) The CES-D and IDS, and the CES-D and HRSD, have the lowest overlap (0.26 each; weak overlap); these two pairs of scales encompass very different items.
(5) A pattern emerged from the data: the correlation between the mean Jaccard coefficient of each scale (the mean overlap a scale with all others) and the length of the scale is 0.57 for the number of specific symptoms captured and 0.29 for the adjusted scale length (Table 2 columns 1 and 2); this implies that longer scales analyzed in this paper overlap more with others and thus feature more representative content.

Table 3
Overlap of item content of 7 depression scales.

	IDS	QIDS	BDI	CES-D	SDS	MADRS	HRSD
IDS	1.00	0.61	0.53	0.26	0.51	0.29	0.57
QIDS	0.61	1.00	0.61	0.28	0.43	0.39	0.50
BDI	0.53	0.61	1.00	0.35	0.50	0.37	0.42
CES-D	0.26	0.28	0.35	1.00	0.33	0.38	0.26
SDS	0.51	0.43	0.50	0.33	1.00	0.35	0.45
MADRS	0.29	0.39	0.37	0.38	0.35	1.00	0.31
HRSD	0.57	0.50	0.42	0.26	0.45	0.31	1.00
Mean overlap	0.39	0.40	0.40	0.27	0.37	0.30	0.36

Note: The Jaccard Index ranges from 0 (no overlap) to 1 (total overlap).

4. Discussion

The analyses identified a total of 52 specific disparate depression symptoms in 7 common depression scales. The overall overlap of item content among questionnaires was low: 40% of all symptoms appeared only in a single scale, only 12% across all instruments. These findings imply that the routine practice of using scales as interchangeable measurements of depression severity is problematic and may pose a major threat to the generalizability and replicability of depression research. Given the high prevalence rates and burden caused by MDD (Kessler et al., 2003), and the size of the research field – a non-exhaustive search of a few databases and a small number of journals identified around 50.000 depression articles published between 1990 and 1999 alone (Santor et al., 2006) – the severity of this situation can hardly be overstated.

4.1. Key findings

Consistent with prior studies that have investigated item content and overlap of depression rating scales (Polaino and Senra, 1991; Santor et al., 2006; Shafer, 2006; Snaith, 1993), the first key finding is the considerable difference in item content across instruments. Despite using a different analytic different approach than Santor et al. (2006), the BDI also emerged in the present study as scale most similar to others (along with the QIDS that was not analyzed by Santor et al., 2006).

The second key finding is that the CES-D stands out somewhat both in terms of idiosyncratic items and lack of overlap. This is not surprising, seeing that the CES-D captures items such as ‘people were unfriendly’ or ‘I felt that people disliked me’ that are not commonly understood to be depression symptoms. Santor et al. (2006) compared how 5 commonly used scales (including the CES-D, BDI, and HRSD) differed from a large pool of other depression scales in terms of the proportion of assessment of 5 symptom domains (such as ‘cognitive symptoms’ and ‘affective symptoms’). They also found the CES-D among the least representative scales, despite their different approach: they focused on domains instead of individual items, and compared scales not against each other, but against a variety of other scales less frequently used.

In the absence of a list of ‘true’ symptoms of MDD, however, the low overlap of CES-D items with other scales does not make the CES-D a bad scale; it merely implies that results identified with the CES-D are less likely to generalize to other scales. In contrast to the CES-D, the QIDS exhibits the highest amount of overlap in the present study, and

does not contain a single idiosyncratic item. Likewise, this does not make the QIDS a good scale, seeing that 90% of the QIDS items are unspecific compound items that do not capture particular aspects of depressive symptomatology and may thus be less informative. Note that the quality of scales here is primarily discussed in the context of assessing the problems patients with depression actually have, which is not the only possible objective. Chekroud et al. (2016), for instance, recently showed that the QIDS items at baseline predict treatment response better than the HRSD items.

4.2. Heterogeneity of the depressive syndrome

Where does the pronounced heterogeneity of depression scales come from? One reason may be that instruments reflect the diversity of clinical opinions regarding what depression is. MDD has been understood, among many others, as a brain disease, a clinical form of grief, an adaptive response to recurrent fitness threats, a set of self-defeating attitudes, and a strategy to conserve resources (Hagen, 2011; Santor et al., 2006; Snaith, 1993). The BDI symptoms, for instance, are based on the particular concept Beck and colleagues had about depression, and the scale features many cognitive symptoms that were central to Beck's theory (Beck et al., 1979). For Hamilton, on the other hand, anxiety and insomnia symptoms were especially important, with a strong overall focus on somatic depression symptoms that are easier to measure in a clinician-rated scale.

A second reason for differences among scales may be that they were developed for different purposes. Radloff (1977) selected the 20 CES-D items from prior scales with the goal to screen for depression in a general population sample, whereas Hamilton (1960) designed the HRSD to assess the severity of symptoms in patients already diagnosed with MDD. This, in turn, differs from the approach of Montgomery and Åsberg (1979) who constructed the MADRS by examining a scale of 17 symptoms and deleting 7 items that were not sufficiently responsive to pharmacological treatment.

4.3. Is lack of content overlap a challenge at all?

Before moving on to implications and future directions, it is important to address two arguments against the notion presented here that lack of content overlap leads to study results that are less likely to generalize.

First, consider an infectious disease that causes 40 disparate symptoms, all of which measure the underlying disorder roughly equally well. In this case, two scales with 20 different symptoms each can be considered to measure the same underlying construct – and to measure it equally well – although they share no single symptom. For MDD, however, such a situation unlikely holds: symptoms are not interchangeable indicators of depression, seeing that they are differentially related to external variables such as impairment, risk factors, or biomarkers (Fried and Nesse, 2015b), and seeing that scales are differentially related to external variables such as diagnosis (Zimmerman et al., 2012). Moreover, scales tap into multiple constructs that do not generalize across different scales (Shafer, 2006; van Loo et al., 2012).

The second argument is that lack of content overlap poses no problem since sum-scores of instruments are substantially interrelated. The evidence for convergent validity of depression instruments is mixed, and correlations ranging from 0 to virtually 1 have been reported in prior investigations (Fried et al., 2016b; Polaino and Senra, 1991). But even if the convergent validity were high – assume a correlation of 0.7 between the 20-item CES-D and 20-item SDS sum-scores – this does not imply that scales measure the same construct. To calculate the correlation among sum-scores, one needs only the inter-item correlation and the length of scales. If we take 40 disparate items that are only minimally related with each other (all inter-item correlations r=0.1), and distribute 20 symptoms each to the CES-D and SDS so that scales share no single item, the calculated correlation among sum-scores is 0.69 (see Supplementary Materials for a reproducible example). This shows that high convergent validity between two scales can be easily achieved even in cases they capture very different symptoms only minimally related.

4.4. Implications and future directions

Since different instruments capture different aspects of the heterogeneous depressive syndrome, there is the risk that the selection of a particular scale for a study may severely bias results (Santor et al., 2006; Snaith, 1993; Zimmerman et al., 2012). Considering the persistent lack of progress in core research areas such as antidepressant efficacy (Khan and Brown, 2015) and biomarkers robustly associated with depression diagnosis (Hek et al., 2013; Kapur et al., 2012), this topic deserves more attention in contemporary research.

The idiosyncrasy of scales also has strong implications for certain disciplines in which particular instruments are especially common. One example is the HRSD that has been used for decades as the gold standard to assess treatment efficacy of antidepressants (Santor et al., 2006). The HRSD encompasses predominantly somatic symptoms such as fatigue, insomnia, sexual dysfunction, and appetite/weight changes that closely resemble adverse effects caused by antidepressant treatment (Fried and Nesse, 2015b). We have to entertain the possibility that study results on the efficacy of antidepressants have been biased for decades because the most commonly used scale is especially sensitive to picking up treatment side-effects.

Furthermore, this study adds another piece to the literature of depression heterogeneity (Fried and Nesse, 2015a; Olbert et al., 2014), which begs the question how useful and informative sum-scores of rating scales are. Not only do symptoms differ from each other in important dimensions (Fried and Nesse, 2015b), scales also encompass many different *types* of symptoms (Snaith, 1993). On top of that, common rating scales are multidimensional, meaning they do not capture *one* underlying construct (Fried et al., 2016b; van Loo et al., 2012), making the routine use of *one* sum-score to reflect depression severity highly problematic. A way forward for now is the assessment of a broad range of symptoms; the 44-item Symptoms of Depression Questionnaire provides a good example and aims to capture the full range of clinical presentations of depression (Pedrelli et al., 2014).

Roughly 2 decades ago, Gullion and Rush (1998) (p. 959) noted that "a critical prerequisite for progress [...] is consensus on a) the array of symptoms that comprise the depressive syndrome and b) the use of adequate measures of these symptoms". We are nowhere near to agreeing which symptoms are good depression symptoms, or which scales are good depression scales. While the present study falls short of facilitating such a distinction, it stresses why future research efforts should be directed towards answering this crucial problem. One important topic is to determine which specific symptoms are common, severe, impairing, and clinically relevant in depressed patients. Only few large-scale studies have analyzed these questions on a symptom-level so far, and future studies are required to replicate and extend these findings. Results of two preliminary investigations of MDD outpatients from the large Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study suggest that the most central depression symptoms (derived from a network analysis) are also the ones patients experience as most impairing (Fried and Nesse, 2014; Fried et al., 2016a). Such symptoms identified as important across multiple domains could be considered especially relevant depression symptoms in future studies, for example regarding prevention and intervention. This means that the assessment of a wide range of symptoms seems crucial to capture the heterogeneity of the depressive syndrome.

The question which scales are adequate scales will require an empirical comparison of instruments regarding their performance and bias in specific domains; that there is currently no dedicated

literature on this problem shows the pervasiveness of the assumption that scales can be used interchangeably. The standard method to establish a depression scale as a good scale is to identify a sufficiently large convergent validity with other well-established instruments; however, as shown above, a high correlation of sum-scores does by no means guarantee that two scales measure the same construct. In addition, gold-standard scales to which novel scales are compared often have insufficient psychometric properties (Bagby et al., 2004). Addressing this challenge empirically would allow the scientific community to better gauge whether study results may be biased due to the selection of a specific instrument; we need to understand better whether scales are differentially related to personality dimensions, gender, or treatment response, to name but a few examples.

Using multiple scales when assessing depression severity goes into a similar direction. This will enable researchers to determine the robustness of their findings: if different depression scales explain widely different amounts of variance in a regression, perform differentially as mediator or moderator, lead to including very different subjects into a study, or result in different levels of treatment response, there is a problem that deserves further investigation. The use of multiple scales is already the established standard in certain fields such as clinical trials; interestingly, one outcome is often declared primary and given analytic precedence over secondary instruments. Until we understand scale performance better, a more conservative approach may be to examine the average impact of treatment across all scales, which also reduces the incentive for post-hoc changes of primary and secondary measures (Le Noury et al., 2015; Pigott, 2011).

Finally, as stated in the Strategic Plan for Mood Disorder Research of the National Institute of Mental Health (NIMH, 2003) (p. 93): "The most widely used instruments in clinical settings have generally failed to provide clear documentation of the symptoms experienced by individuals and [do not deal] effectively with the heterogeneity of depression [...]. Refined measurement is a prerequisite for studies that examine such associations, link genetic diatheses to particular forms of disorder, or guide the precise tailoring of future therapeutics". This implies that we need scales focused on the reliable assessment *depression symptoms*. Such scales should have the following important characteristics: (A) they should encompass a large variety of different specific symptoms; (B) they should, similar to other fields of psychometric assessment, assess each symptom with multiple questions to control for measurement error; (C) they should be constructed with the goal to measure specific symptoms reliably, not to measure *one* underlying construct for which there is no evidence after half a century of psychometric literature (Fried et al., 2016b). Such symptom-based instruments may provide a better rough index of depression severity (i.e. assessing general psychopathological load, not reflecting the severity of one underlying disorder), and facilitate symptom-based investigations such as the question whether specific antidepressants are particularly efficacious for patients suffering from specific symptoms. This is consistent with the agenda of the NIMH's Research Domain Criteria initiative (RDoC) (Insel et al., 2010); the framework encourages the examination of specific endophenotypes that may cut across current diagnostic categories, and "depression" symptoms such as insomnia, fatigue, or concentration difficulties that feature in various psychiatric disorders likely represent such dimensions worth investigating (Hasler et al., 2004).

4.5. Strengths and limitations

The present report goes beyond previous research examining item content in several aspects. First, the investigation focuses on *specific* symptoms in contrast to research on compound symptoms or dimensions. Snaith (1993), for example, grouped weight changes, appetite changes, and libido into one category, or the items productivity, activity, fatigue, and retardation. The focus on particular symptoms such as psychomotor retardation or different types of insomnia is crucial because they differ in important properties and should not be collapsed into compounds (Fried and Nesse, 2015b). Psychomotor retardation, for example, seems to respond much better to treatment than psychomotor agitation (Hieronymus et al., 2015), and retardation is more than 4 times as impairing as agitation (Fried and Nesse, 2014). Another example is that biological markers specific to particular insomnia symptoms (such as midnocturnal awakening) have been found (Myung et al., 2012). In contrast to prior reports, the current study also examines idiosyncrasy of scales along with the degree to which scales capture specific vs. compound symptoms.

The results of the report also have to be interpreted in the light of several limitations. First, the question arises whether the choice of the specific scales analyzed, or the fact that I included 7 (instead of 3 or 30) scales biased the results. Particular scales were chosen 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). A moderate amount of scales were analyzed because of my interest in the rate of idiosyncratic symptoms; in a study of 30 instruments, such items would be very rare. Furthermore, with 125 items across 7 scales, 1185 comparisons *within* scales were performed, ('k choose 2' for each scale, k being the number of items in a given scale), and 6242 comparisons *across* scales, leading to 7427 comparisons overall. More scales would have increased the burden of these comparisons exponentially. In conclusion, including more scales would have *increased* heterogeneity and unlikely decreased the low degree of overlap, and including *different* scales would unlikely change the results towards less heterogeneity or more overlap.

Second, and related, the content analysis entailed 7427 comparisons of symptom pairs: is symptom A1 from scale A similar enough to symptom B1 from scale B to consider them the same or not? While it would have been much preferable to have multiple raters compare all item pairs, this was not feasible. To counter this limitation, the author was very conservative and erred, whenever possible, on the side of caution, i.e.: considering symptoms rather too similar than too different. In addition, items were first combined within scales if they were worded very similarly, and symptoms considered to overlap across scales very broadly. While other raters would likely have come to somewhat different conclusions regarding some of the various choices, the current content analysis, if anything, underestimates the amount of heterogeneity and lack of overlap among scales.

Acknowledgements

I would like to extend my sincerest thanks to: Jana Jarecki, for help with Fig. 1; Sophie van der Sluis, for the calculation of sum-score correlations given scale length and inter-item correlation; and Don Robinaugh and Lauren Bylsma, for the very helpful comments on previous versions of this manuscript.

During the preparation of this manuscript, EIF was supported in part by the Research Foundation Flanders (G.0806.13), the Belgian Federal Science Policy within the framework of the Interuniversity Attraction Poles program (IAP/P7/06), the Grant GOA/15/003 from University of Leuven, and the European Research Council Consolidator Grant no. 647209.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jad.2016.10.019.

References

APA, 2013. Diagnostic and Statistical Manual of Mental Disorders Fifth Edition. American Psychiatric Association, Washington, DC.
Bagby, R.M., Ryder, A.G., Schuller, D.R., Marshall, M.B., 2004. Reviews and overviews

The Hamilton depression rating scale: has the gold standard become a lead weight? Am. J. Psyc 161, 2163–2177.
Beck, A.T., Rush, A.J., Shaw, F.S., Emery, G., 1979. Cognitive Therapy of Depression. Guilford Press, New York.
Beck, A.T., Steer, R.A., Ball, R., Ranieri, W., 1996. Comparison of beck depression inventories -IA and -II in psychiatric outpatients. J. Pers. Assess. 67, 588–597.
Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J., 1961. An inventory for measuring depression. Arch. Gen. Psychiatry 4, 561–571.
Cassano, P., Fava, M., 2002. Depression and public health. J. Psychosom. Res. 53, 849–857.
Chekroud, A.M., Zotti, R.J., Shehzad, Z., Gueorguieva, R., Johnson, M.K., Trivedi, M.H., Cannon, T.D., Krystal, J.H., Corlett, P.R., 2016. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry 0366, 1–8.
Evans, J.D., 1996. Straightforward Statistics for the Behavioral Sciences. Brooks/Cole Publishing, Pacific Grove, CA.
Fried, E.I., Epskamp, S., Nesse, R.M., Tuerlinckx, F., Borsboom, D., 2016a. What are "good" depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. J. Affect Disord. 189, 314–320.
Fried, E.I., Nesse, R.M., 2014. The impact of individual depressive symptoms on impairment of psychosocial functioning. PLoS One 9, e90311.
Fried, E.I., Nesse, R.M., 2015a. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. J. Affect Disord. 172, 96–102.
Fried, E.I., Nesse, R.M., 2015b. Depression sum-scores don't add up: why analyzing specific Depression symptoms is essential. BMC Med. 13, 1–11.
Fried, E.I., van Borkulo, C.D., Epskamp, S., Schoevers, R.A., Tuerlinckx, F., Borsboom, D., 2016b. Measuring Depression Over Time ... or not? Lack of Unidimensionality and Longitudinal Measurement Invariance in Four Common Rating Scales of Depression. Psychol Assess.
Gullion, C.M., Rush, A.J., 1998. Toward a generalizable model of symptoms in major depressive disorder. Biol. Psychiatry 44, 959–972.
Hagen, E.H., 2011. Evolutionary theories of depression: a critical review.
Hamilton, M., 1960. A rating scale for depression. J. Neurol. Neurosurg. Psychiatry 23, 56–62.
Hasler, G., Drevets, W.C., Manji, H.K., Charney, D.S., 2004. Discovering endophenotypes for major depression. Neuropsychopharmacology 29, 1765–1781.
Hek, K., Demirkan, A., Lahti, J., Terracciano, A., 2013. A genome-wide association study of depressive symptoms. Biol. Psychiatry 73 (7), 667–678.
Hieronymus, F., Emilsson, J.F., Nilsson, S., Eriksson, E., 2015. Consistent superiority of selective serotonin reuptake inhibitors over placebo in reducing depressed mood in patients with major depression. Mol. Psychiatry, 1–8.
Hieronymus, F., Nilsson, S., Eriksson, E., 2016. A mega-analysis of fixed-dose trials reveals dose-dependency and a rapid onset of action for the antidepressant effect of three selective serotonin reuptake inhibitors. Transl. Psychiatry (6), e834.
Insel, T.R., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., Wang, P., 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am. J. Psychiatry 167, 748–751.
Kapur, S., Phillips, A.G., Insel, T.R., 2012. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? Mol. Psychiatry 17, 1174–1179.
Kessler, R.C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K.R., Rush, A.J., Walters, E.E., Wang, P.S., 2003. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). JAMA 289, 3095–3105.
Khan, A., Brown, W.A., 2015. Antidepressants versus placebo in major depression: an overview. World Psychiatry 14, 294–300.
Le Noury, J., Nardo, J.M., Healy, D., Jureidini, J., Raven, M., Tufanaru, C., Abi-Jaoude, E., 2015. Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. bmj, 101006.
Montgomery, S.A., Asberg, M., 1979. A new depression scale designed to be sensitive to change. Br. J. Psychiatry 134, 382–389.
Myung, W., Song, J., Lim, S.-W., Won, H.-H., Kim, S., Lee, Y., Kang, H.S., Lee, H., Kim, J.-W., Carroll, B.J., Kim, D.K., 2012. Genetic association study of individual symptoms in depression. Psychiatry Res 198 (3), 400–406.
National Institute of Mental Health, 2003. Breaking Ground, Breaking Through: The Strategic Plan for Mood Disorders Research (NIH Publication No. 03–5121). National Institutes of Health, Washington, DC.
Olbert, C.M., Gala, G.J., Tupler, L.A., 2014. Quantifying heterogeneity attributable to polythetic diagnostic criteria: theoretical framework and empirical application. J. Abnorm Psychol. 123, 452–462.
Pedrelli, P., Blais, M.A., Alpert, J.E., Shelton, R.C., Walker, R.S.W., Fava, M., 2014. Reliability and validity of the Symptoms of Depression Questionnaire (SDQ). CNS Spectr. Oct., 1–12.
Pigott, H.E., 2011. STAR*D: a tale and trail of bias. Ethic Hum. Psychol. Psychiatry 13, 6–28.
Polaino, A., Senra, C., 1991. Measurement of depression: comparison between self-reports and clinical assessments of depressed outpatients. J. Psychopathol. Behav. Assess. 13, 313–324.
Radloff, L.S., 1977. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol. Meas. 1, 385–401.
Rush, A.J., Gullion, C.M., Basco, M.R., Jarrett, R.B., Trivedi, M.H., 1996. The inventory of depressive Symptomatology (IDS): psychometric properties. Psychol. Med. 26, 477–486.
Rush, A.J., Trivedi, M.H., Ibrahim, H.M., Carmody, T.J., Arnow, B., Klein, D.N., Markowitz, J.C., Ninan, P.T., Kornstein, S., Manber, R., Thase, M.E., Kocsis, J.H., Keller, M.B., 2003. The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol. Psychiatry 54 (5), 573–583.
Santor, D.A., Gregus, M., Welch, A., 2006. Eight Decades of Measurement in. Depress. Meas. 4, 135–155.
Shafer, A.B., 2006. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. J. Clin. Psychol. 62, 123–146.
Snaith, P., 1993. What do depression rating scales measure? Br. J. Psychiatry 163, 293–298.
van Loo, H.M., de Jonge, P., Romeijn, J.-W., Kessler, R.C., Schoevers, R.A., 2012. Data-driven subtypes of major depressive disorder: a systematic review. BMC Med. 10, 156.
van Noorden, R., Maher, B., Nuzzo, R., 2014. The top 100 papers. Nature 514, 550–553.
Zimmerman, M., Martinez, J.H., Friedman, M., Boerescu, D., Attiullah, N., Toba, C., 2012. How can we use depression severity to guide treatment selection when measures of depression categorize patients differently? J. Clin. Psychiatry 73, 1287–1291.
Zung, W.W.K., 1965. A self-rating depression scale. Arch. Gen. Psychiatry 12, 63–70.