Measuring depression over time . . . Or not?

Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression

Fried, E.I.; van Borkulo, C.D.; Epskamp, S.; Schoevers, R.A.; Tuerlinckx, F.; Borsboom, D.

Published in:
Psychological Assessment

DOI:
10.1037/pas0000275

Measuring Depression over Time … or not? Lack of Unidimensionality and Longitudinal Measurement Invariance in Four Common Rating Scales of Depression

Eiko I. Fried[1*]

Claudia D. van Borkulo[2,3]

Sacha Epskamp[3]

Robert A. Schoevers[2]

Francis Tuerlinckx[1]

Denny Borsboom[3]

1. University of Leuven, Faculty of Psychology and Educational Sciences, Leuven, Belgium.

2. University of Groningen, University Medical Center Groningen, Department of Psychiatry, Groningen, The Netherlands.

3. University of Amsterdam, Department of Psychology, Amsterdam, The Netherlands.

* Corresponding author: Dr. Eiko Fried, Faculty of Psychology and Educational Sciences, Tiensestraat 102, University of Leuven, Leuven, Belgium. Email: eiko.fried@gmail.com.

Preprint version of manuscript published in Psychological Assessment

Cite as: Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring Depression over Time … or not? Lack of Unidimensionality and Longitudinal Measurement Invariance in Four Common Rating Scales of Depression. *Psychological Assessment.* Advance online publication. DOI: 10.1037/pas0000275.

Abstract

In depression research, symptoms are routinely assessed via rating scales and added to construct sum-scores. These scores are used as a proxy for depression severity in cross-sectional research, and differences in sum-scores over time are taken to reflect changes in an underlying depression construct. To allow for such interpretations, rating scales must (1) measure a single construct, and (2) measure that construct in the same way across time. These requirements are referred to as unidimensionality and measurement invariance. We investigated these two requirements in two large prospective studies (combined *n*=3,509) in which overall depression levels decrease, examining four common depression rating scales (one self-report, three clinician-report) with different time intervals between assessments (between 6 weeks and 2 years). A consistent pattern of results emerged. For all instruments, neither unidimensionality nor measurement invariance appeared remotely tenable. At least 3 factors were required to describe each scale, and the factor structure changed over time. Typically, the structure became less multifactorial as people improved (without however reaching unidimensionality). The decrease in the sum-scores was accompanied by an increase in the variances of the sum-scores, and sharp increases in internal consistency. These findings challenge the common interpretation of sum-scores and their changes as reflecting one underlying construct. We discuss the possible causes of the violations such as response shift bias, restriction of range, and regression to the mean. The violations of common measurement requirements are sufficiently severe to suggest alternative interpretations of depression sum-scores as formative instead of reflective measures.

Keywords: exploratory structural equation modeling; major depression; measurement invariance; unidimensionality

Introduction

One of the primary goals of the social sciences is to describe, explain, and predict psychological constructs and their changes across time. While some constructs, like personality traits and intelligence, appear quite stable within individuals across adulthood (Costa & McCrae, 1997; Deary, 2012), other constructs, such as major depression (MD) can change dramatically: a person may be diagnosed with depression at one time point, but may not fulfill the diagnostic criteria for an episode of MD (APA, 2013) at a later point in time.

Measurement of such psychological constructs is complicated, not only because of their dynamic quality, but also because they do not admit direct observation such as a person's weight or body temperature do. Psychological traits in general, and mental disorders in particular, are routinely conceptualized as *latent variables* for this reason. More specifically, *reflective* latent variable models (Bollen & Lennox, 1991) are used to study mental disorders. Such models assume that psychological constructs cannot be directly measured, but can be assessed indirectly by examining their observable consequences; typically, these are taken to be reflective indicators, in the sense that their covariance can be explained in terms of the common influence of a latent variable. In other words, depression is assumed to cause the systematic covariation between its symptoms (Fried, 2015). These symptoms are commonly assessed with rating scales such as the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) or the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960). These instruments encompass a number of observable symptoms like sad mood, fatigue, concentration problems, and feelings of worthlessness or suicidality, all of which are understood as *measurements* of an underlying condition. After the assessment of depression symptoms, they are typically added to construct an unweighted sum-score. The common interpretation of this sum-score is that it reflects the severity

of the underlying condition, and that observed differences in total scores across time—e.g., a reduction by 50% in a group of patients—reflect changes in the underlying construct.

The veracity of this routine interpretation, that depression rating scales measure one underlying depression construct, is central to virtually all current research in the psychology, neurobiology, and genetics of depression, as well as to the assessment of treatment interventions. However, this notion is not a psychometric free lunch, and two important conditions are required for the assumption to hold: unidimensionality and temporal invariance.

Unidimensionality

The large majority of depression studies sum up all items to a single score, and interpret this score as reflecting *one* underlying construct. For this assumption to hold, scales should exhibit *unidimensionality*, which means that all items should load strongly on one primary factor. If the factor structure is multifactorial, however, the sum-score represents a mixture of several constructs that may, in some cases, not even be correlated. In such cases, adding items becomes very problematic. Prior studies are highly inconsistent when it comes to the question of unidimensionality in depression rating scales. For the BDI and the HRSD, anything from 1-7 factors have been extracted (cf. Gullion & Rush, 1998), and factor solutions rarely generalize across samples (Bagby, Ryder, Schuller, & Marshall, 2004). For the Center of The Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977) and the Montgomery-Åsberg depression rating scale (MADRS; Montgomery & Asberg, 1979), 1-4 factors have been extracted (cf. Fried & Nesse, 2015a; Quilty et al., 2013), and the literature is similarly inconsistent for many other scales (e.g., Elhai et al., 2012; Shafer, 2006; Wardenaar et al., 2010).

Measurement invariance

The second important psychometric assumption necessary to interpret changes of sum-scores as changes in depression is that MD is assessed in the same way at multiple occasions, a

requirement known as *temporal invariance* or *longitudinal measurement invariance* (Meredith, 1993; Widaman, Ferrer, & Conger, 2010). Measurement invariance implies that the relation between the latent variable and (the probability distribution of) its manifest indicators is invariant across occasions. If measurement invariance holds, changes in the sum-score of a given sample represent actual differences in the construct measured through the rating scale. If, on the other hand, measurement invariance is violated, observed differences over time do not necessarily reflect changes of the latent variable, and thus may offer limited or even misleading insights into the structure and causes of the true progress patients make. Similar to unidimensionality, prior literature on temporal invariance is inconsistent. While longitudinal measurement invariance has been established in a number of prior studies (e.g., non-clinical adolescents samples: Brunet et al., 2014; Motl, 2005; mothers with children suffering from epilepsy: Ferro & Speechley, 2013), other reports detected violations of measurement invariance (e.g., depressed patients: Uher et al., 2008; Rocca et al., 2002; Galinowski & Lehert, 1995; non-clinical sample of children: Lei et al., 2014; elderly twins: Wetherell, Gatz, & Pedersen, 2001).

The present study

The purpose of the present study is to systematically examine whether unidimensionality and temporal invariance are tenable assumptions in typical studies of depression. To do so, we test these two conditions in two large prospective datasets with a total sample of 3,509 participants, in four widely used depression rating scales (one self-report and three clinician-report instruments), with varying intervals between measurement points (ranging from 6 weeks to 2 years).

We also address two potential limitations of prior studies on temporal invariance, related to the model family of confirmatory factor analysis (CFA). This approach requires a sufficient number of fixed zero-loadings (i.e. simple structure) that may be overly restrictive and

inappropriate for many psychological data, including depression data (Dolan, Oort, Stoel, & Wicherts, 2009; Marsh, Morin, Parker, & Kaur, 2014). Furthermore, the specification of CFA models is often derived from the literature, but factor solutions are highly volatile for depression instruments. As described above, many different factor solutions have been extracted for most scales, and choosing one specific model for a rating scale over all the others *a priori* is very difficult. The recent study of Quilty et al. (2013) on the MADRS, in which the authors fit 12 CFA models established in the prior literature to data, exemplifies the problem. No model described the data well, and only one provided acceptable fit. It is thus conceivable that violations of measurement invariance could be caused by the fact that auxiliary hypotheses, like simple structure, are not tenable and compromise model fit. In our study, we therefore lift the assumption of simple structure, as well as the reliance on volatile a priori models, by using exploratory structure equation models (ESEM).

Methods

Samples

Two datasets were analyzed for this report. The first dataset was version 3 of the NIH-supported "Sequenced Treatment Alternatives to Relieve Depression" (STAR*D) study (Rush et al., 2004), a multisite randomized clinical trial to investigate treatment efficacy for nonpsychotic MDD outpatients. The first treatment stage encompassed 4,041 patients who received the antidepressant citalopram. STAR*D was approved and monitored by the institutional review boards at each of the 14 participating institutions, and all participants provided written informed consent at study entry. Participants for STAR*D had to be between 18 and 75 years old, fulfill DSM-IV criteria for single or recurrent nonpsychotic MDD, and have at least moderately severe

depression (at least 14 points on the HRSD). Exclusion criteria were a history of bipolar disorder, schizophrenia, schizoaffective disorder, or psychosis, as well as current anorexia, bulimia, or primary obsessive compulsive disorder. Further details on the STAR*D study can be found elsewhere (Rush et al., 2004). Our analyses are limited to the 2,745 individuals who provided data during the first treatment stage.

We analyzed specifically the STAR*D dataset to test for unidimensionality and measurement invariance for a number of reasons. First, it is one of the largest antidepressant trials conducted so far, and thus offers are large sample size and multiple timepoints. Second, the raw data for STAR*D can be obtained through the NIMH, which is not the case for the large majority of clinical trials. Third, STAR*D used comparably broad inclusion criteria, leading to a sample that is considered representative of the depressed population in general; this is important because most other depression studies examine rather artificial populations, often without any comorbidities, that are not considered to reflect the large majority of depressed patients that seek psychiatric help (Preskorn, Macaluso, & Trivedi, 2015; Wisniewski, Rush, Nierenberg, & Gaynes, 2009). Finally, STAR*D tracked antidepressant response with multiple rating scales, allowing us to test the robustness of the findings across different instruments.

The second dataset for which we investigated unidimensionality and measurement invariance was the "Netherlands Study of Depression and Anxiety" (NESDA) dataset (Penninx et al., 2008). This prospective cohort study examines the long-term course and consequences of mood and anxiety disorders. Participants (aged 18-65 years) were included from the community (19%), general practice (54%), and secondary mental health care (27%). At baseline, individuals with a current or history of mood and/or anxiety disorders were enrolled, along with a healthy control group (total n=2,981). The ethical boards of the participating centers approved the study and all participants provided written informed consent. For the current report, we selected all 649

participants with past-month DSM-IV MDD diagnosis at baseline who also participated in the two-year follow-up.

We analyze the NESDA data in addition to the STAR*D dataset for reasons of robustness. An important difference between datasets is that NESDA used a self-report rating scale to gauge depression severity, while STAR*D employed three clinician-rated scales that are described below in detail. Furthermore, the NESDA data encompass a prospective period of two years, compared to a maximum of 11 weeks in the STAR*D dataset. If results generalize across samples, scales, and time-frames, findings can be considered more robust.

Outcomes measures

Overall, we examined four outcome measures, three that were assessed in the STAR*D study, and one that was used in the NESDA study.

Hamilton Rating Scale for Depression (HRSD). STAR*D used the clinician-rated 17-item version of the HRSD at entry (week 0) and exit (week 11) of the first treatment stage. The HRSD is still one of the most commonly used rating scales for depression (Bagby et al., 2004; Santor, Gregus, & Welch, 2009), despite its problematic psychometric properties. Extensive reviews reported factor structures ranging from 1 to 7 factors with poor replication across samples, reliability estimates in a range from 0.46 to 0.97, and poor interrater reliability, retest reliability, and content validity (Bagby et al., 2004; Gullion & Rush, 1998). Of note, Rush et al. (1996) observed that the reliability of the HRSD in a mixed sample of both healthy and depressed participants was substantially higher (0.88-0.89) compared to a sample of only depressed participants (0.53-0.56).

Inventory of Depressive Symptoms, clinician-rated version (IDC-C). The 30-item clinician-rated Inventory of Depressive Symptoms (IDS-C) was assessed alongside the HRSD at study entry and exit (Rush et al., 1996) in STAR*D. The IDS-C covers 30 items that include the

DSM-5 (APA, 2013) criterion symptoms of MD, and other symptoms such as anxiety and irritability common among depressed patients. The IDS-C was developed, among other reasons, because the authors suggested that other instruments such as the HRSD have poor item content (Rush et al., 1996). The items 'weight gain' and 'weight loss', and 'appetite increase' and 'appetite decrease', are combined into 'weight problems' and 'appetite problems', since only one symptom in each domain was scored in the STAR*D study. Overall, this leads to an 28 items. All other DSM-5 compound symptoms, such as 'psychomotor agitation or psychomotor retardation', or 'loss of interest or loss of pleasure', are queried by the IDS-C as individual items, providing a large amount of specific symptom information. The IDS-C has satisfactory psychometric properties, with reliability estimates between 0.94 in a mixed sample (including both depressed and healthy participants) to 0.67 in a depressed sample (Rush et al., 1996). Unidimensionality of the IDC-C could not be established in a sample of 4041 depressed outpatients (Bech, Fava, Trivedi, Wisniewski, & Rush, 2011).

Quick Inventory of Depressive Symptoms, clinician-rated version (QIDS-C). STAR*D also used the clinician-rated version of the 16-item Quick Inventory of Depressive Symptoms (QIDS-C) to assess depressive symptoms every two weeks throughout the first treatment phase during clinical visits. The QIDS-C is the short version of the IDS-C (Rush et al., 2003) and focuses only on DSM criteria. While the QIDS-C assesses disaggregated symptoms such as 'psychomotor agitation' and 'psychomotor retardation', the 16 items are commonly scored into the 9 DSM symptom dimensions of MD, a procedure that loses a large amount of information. Combining opposites such as 'hypersomnia' and 'insomnia', or 'psychomotor agitation' and 'psychomotor retardation', into composite scores seems problematic, considering that sub-symptoms have important distinct properties (e.g., psychomotor agitation is about four times as much impairing as psychomotor retardation; for a review, see Fried & Nesse, 2015b) and

are differentially severe in depressed populations (e.g., insomnia is about four times as severe as hypersomnia in STAR*D patients; Fried & Nesse, 2014). Consistent with a prior study (Fried & Nesse, 2014), we therefore retained 14 instead of just 9 symptoms in our analyses; analogous to the IDS-C, the items 'weight gain' and 'weight loss', and 'appetite increase' and 'appetite decrease', had to be combined into 'weight problems' and 'appetite problems', because only one of the two respective items was queried in STAR*D). Our decision for this specific coding of the items was motivated to conserve item content, but also for reasons of consistency across scales (we coded the same items in exactly the same way in the QIDS-C and IDS-C). For the QIDS-C, we analyzed the timeframe between study entry (week 0) and midpoint of the first treatment stage (week 6) to have a timeframe different from the other two STAR*D instruments that encompass the full first treatment stage period (11 weeks). We consider this difference in timeframes part of our robustness analysis: if the results generalize across different rating scales for different timeframes in different populations, result can be understood to be more robust. Prior studies established good psychometric properties of the QIDS-C, with a reliability estimate of 0.85 among MD patients (Trivedi et al., 2004), along with unidimensionality of the 9 symptom domains (Rush et al., 2006). It should be noted that both reports used the exit time points of longitudinal treatment studies to establish reliability and unidimensionality, implying that these psychometric qualities were identified in mixed samples consisting of depressed and remitted participants, but not in a population of only depressed patients.

Inventory of depressive symptoms, self-rated version (IDS-SR). In the NESDA study, the self-report version of the IDS (IDS-SR; Rush et al., 1996) was used to assess depressive symptoms at baseline and two-year follow-up; the 30 items are identical to the IDS-C described above, and were coded into 28 items analogous to the IDS-C. Between 1 and 4 factors have previously been extracted for the IDS-SR (Wardenaar et al., 2010), and reliability estimates seem

to differ depending on the type of sample studied. One study reported that the reliability of the IDS-SR increased substantially (from 0.57 to 0.85) in a depressed sample during 12 weeks of treatment (Rush et al., 2003), while another study reported a reliability of 0.93 to 0.94 in a mixed sample of both healthy and depressed participants, and a reliability of 0.77 in a purely depressed sample (Rush et al., 1996).

Statistical Analysis

Unidimensionality. To test for unidimensionality, we examined the number of factors needed to describe each rating scales at each time point. We conducted a series of exploratory factor analyses (EFA), and determined the optimum number of factors that should be extracted via a parallel analysis. Parallel analysis compares the observed eigenvalues with eigenvalues of randomly drawn data, and we extracted factors for which the eigenvalues exceeded the randomly generated eigenvalues (O'Connor, 2000). We generated 50 parallel datasets for each analysis, and used 95% eigenvalue percentiles[1].

Moreover, we fit a 1-factor confirmatory factor analysis (CFA) to each instrument at each measurement point to assess the fit of unidimensional models; we used the root mean square error of approximation (RMSEA; ≤ 0.06 indicating a good fit) and the comparative fit index (CFI; ≥ 0.95 indicating a good fit) (Hu & Bentler, 1999).

Measurement Invariance. Our second goal was to investigate longitudinal measurement invariance. As discussed above, temporal invariance has mostly been tested using the framework of CFA, requiring simple structure and the specification of models based on the highly inconsistent literature. Therefore, we face the major challenge to specify factor models for each of the four rating scales analyzed here that are not arbitrary. ESEM offer a potential solution to

[1] Note that we generated eigenvalues using both the resampling and the simulation method provided by the function *fa.parallel* from the R-package *Psych* (Revelle, 2015); in all cases, both methods yielded the same number of factors to extract.

this problem. In ESEM, an EFA is carried out to search for a measurement model that describes the data best, and items are allowed to load on multiple factors (Dolan et al., 2009; Marsh et al., 2014). Similar to traditional measurement invariance testing, a hierarchical set of equality constraints is employed when investigating temporal invariance in the ESEM framework. It is of note that we do not use these exploratory models to obtain a substantively plausible model, or an interpretable model; in fact, because all items are allowed to load on all factors, ESEM models can be very difficult to interpret. We also do not advance ESEM as a plausible description of the data or data-generating mechanisms—the models exclusively function as the most feasible way to test measurement invariance in a dataset in which a dramatic change of the factor structure can be expected from prior findings. We thus use ESEM to achieve a baseline model that is as neutral as possible, because of the difficulties to justify a specific a priori model based on the literature as discussed above. ESEM gives the model all possibilities to fit the data, but we do not mean to say that the resulting model is a reasonable model, or will replicate in this specific form in other datasets. ESEM provides the least restrictive model that still allows for tests of measurement invariance, and hence minimizes the risk that measurement invariance is rejected because auxiliary hypotheses like simple structure are not tenable.

Similar to measurement invariance testing in the realm of confirmatory models, four models with increasing constraints are estimated, and each model is compared to the previous one using a χ^2 difference test. If introducing equality constraints decreases the fit significantly, measurement invariance is rejected. First, a configural invariance model M1 is fit to the data of all measurement points per rating scales that imposes no equality constraints on the parameters and only restricts the number of factors to be equal across time. In the next step, the weak factorial invariance model M2 is estimated that constrains item loadings to be equal across time.

The strong factorial invariance model M3 additionally constrains thresholds to be equal across time, and the strict invariance model M4 forces all residual invariances to be equal on top of all previous constraints. measurement invariance can be established only if M4 is not rejected in this iterative procedure (Meredith, 1993).

In general, measurement invariance refers to the invariance of the probability distribution (or density) of the observed scores, given the latent variable. Strong measurement invariance exists when this distribution is entirely invariant over groups; weak measurement invariance means that only the first moment (the expected value) is invariant. These requirements cannot be tested directly, because we cannot observe the latent variable; however, within a given family of models (e.g. the factor model) we can work out the model restrictions that, if true, would guarantee measurement invariance. In the factor analysis context, it turns out that, for measurement invariance to hold, strict factorial invariance is required (Meredith, 1993). Thus, any interpretation of scores that is based on strong measurement invariance requires invariance of intercepts, factor loadings, and error variances. If only the means of the observed scores are interpreted, strong factorial invariance (invariance of intercepts and factor loadings) is sufficient. Thus, to e.g. interpret mean differences over groups, it is generally necessary to have strict factorial invariance. However, there are various weaker goals of test use that do not require even strong factorial invariance; e.g. comparing the direction of correlations in groups is possible with only configural invariance (i.e. invariant factor loadings). Borsboom (2006) provides an overview of the way different kinds of test use correspond to different kinds of invariance.

Modeling standards. All analyses were based on polychoric correlations to account for the ordered-categorical nature of the symptoms. We used the oblique geomin rotation to rotate factors, and the weighted least squares means and variance adjusted (WLSMV) estimator, a robust estimator that provides the best option for modeling ordered-categorical variables (Brown,

2006). For the temporal invariance models, we used the theta parameterization and followed the measurement invariance model specifications described in detail by Millsap (2011). χ^2 values derived from the WLSMV estimator do not have the same behavior as other χ^2 statistics and cannot be used in the context of standard χ^2 difference testing for purposes of model comparison; instead, we used the Mplus *difftest* procedure, which is specifically developed for this situation. There were serious convergence problems when item residuals were not allowed to be correlated across time; an inspection of the modification indices revealed that these missing residual correlations were a major source of model misfit. In line with previous papers (Fokkema, Smits, Kelderman, & Cuijpers, 2013; Oort, 2005; Vandenberg & Lance, 2000), we therefore allowed the residuals of each item to be correlated across time. All structural equation models (EFA, CFA, and ESEM) and tests of measurement invariance were estimated in Mplus 7.3 (Muthén & Muthén, 2012); all other analyses were conducted in R 3.1 (R Development Core Team, 2008).

Results

Descriptive statistics

Final analytic samples were obtained through listwise deletion; no participants were excluded for other reasons than missing data. From the STAR*D data, we included 1,938 participants queried on the QIDS-C at weeks 0 and 6, and 2,522 individuals who provided data on the HRSD and IDS-C at weeks 0 and 11. Together, these two sub-samples represent a total sample of n=3,013, seeing that 1,447 participants are included in both subsamples. From the NESDA study, 496 participants were included in the final sample that provided data at measurement points two years apart. In the STAR*D study, the mean age of the 3,013

participants was 41.7 (*SD*=13.2), and 62.2% were female. The NESDA sample comprised 496 participants with a mean age of 39.7 (*SD*=12.6), of whom 66% were female.

Information about changes measured via the different rating scales over time is presented in Table 1. Overall, paired *t*-tests revealed that all sum-scores decreased significantly across occasions (QIDS-C: $t(1,937)=62.97$; HRSD: $t(2,521)=51.30$; IDS-C: $t(2,521)=51.87$; IDS-SR: $t(495)=20.19$; all $p<0.001$) (Table 1, row 2); the *SD* of the total scores, in contrast, increased in all scales (Table 1, row 3). The density plots in Figure 1 show the distribution of the total scores of all rating scales; they were fairly normally distributed at baseline levels and became increasingly skewed over time.

Table 1. *Changes across time*

#		QIDS-C		HRSD		IDS-C		IDS-SR	
1	Time (weeks)	0	6	0	11	0	11	0	104
2	*M* sum-score	16.27	9.53	19.68	11.51	35.50	20.60	35.11	24.55
3	*SD* sum-score	3.36	4.87	6.40	8.44	11.32	15.26	11.14	12.70
4	α	0.68	0.85	0.81	0.92	0.85	0.95	0.87	0.92
5	Mean *r*	0.12	0.28	0.16	0.33	0.16	0.39	0.19	0.30

Note: *M*, mean; *SD*, standard deviation; α, Cronbach's alpha; mean *r*, mean of the polychoric correlations of all items.

The correlations among sum-scores of different scales in case of overlapping samples were: 0.88 between HRSD and IDS-C at baseline; 0.53 between HRSD and QIDS-C at baseline; 0.58 between IDS-C and QIDS-C at baseline; and 0.96 between HRSD and IDS-C at week 11.

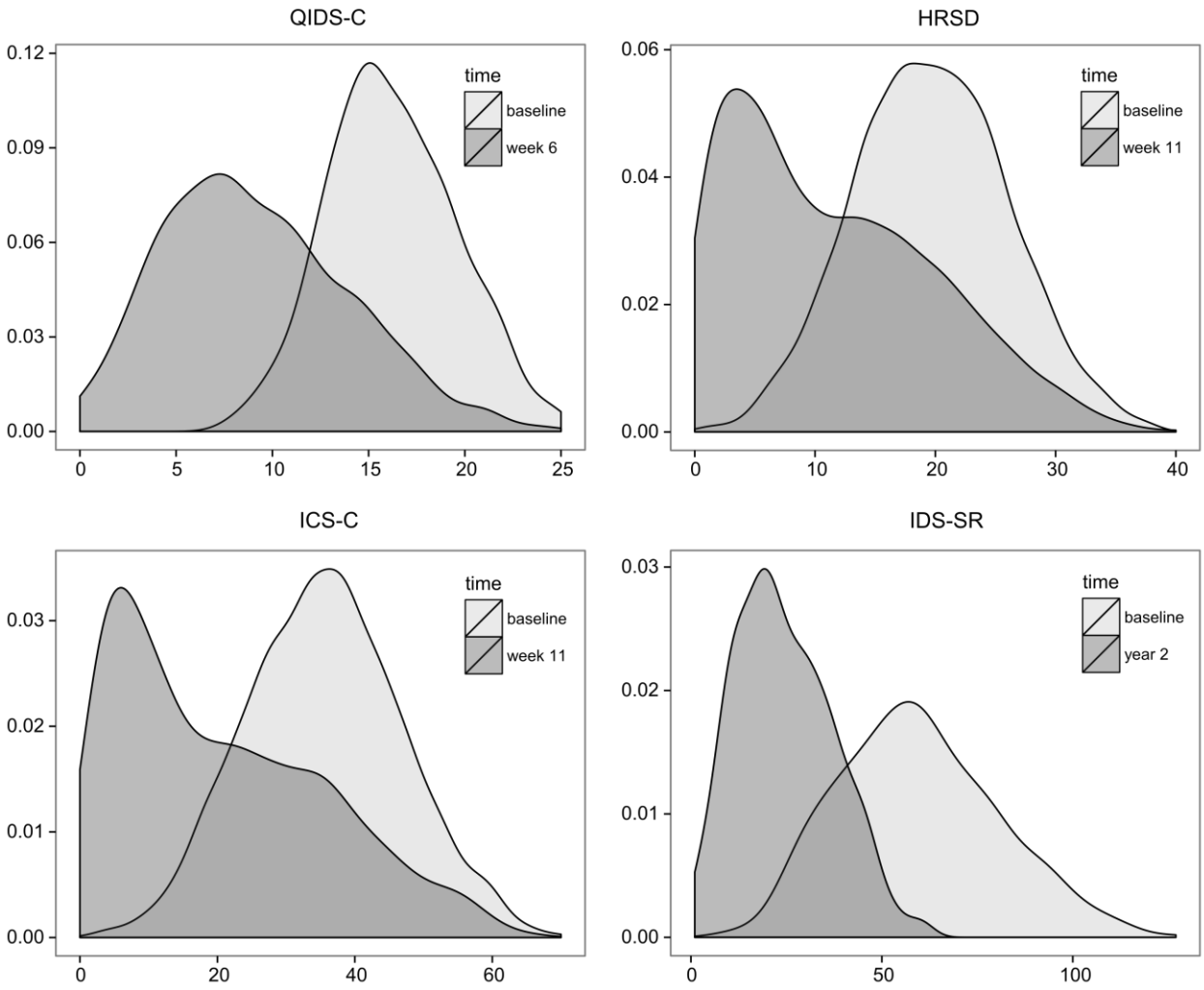

Figure 1. Density plots of the total scores of all rating scales at all measurement points.

Unidimensionality

To determine whether the rating scales exhibited a unidimensional factor structure, we compared the eigenvalues of the factors with the results of the parallel analysis (Figure 2). While one dominant factor emerged for all instruments, the eigenvalues of several factors beyond the first one exceeded the eigenvalues derived from a parallel analysis of random data, implying a more complex factorial structure. In detail, the parallel analyses suggested (a) for the QIDS-C to extract 3 factors at both measurement occasions, (b) for the HRSD to extract 4 factors at baseline

and at least 3 at follow-up, (c) for the IDS-C to extract 6 factors at baseline and 3 at follow-up, and (d) for the IDS-SR to extract 4 factors at baseline and 3 at follow-up.

The results of the unidimensional CFA presented in Table 2 are consistent with the findings of the parallel analysis. A 1-factor solution did not describe any of the rating scales well at any occasion. It is of note that the fit of these unidimensional CFA improved over time (i.e., as the average sum-score of individuals improved) for all instruments, which is consistent with the larger eigenvalues of the primary factors of all instruments at later time points (see Figure 2), indicating a *decrease* in dimensionality. Despite an increase in fit, however, none of the models did meet conventional cutoffs for good fit.

Table 2. *Unidimensional confirmatory factor analyses*

Instrument	Time	χ^2	*df*	RMSEA	CFI
QIDS-C	Week 0	1595.78	77	0.10	0.58
QIDS-C	Week 6	1778.78	77	0.11	0.87
HRSD	Week 0	3053.67	119	0.10	0.69
HRSD	Week 11	2348.07	119	0.09	0.93
IDS-C	Week 0	5032.20	350	0.07	0.76
IDS-C	Week 11	4470.54	350	0.07	0.94
IDS-SR	Year 0	1190.35	350	0.07	0.84
IDS-SR	Year 2	1128.32	350	0.07	0.92

Note: χ^2, chi-square statistic; *df*, degrees of freedom; RMSEA, root mean square error of approximation; CFI, comparative fit index.

Detailed information on item proportions and counts for the categorical symptoms, factor loadings, thresholds, and item intercorrelations of all CFA models are available in the online supplementary materials.

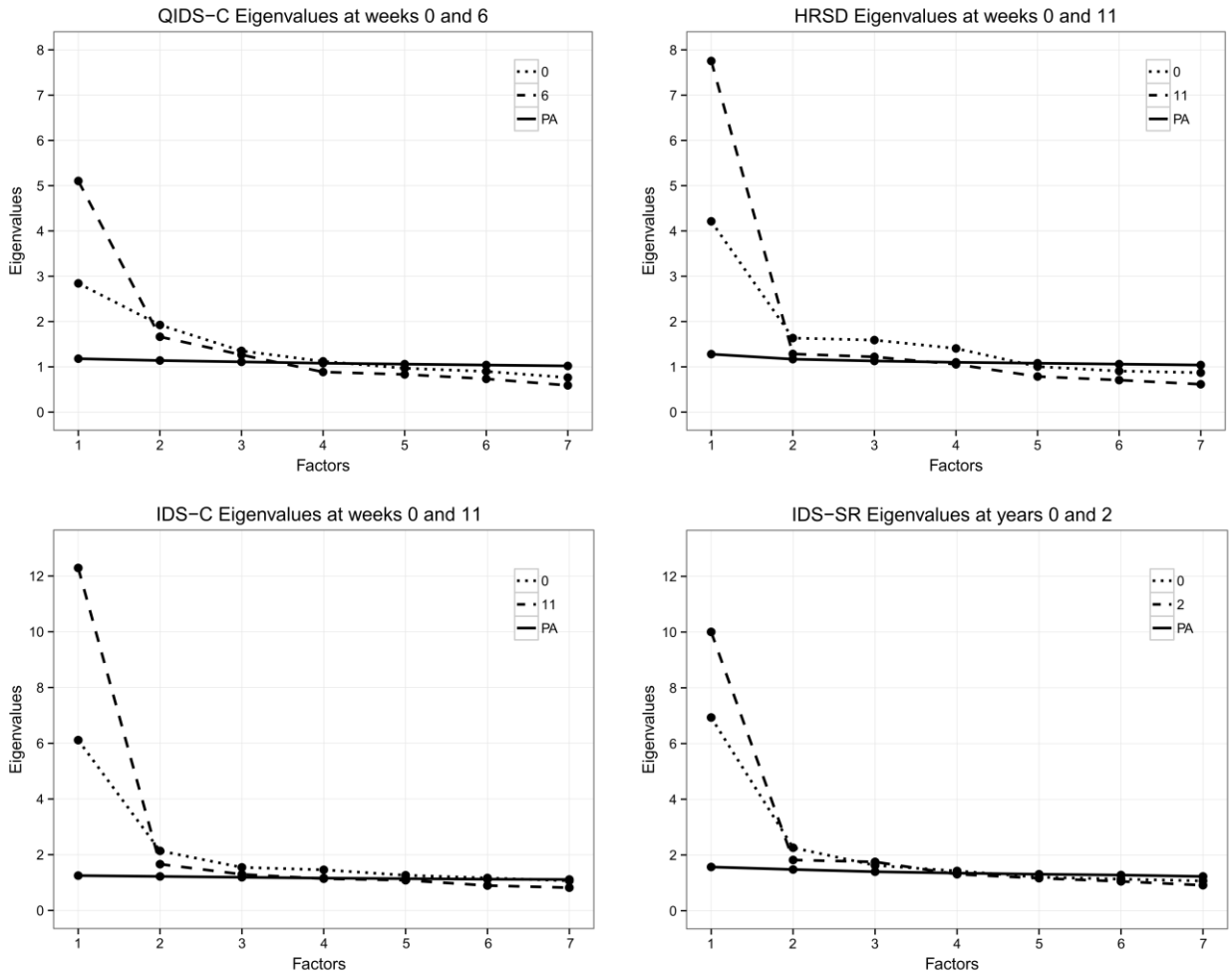

Figure 2. The plots show the eigenvalues of the factors observed in the data compared to eigenvalues generated via parallel analyses (PA). Eigenvalues larger than the ones resulting from the parallel analyses should be extracted.

Measurement invariance

Configural invariance – the question whether the same number of factors describe a given rating scales equally well in repeated measures – could only be established for the QIDS-C, seeing that the parallel analysis suggested to extract 3 factors at all time points; configural invariance is necessary, but not sufficient, for measurement invariance to hold. For the other three

scales, the number of factors recommended by the parallel analysis differed across time, implying a severe violation of measurement invariance.

If the structure of the latent space is not invariant across time, measurement invariance must be rejected. However, to assess to what extent temporal invariance is further violated given a reasonable choice for a common model across time points, we decided to fit measurement invariance models using an ESEM model that comprises 3 factors for each instrument, representing the lower bound of the factor estimation per rating scale. Table 3 provides an overview of the measurement invariance models, the fit indices, and model comparison tests[2]. Detailed information for all ESEM models, such as factor loadings, thresholds, residual variances, factor means, factor variances, and factor correlations are available in the online supplementary materials.

As expected, all four ESEM baseline models M1 fit the data well, as they are extremely flexible and impose few restrictions on the data. However, even using this technique, M2 (weak invariance), M3 (strong invariance), and M4 (strict invariance) were still rejected in all depression scales. This means that we cannot assume measurement invariance to hold: first, because the factor structure is clearly not invariant over time, and second, because even with respect to an "average" amalgam model, parameters of that model are not invariant.

The results suggest that unidimensionality and measurement invariance are untenable assumptions for the rating scales analyzed here. However, they also suggest that, below the surface of sum-scores, there is a highly interesting psychometric pattern of change that may offer

[2] Not all measurement invariance models for the IDS-C converged. This is not surprising, given that the structure of the latent space is not invariant across time. We decided to reduce the complexity of the ESEM models for the IDS-C by only allowing for correlated item residuals across time that would otherwise lead to a very severe misfit (arbitrarily defined as a change in $\chi^2 > 100$; these were items 1, 4, 9, 18, 22, 24, 25, 28, 29, and 30). After this change, all IDS-C models converged.

important insights into the structure of depression. To study that pattern of change, we complemented the findings of decreased dimensionality across time with two exploratory analyses. First, we found that the reliability of each scale increased sharply, especially in the QIDS-C (Table 1, row 4). Second—consistent with this result—the average correlation among symptoms increased in all scales across time (Table 1, row 5).

Table 3. *Exploratory structural equation modeling measurement invariance analyses*

Rating scale	Model	χ^2	*df*	RMSEA	CFI	χ^2-test	χ^2_{diff}	df_{diff}	*p*
QIDS-C	M1	854.67	277	0.03	0.97	-	-	-	-
	M2	935.56	307	0.03	0.97	M1 vs. M2	107.54	30	< 0.001
	M3	1461.19	335	0.04	0.94	M2 vs. M3	526.08	28	< 0.001
	M4	1518.33	349	0.04	0.94	M3 vs. M4	92.77	14	< 0.001
HRSD	M1	2420.11	439	0.04	0.95	-	-	-	-
	M2	3505.92	479	0.05	0.93	M1 vs. M2	975.06	40	< 0.001
	M3	4418.60	515	0.06	0.91	M2 vs. M3	977.59	36	< 0.001
	M4	4299.74	532	0.05	0.91	M3 vs. M4	90.43	17	< 0.001
IDS-C	M1	5854.21	1359	0.04	0.95	-	-	-	-
	M2	6115.91	1429	0.04	0.95	M1 vs. M2	409.68	70	< 0.001
	M3	6352.35	1480	0.04	0.94	M2 vs. M3	352.16	51	< 0.001
	M4	6294.91	1508	0.04	0.95	M3 vs. M4	482.51	79	< 0.001
IDS-SR	M1	1935.05	1341	0.03	0.96	-	-	-	-
	M2	2025.68	1411	0.03	0.96	M1 vs. M2	132.52	70	< 0.001
	M3	2078.10	1463	0.03	0.96	M2 vs. M3	76.57	52	< 0.05
	M4	2104.64	1491	0.03	0.96	M3 vs. M4	55.63	28	< 0.01

Note: All models are 3-factor exploratory structural equation models. M1, configural invariance; M2, weak invariance; M3, strong invariance; M4, strict invariance; χ^2, chi-square statistic (due to the WLSMV estimator not directly interpretable); *df*, degrees of freedom; RMSEA, root mean square error of approximation; CFI, comparative fit index; TLI, Tucker-Lewis Index; χ^2_{diff}, statistic of the χ^2 difference test; df_{diff}, degrees of freedom of the χ^2 difference test; *p*, *p* value of the χ^2 difference test.

Discussion

In studies of depression, sum-scores are routinely used to reflect the severity of one underlying disorder and changes in sum-scores are used to represent differences in true scores. This interpretation is valid only to the extent that rating scales are unidimensional and measurement invariant across time. In our analyses of four common depression instruments, assessed in two large samples with repeated measures with a total number of 3,509 depressed participants, we could establish neither unidimensionality nor temporal invariance. Specifically, we found that the means of the total scores systematically decreased while their variances increased, the factor structure of the depression scales became less multifactorial (but not unidimensional), and reliability increased as the inter-correlation among items increased. Of note, the results were consistent across two different samples, various time frames ranging from 6 weeks to 2 years, and across four self-report and clinician-rated scales.

This suggests that the instruments analyzed in this report (a) do not assess a single underlying construct, and (b) do not measure the same (set of) construct(s) in the same way across time. These two assumptions, however, are crucial to the validity of routine interpretations of nearly all current research on depression—including the study of treatment approaches. In the following two sections, we discuss this pattern of results in relation to prior literature.

Unidimensionality

As reviewed in the introduction, the lack of unidimensionality of common depression rating scales will not come as a surprise to those familiar with the psychometric literature (e.g., Bagby et al., 2004; Gullion & Rush, 1998; Shafer, 2006), and developers of the instruments themselves have often acknowledged this. Radloff, for instance, recommended to describe the CES-D with 4 factors, Beck suggested to extract at least two factors for the BDI, and Rush et al.

identified 3 factors in their psychometric study of the IDS-30 (Rush et al., 1996). Multifactorial results also emerge when symptoms of several questionnaire are pooled (Uher et al., 2008). Given the strong evidence for a lack of unidimensionality, sum-scores should not be interpreted as reflecting the severity of *one* underlying condition. In our study, one strong factor emerged for all scales, but several additional factors were required to explain the covariance among items. Consistent with this observation, unidimensional CFA did, overall, not describe the data well, although it can be argued that some of the models do show acceptable levels of fit (especially at the second measurement point), depending on the particular thresholds used to determine fit (c.f. Kline, 2005).

Considering the dramatic heterogeneity of the depressive syndrome (Fried & Nesse, 2015a; Olbert, Gala, & Tupler, 2014; Zimmerman, Ellison, Young, Chelminski, & Dalrymple, 2014), multifactorial results in depression instruments cannot come as a surprise. There is a large number of disparate psychiatric symptoms of depression—including (a) symptoms considered typical for depression such as sad mood and anhedonia, (b) symptoms that are common in many other mental and medical conditions such as fatigue and insomnia, (c) bi-directional symptoms such as weight loss vs weight gain, psychomotor agitation vs retardation, and insomnia vs hypersomnia, and (d) more exotic symptoms such as hypochondriasis, loss of insight, and genital problems. In a recent study, we identified 1030 unique DSM symptoms profiles of MD in 3703 depressed patients; complex factorial structure can be expected in such a highly heterogeneous syndrome.

Temporal invariance

As described in more detail in the introduction, the prior literature is inconsistent regarding longitudinal measurement invariance of depression scales. Armed with the consistent and intriguing results identified in this report, previously unconnected pieces from the literature

suddenly fell into place and we saw a pattern: the dimensionality of depression scales may vary as a function of the *severity* of the studied samples.

Apart from our report, there are several independent sources that substantiate this conclusion. First, cross-sectional studies that identified especially high reliability of MD rating scales have often analyzed the *exit time point* of clinical trials, i.e. the measurement point for which depression was less severe in the population because a number of remitted (healthy) participants were included (e.g., Rush et al., 2006; Trivedi et al., 2004). Second, one study reported the reliability of four different depression scales in two different samples—a mixed sample, including both healthy and depressed participants, and a depressed sample (Rush et al., 1996). While the authors do not discuss the pattern of observations, they found that the reliability was consistently higher in the mixed sample (between 0.88 and 0.94) than in the depressed sample (from 0.53 to 0.83). Although reliability does not bear a direct analytic connection to unidimensionality (Sijtsma, 2009), if these changes occur in the same way as the changes we observed, this suggests that the dimensionality of the data may have decreased in these cases as it did in the ones we studied. Third, there is some evidence from longitudinal research that dimensionality may decrease (along with an increase in reliability) across the study period of clinical trials for depression (Fokkema et al., 2013; Galinowski & Lehert, 1995; Quilty et al., 2013; Rocca et al., 2002; Rush et al., 2003). Of note, we are aware of only two reports that specifically tested measurement invariance (Fokkema et al., 2013; Quilty et al., 2013), while the other studies compared the reliability or factor structure across time. Finally, while most studies that did establish temporal invariance did not report whether severity of depression changed across time, levels of symptomatology were likely fairly stable due to the nature of the samples chosen in these studies (e.g., Brunet et al., 2014; Ferro & Speechley, 2013; Motl, 2005),

supporting the notion that decreases in depression severity may be related to the pattern of observations we found.

Of the studies mentioned above, we discuss the two prospective reports we consider most relevant below. In a study of 155 depressed individuals undergoing different forms of treatment (Fokkema et al., 2013), the authors identified an average recovery rate of 47% across all treatment modalities (as measured by the self-report instrument BDI). This is similar to the decreases in sum-scores by 41% (QIDS-C), 42% (HRSD), 42% (IDS-C), and 30% (IDS-SR) observed in our report. Along with this decrease, the authors identified an increase in variance of the total score, a substantial increase in reliability, and violations of measurement invariance—a pattern that closely resembles our results. In a study by Quilty et al. (2013), the authors examined the efficacy of combined pharmacotherapy and psychotherapy in a sample of 821 depressed outpatients using the MADRS. Between enrollment and month 3, the sum-score decreased by 58%, and the authors detected an increase in the variance of the sum-scores, a drastic increase of reliability, and violations of scalar and strict temporal invariance.

Overall, we conclude that the lack of invariance of the latent space may be especially pronounced in studies that track populations with substantial changes in depression scores. This may explain the inconsistencies in the prior literature regarding factor solutions, unidimensionality, and temporal invariance.

A new perspective on depression sum-scores

Consistent with the majority of the literature, our findings provide strong evidence against the common assumption that depression sum-scores adequately represent the severity of one underlying disease; the routine reflective latent variable interpretation of depression as a latent disease that is responsible for the covariation among symptoms is questionable.

This, in turn, implies that symptoms are unlikely to be *measurements* of one underlying disorder (Borsboom, 2008; Fried, 2015), which is consistent with research documenting that individual depression symptoms differ in important dimensions such as their risk factors (Fried, Nesse, Zivin, Guille, & Sen, 2014; Lux & Kendler, 2010), impact on impairment of psychosocial functioning (Fried & Nesse, 2014), antidepressant response (Hieronymus, Emilsson, Nilsson, & Eriksson, 2015), and genetic as well as neuroimaging correlates (Kendler, Aggen, & Neale, 2013; Myung et al., 2012; Webb et al., 2015) (for a review, see Fried & Nesse, 2015b). To put it differently, it seems unlikely that depression symptoms are interchangeable measurements of one depression construct due to their pronounced differences in relation to important constructs. Instead of analyzing sum-scores, putting the emphasis on the study of individual symptoms (Costello, 1993; Fried & Nesse, 2015b; Fried, 2015)—or related concepts such as endophenotypes (Hasler & Northoff, 2011; Webb et al., 2015) or dimensions defined by neurobiology and behavioral measures that cut across disorders as proposed by the RDoC framework (Cuthbert, 2014)—promise important insights. The limited reliability of single-item measures poses a significant challenge for the investigating of individual symptoms; most scales were not developed for the analysis of individual items. A potential solution is to increase the reliability of symptom assessment by measuring symptoms with multiple items, and some depression rating scales such as the Inventory of Depression and Anxiety Symptoms follow this standard psychometric approach (Watson et al., 2007); for instance, suicidal ideation is measured with 6 items, and one could construct a more reliable latent variable that accounts for the covariation of the individual suicide items.

A related problematic assumption of reflective models is that correlations among symptoms, such as the relationship between insomnia and fatigue, should vanish once the latent variable is controlled for (i.e., conditional independence) (Schmittmann et al., 2013). A common

example for conditional independence is temperature. If we spread 10 thermometers (the indicators) across a large hall and aim to measure the temperature (the latent variable), the measurements will be highly correlated because they originate from the same common cause (the reflective latent variable temperature); the correlations among thermometers are spurious and disappear once we condition on the latent variable. For depression, however, the assumption that symptom correlations are spurious is not only inconsistent with common sense (insomnia -> fatigue -> concentration problems) and residual dependencies among symptoms, but also contrasts with studies demonstrating that symptoms influence each other directly in complex dynamic systems (Borsboom & Cramer, 2013; Bringmann, Lemmens, Huibers, Borsboom, & Tuerlinckx, 2015; Fried, 2015).

These problems of the routine interpretation of sum-scores do not, however, imply that total scores are not useful, or that they should not be interpreted. In contrary, the sum of symptoms certainly does provide some information about the general psychopathological burden people carry, and we can safely assume an inverse relation between the number of symptoms and the well-being of a person (e.g., Faravelli, Servi, Arends, & Strik, 1996). In other words, one does not need a latent variable to suffer from the symptoms assessed in depression questionnaires. Furthermore, evidence shows that treatment of depression should strive to achieve full remission, often defined as scoring below a certain cut-off on a rating scales, because patients with remission have a better prognosis than those with remaining symptoms (Kennedy & Foy, 2005). However, a more promising approach may be to understand sum-scores as nothing more than the sum of a number of problems, an index similar to socioeconomic status (SES). SES is a composite of indicators such as income, job, and neighborhood. SES predicts adverse social and health outcomes in children and adolescents such as growth retardation, disability, injuries, and poverty (Bradley & Corwyn, 2002), and has been identified as the overall strongest and most

consistent predictor of both morbidity and mortality in adults (Winkleby, Jatulis, Frank, & Fortmann, 1992). As such, it is a score that carries important information, and we suggest thinking about depression total scores in a similar way. If we stay in the domain of latent variable approaches, such index scores are commonly estimated with formative models (Bollen & Lennox, 1991). Indicators like income or education for SES—or symptoms like fatigue, sadness, and insomnia for depression—contribute to a formative construct, and the sum-score merely represents an unweighted or weighted composite of this formative variable. In contrast to reflective models in which the latent variable *causes* the covariance among the indicators, the opposite is the case for formative models: the latent variable is *constructed* by the indicators. In other words, a high SES does not cause higher income, but changes in income influence the SES value for a person. For depression, it is very unlikely that depression causes depression symptoms (Fried, 2015), and depression sum-scores may be instead better understood as composite scores of psychopathological problems. This also offers an explanation why depression rating scales are often only moderately correlated and lead to idiosyncratic results depending on the particular scale used in a study (Gullion & Rush, 1998; Santor et al., 2009; Zimmerman et al., 2012). From a reflective position, they all measure the same disease, and should therefore be highly correlated. From a formative point of view, however, it is obvious that the various different symptoms used in different depression scales (Santor et al., 2009; Shafer, 2006) lead to different composite scores.

In the formative world, the main question is not how ‘real’ a construct is (Kendler, 2015)—SES, for instance, may not have biomarkers, but is an important predictor for well-being nonetheless. Much more crucial is the question how *useful* such an index score is (Kendler, Zachar, & Craver, 2010; Zachar, 2002). For mental disorders such as MD, *useful* can mean, for instance, that the sum-score provides information about the course of illness, treatment response

and success, recovery and relapse rates, potential complications, and so on. While it is much-debated whether the current operationalization of MD as described in the DSM meets orthodox standards for a useful construct (Fried, 2015; Kupfer, First, & Regier, 2002; Kupfer, 2013; Parker, 2005), a more detailed discussion of this topic is beyond the scope of this paper.

Possible explanations for the observed changes in the factor structure

Although a formative perspective provides a tentative idea on how to better interpret sum-scores, it does not answer the question what causes the observed changes in the data. Why does the factor structure change substantially? We see several possible explanations.

First, in the treatment study by Fokkema et al. (2013) that used the (self-report) BDI, the authors detected changes in the latent space similar to our findings. They interpreted these shifts as *response shift bias* (Oort, 2005). Seeing that psychological treatments of MD are often aimed at influencing patients' values or their frame of reference (Beck, Rush, Shaw, & Emery, 1979), measurement invariance violations such as changes in loadings and thresholds could reflect a shift of the participants' perception of how items and the underlying construct relate to each other. In our report, the changes of the latent space occur in both self-report and clinician-report instruments. Considering that it is not intuitive to assume that clinicians (i.e. expert raters) change their perspective on the relation between symptoms and the underlying disease, we conclude it is unlikely that response shift bias can fully explain the consistent and severe patterns of measurement invariance violations. We suggest for future studies to explore whether the rating of clinicians can be affected by participants' response shift bias.

The second group of explanations is a combination of restriction of range, selection, and regression toward the mean (RTTM). In depression studies, individuals are enrolled based on (high) symptom sum-scores. Because measurement is influenced by chance to some degree, scores at the upper end of a given scale are likely influenced in the upward direction (Barnett, van

der Pols, & Dobson, 2005). This means that the true values of high-scorers at baseline may be overestimated, and decreased scores are often observed in repeated measures due to the redistribution of chance. RTTM is known to influence the results of antidepressants trials, and, as expected, especially severe in studies with very high severity thresholds (Fava, Evins, Dorer, & Schoenfeld, 2003). We found an increase in variability of the sum-scores across time, which speaks to the possibility of a restricted range that could lead to RTTM. However, the density plots of the total scores at baseline approximate normal distributions for all rating scales, which makes restriction of range (and thus RTTM) unlikely, at least as main explanation for the severe lack of temporal invariance.

Another potential explanation is a decrease of variability of symptoms across time. If items approach a mean of zero during repeated measures, their *SD*s may decrease to a degree where they cannot exhibit pronounced correlations anymore. This could impact on the covariance matrix and thus impact on the factor structure. However, an inspection of the development of the variability of all individual symptoms revealed that *SD*s were very stable across time[3]. Moreover, we observed an *increase* of the item-intercorrelations and Cronbach's alpha, which contrasts with the explanation of decreasing item variability.

An anonymous reviewer of the manuscript pointed us towards yet another possibility: a formative-reflective continuum. Scales at baseline, when individuals express more specific items, could be more formative, while they represent more reflective constructs at later timepoints when such specific items disappear and mostly normal variation is expressed. Follow-up studies will be required to test this specific and intriguing possibility.

[3] Mean of the *SD*s of all symptoms: QIDS-C: t_1=0.89, t_2=0.87; HRSD: t_1=0.83, t_2=.0.83; IDS-C: t_1=1.00, t_2=.96; IDS-SR: t_1=0.92, t_2=0.89.

Finally, antidepressant side-effects may have impacted on the symptoms. All participants in the first treatment stage of STAR*D received citalopram for which adverse side effects are not uncommon (Bet, Hugtenburg, Penninx, & Hoogendijk, 2013), and some individuals in the NESDA sample also took antidepressants (Penninx et al., 2008). Some of the drug side-effects resemble depression symptoms such as weight change, fatigue, sleep problems, or agitation (Fried & Nesse, 2015b), which could have contributed to measurement invariance. However, a clinical trial with a similar timeframe to STAR*D in which participants received either escitalopram or nortriptyline documented that the factor structure was largely invariant across time (Uher et al., 2008).

Overall, these possibilities seem unlikely to fully explain the causes of the pronounced and consistent shifts of the factorial space observed in this report, although they may each contribute somewhat. In other words, while we have provided a thorough description of the crime scene, we have no good idea who the main suspect may be. Our hope is that future investigations of depression data—especially clinical trials or prospective stress studies (Sen et al., 2010) in which depression sum-scores decrease or increase substantially—will help to eventually identify the culprit.

Limitations

The results of this study have to be interpreted in the light of several limitations. First, the different subsamples of the STAR*D study analyzed in this report are not independent, seeing that there is a substantial overlap of participants that provided data on different scales. Furthermore, the IDS-C, IDS-SR, and QIDS-C are not independent instruments because the QIDS-C is a short version of the IDS, and the IDS-C and IDS-SR have identical item content (although one is self-rated and the other is clinician-rated). While we find consistent results across several instruments and subsamples, and while there is evidence for similar patterns in

other rating scales in clinical trials (e.g., Quilty et al., 2013; Rush et al., 1996), follow-up studies are required to examine the generalizability of our results regarding depression instruments, different populations, and psycho- vs. pharmacotherapy.

Second, as mentioned in the Methods section, the QIDS-C is usually scored into 9 items, while we coded the questionnaire into 14 items to retain information content otherwise obfuscated. It is possible that this choice led to a higher-dimensional factor structure compared to the standard scoring system (Rush et al., 2006).

Third, as described in more detail above, depression is a highly heterogeneous disease category (Fried & Nesse, 2015a; Olbert et al., 2014), and the datasets analyzed here encompass a large group of individuals with very different characteristics (in terms of, for instance, age and socioeconomic status). Studying more homogeneous samples may offer opportunities for future research.

Finally, we do not have a solid explanation for the cause of the interesting pattern of observed results (i.e. decreasing mean of the sum-score, increasing variance of the sum-score, increasing intercorrelations among items, and decrease in the number of factors).

Conclusion

From our analyses of four common rating scales of depression in two datasets over time frames ranging from 6 weeks to 2 years, we can conclude that there is a pronounced lack of unidimensionality and longitudinal measurement invariance. Given the findings reported here, we suggest that depression may be best interpreted as formative, and not as reflective, construct. The sources of the changing patterns of covariance among symptoms invite further exploration in follow-up research. Specifically, three crucial questions remain: What is the cause of this intriguing consistent pattern of observations? Is the shift in depression severity over time

responsible for the changes of the latent space, and does the inverse pattern of observations occur—increases in dimensionality, increases of the sum-score, decreases of the variability of sum-score, and decreases in rating scales reliability—in prospective studies in which depression severity increases over time? Finally, can this pattern of changes be observed only in depression, or does it generalize to other mental disorders, or even psychological constructs in general?

References

APA. (2013). *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. Washington, DC: American Psychiatric Association.

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 ? *American Journal of Psyc*, *161*(12), 2163–2177.

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–20. doi:10.1093/ije/dyh299

Bech, P., Fava, M., Trivedi, M. H., Wisniewski, S. R., & Rush, A. J. (2011). Factor structure and dimensionality of the two depression scales in STAR*D using level 1 datasets. *Journal of Affective Disorders*, *132*(3), 396–400. doi:10.1016/j.jad.2011.03.011

Beck, A. T., Rush, A. J., Shaw, F. S., & Emery, G. (1979). *Cognitive Therapy of Depression*. New York: Guilford Press.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–71.

Bet, P. M., Hugtenburg, J. G., Penninx, B. W. J. H., & Hoogendijk, W. J. G. (2013). Side effects of antidepressants during long-term use in a naturalistic setting. *European Neuropsychopharmacology*, (2013), 1–9. doi:10.1016/j.euroneuro.2013.05.001

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11).

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, *64*(9), 1089–1108. doi:10.1002/jclp

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121. doi:10.1146/annurev-clinpsy-050212-185608

Bradley, R., & Corwyn, R. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, *53*, 371–99. doi:10.1146/annurev.psych.53.100901.135233

Bringmann, L. F., Lemmens, L. H. J. M., Huibers, M. J. H., Borsboom, D., & Tuerlinckx, F. (2015). Revealing the dynamic network structure of the Beck Depression Inventory-II. *Psychological Medicine*, *45*(4), 747–57. doi:10.1017/S0033291714001809

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford

Press.

Brunet, J., Sabiston, C. M., Chaiton, M., Low, N. C. P., Contreras, G., Barnett, T. a, & O'Loughlin, J. L. (2014). Measurement invariance of the depressive symptoms scale during adolescence. *BMC Psychiatry*, *14*, 95. doi:10.1186/1471-244X-14-95

Costa, P. T., & McCrae, R. R. (1997). Longitudinal stability of adult personality. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 269–290). San Diego, US: Academic Press.

Costello, C. (1993). The advantages of the symptom approach to depression. In C. Costello (Ed.), *Symptoms of Depression* (pp. 1–21). New York: John Wiley and Sons.

Cuthbert, B. N. (2014). The RDoC framework: Facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry*, *13*(1), 28–35. doi:10.1002/wps.20087

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology*, *63*, 453–82. doi:10.1146/annurev-psych-120710-100353

Dolan, C. V., Oort, F. J., Stoel, R. D., & Wicherts, J. M. (2009). Testing Measurement Invariance in the Target Rotated Multigroup Exploratory Factor Model. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(2), 295–314. doi:10.1080/10705510902751416

Elhai, J. D., Contractor, A. A., Tamburrino, M., Fine, T. H., Prescott, M. R., Shirley, E., … Calabrese, J. R. (2012). The factor structure of major depression symptoms: A test of four competing models using the Patient Health Questionnaire-9. *Psychiatry Research*, *199*(3), 169–173. doi:10.1016/j.psychres.2012.05.018

Faravelli, C., Servi, P., Arends, J., & Strik, W. (1996). Number of symptoms, quantification, and qualification of depression. *Comprehensive Psychiatry*, *37*(October), 307–315.

Fava, M., Evins, A. E., Dorer, D. J., & Schoenfeld, D. A. (2003). The Problem of the Placebo Response in Clinical Trials for Psychiatric Disorders: Culprits, Possible Remedies, and a Novel Study Design Approach. *Psychotherapy and Psychosomatics*, *72*(3), 115–127. doi:10.1159/000069738

Ferro, M. A., & Speechley, K. N. (2013). Factor structure and longitudinal invariance of the Center for Epidemiological Studies Depression Scale (CES-D) in adult women: Application in a population-based sample of mothers of children with epilepsy. *Archives of Women's Mental Health*, *16*, 159–166. doi:10.1007/s00737-013-0331-5

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: an illustration of longitudinal measurement invariance. *Psychological Assessment*, *25*(2), 520–31. doi:10.1037/a0031669

Fried, E. I. (2015). Problematic assumptions have slowed down depression research: why symptoms, not syndromes are the way forward. *Frontiers in Psychology*, *6*(306), 1–11. doi:10.3389/fpsyg.2015.00309

Fried, E. I., & Nesse, R. M. (2014). The Impact of Individual Depressive Symptoms on Impairment of Psychosocial Functioning. *PLoS ONE*, *9*(2), e90311. doi:10.1371/journal.pone.0090311

Fried, E. I., & Nesse, R. M. (2015a). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *Journal of Affective Disorders*, *172*, 96–102. doi:10.1016/j.jad.2014.10.010

Fried, E. I., & Nesse, R. M. (2015b). Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(72), 1–11. doi:10.1186/s12916-015-0325-4

Fried, E. I., Nesse, R. M., Zivin, K., Guille, C., & Sen, S. (2014). Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. *Psychological Medicine*, *44*, 2067–2076. doi:10.1017/S0033291713002900

Galinowski, A., & Lehert, P. (1995). Structural validity of MADRS during antidepressant treatment. *International Clinical Psychopharmacology*, *10*(3), 157–161. doi:10.1097/00004850-199510030-00004

Gullion, C. M., & Rush, A. J. (1998). Toward a generalizable model of symptoms in major depressive disorder. *Biological Psychiatry*, *44*(10), 959–72.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, *23*, 56–62.

Hasler, G., & Northoff, G. (2011). Discovering imaging endophenotypes for major depression. *Molecular Psychiatry*, *16*(6), 604–19. doi:10.1038/mp.2011.23

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. *Molecular Psychiatry*, 1–8. doi:10.1038/mp.2015.53

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. doi:10.1080/10705519909540118

Kendler, K. S. (2015). Toward a limited realism for psychiatric nosology based on the coherence theory of truth. *Psychological Medicine*, *45*(06), 1115–1118. doi:10.1017/S0033291714002177

Kendler, K. S., Aggen, S. H., & Neale, M. C. (2013). Evidence for Multiple Genetic Factors

Underlying DSM-IV Criteria for Major Depression. *American Journal of Psychiatry, 70(6)*, 599–607. doi:10.1001/jamapsychiatry.2013.751

Kendler, K. S., Zachar, P., & Craver, C. (2010). What kinds of things are psychiatric disorders? *Psychological Medicine*, *41*(06), 1143–1150. doi:10.1017/S0033291710001844

Kennedy, N., & Foy, K. (2005). The impact of residual symptoms on outcome of major depression. *Current Psychiatry Reports*, *7*(6), 441–446. doi:10.1007/s11920-005-0065-9

Kline, R. (2005). *Principles and Practice of Structural Equation Modeling*. New York, NY: Guildford.

Kupfer, D. J. (2013). Field Trial Results Guide DSM Recommendations. *Huffington Post*. Retrieved from http://www.huffingtonpost.com/david-j-kupfer-md/dsm-5_b_2083092.html

Kupfer, D. J., First, M. B., & Regier, D. A. (2002). *A Research Agenda For DSM V*. Washington, DC: American Psychiatric Association.

Lei, H., Yao, S., Zhang, X., Cai, L., Wu, W., Yang, Y., … Zhu, X. (2014). Longitudinal Invariance of the Children's Depression Inventory for Urban Children in Hunan, China. *European Journal of Psychological Assessment 2014*, 1–10. doi:10.1027/1015-5759/a000195

Lux, V., & Kendler, K. S. (2010). Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria. *Psychological Medicine*, *40*(10), 1679–90. doi:10.1017/S0033291709992157

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: an integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*(Mimic), 85–110. doi:10.1146/annurev-clinpsy-032813-153700

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Taylor and Francis Group.

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, *134*(4), 382–389. doi:10.1192/bjp.134.4.382

Motl, R. W. (2005). Longitudinal Invariance of the Center for Epidemiologic Studies-Depression Scale among Girls and Boys in Middle School. *Educational and Psychological Measurement*, *65*(1), 90–108. doi:10.1177/0013164404266256

Muthén, B. O., & Muthén, L. (2012). *Mplus User's Guide, seventh edition*. Los Angeles: Muthén

& Muthén.

Myung, W., Song, J., Lim, S.-W., Won, H.-H., Kim, S., Lee, Y., … Kim, D. K. (2012). Genetic association study of individual symptoms in depression. *Psychiatry Research, 198(3)*, 400–6. doi:10.1016/j.psychres.2011.12.037

O'Connor, B. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32*, 396–402.

Olbert, C. M., Gala, G. J., & Tupler, L. A. (2014). Quantifying Heterogeneity Attributable to Polythetic Diagnostic Criteria : Theoretical Framework and Empirical Application. *Journal of Abnormal Psychology*, *123*(2), 452–462. doi:10.1037/a0036068

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14*, 587–598.

Parker, G. (2005). Beyond major depression. *Psychological Medicine*, *35*(4), 467–74.

Penninx, B. W. J. H., Beekman, A. T. F., Smit, J. H., Zitman, F. G., Nolen, W. A., Spinhoven, P., … Van Dyck, R. (2008). The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *International Journal of Methods in Psychiatric Research*, *17*(3), 121–40. doi:10.1002/mpr.256

Preskorn, S. H., Macaluso, M., & Trivedi, M. (2015). How Commonly Used Inclusion and Exclusion Criteria in Antidepressant Registration Trials Affect Study Enrollment. *Journal of Psychiatric Practice*, *21*(4), 267–274. doi:10.1097/PRA.0000000000000082

Quilty, L. C., Robinson, J. J., Rolland, J.-P., De Fruyt, F., Rouillon, F., & Bagby, R. M. (2013). The structure of the Montgomery–Åsberg depression rating scale over the course of treatment for depression. *International Journal of Methods in Psychiatric Research*, *22*(3), 175–184. doi:10.1002/mpr

R Development Core Team. (2008). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Radloff, L. S. (1977). The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*, *1*(3), 385–401. doi:10.1177/014662167700100306

Revelle, W. (2015). psych: Procedures for Psychological, Psychometric, and Personality Research.

Rocca, P., Fonzo, V., Ravizza, L., Rocca, G., Scotta, M., Zanalda, E., & Bogetto, F. (2002). A comparison of paroxetine and amisulpride in the treatment of dysthymic disorder. *Journal of Affective Disorders*, *70*(3), 313–317. doi:10.1016/S0165-0327(01)00327-5

Rush, A. J., Bernstein, I. H., Trivedi, M. H., Carmody, T. J., Wisniewski, S., Mundt, J. C., … Fava, M. (2006). An Evaluation of the Quick Inventory of Depressive Symptomatology and the Hamilton Rating Scale for Depression: A Sequenced Treatment Alternatives to Relieve Depression Trial Report. *Biological Psychiatry*, *59*(6), 493–501. doi:10.1016/j.biopsych.2005.08.022

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., … Niederehe, G. (2004). Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Controlled Clinical Trials*, *25*(1), 119–142. doi:10.1016/S0197-2456(03)00112-0

Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B., & Trivedi, M. H. (1996). The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychological Medicine*, *26*(3), 477–86.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., … 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. *Biological Psychiatry*, *54(5)*, 573–583. doi:10.1016/S0006-3223(03)01866-8

Santor, D. A., Gregus, M., & Welch, A. (2009). Eight Decades of Measurement in Depression. *Measurement*, *4*(3), 135–155. doi:10.1207/s15366359mea0403

Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, *31*(1), 43–53. doi:10.1016/j.newideapsych.2011.02.007

Sen, S., Kranzler, H. R., Krystal, J. H., Speller, H., Chan, G., Gelernter, J., & Guille, C. (2010). A prospective cohort study investigating factors associated with depression during medical internship. *Archives of General Psychiatry*, *67*(6), 557–65. doi:10.1001/archgenpsychiatry.2010.41

Shafer, A. B. (2006). Meta-analysis of the Factor Structures of Four Depression Questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology*, *62*(1), 123–146. doi:10.1002/jclp

Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach. *Psychometrika*, *74*(1), 107–120. doi:10.1007/s11336-008-9101-0

Trivedi, M. H., Rush, A. J., Ibrahim, H. M., Carmody, T. J., Biggs, M. M., Suppes, T., … Kashner, T. M. (2004). The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive

Symptomatology, Clinician Rating (QIDS-C) and Self-Report (QIDS-SR) in public sector patients with mood disorders: a psych. *Psychological Medicine*, *34*(1), 73–82. doi:10.1017/S0033291703001107

Uher, R., Farmer, A., Maier, W., Rietschel, M., Hauser, J., Marusic, A., … Aitchison, K. J. (2008). Measuring depression: comparison and integration of three scales in the GENDEP study. *Psychological Medicine*, *38*(2), 289–300. doi:10.1017/S0033291707001730

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4–70. doi:10.1177/109442810031002

Wardenaar, K. J., van Veen, T., Giltay, E. J., den Hollander-Gijsman, M., Penninx, B. W. J. H., & Zitman, F. G. (2010). The structure and dimensionality of the Inventory of Depressive Symptomatology Self Report (IDS-SR) in patients with depressive disorders and healthy controls. *Journal of Affective Disorders*, *125*(1-3), 146–54. doi:10.1016/j.jad.2009.12.020

Watson, D., O'Hara, M. W., Simms, L. J., Kotov, R., Chmielewski, M., McDade-Montez, E. a, … Stuart, S. (2007). Development and validation of the Inventory of Depression and Anxiety Symptoms (IDAS). *Psychological Assessment*, *19*(3), 253–268. doi:10.1037/1040-3590.19.3.253

Webb, C. A., Dillon, D. G., Pechtel, P., Goer, F., Murray, L., Huys, Q. J. M., … Pizzagalli, D. A. (2015). *Neural Correlates of Three Promising Endophenotypes of Depression: Evidence from the EMBARC Study*. *Neuropsychopharmacology*. Nature Publishing Group. doi:10.1038/npp.2015.165

Wetherell, J. L., Gatz, M., & Pedersen, N. L. (2001). A longitudinal analysis of anxiety and depressive symptoms. *Psychology and Aging*, *16*(2), 187–195. doi:10.1037//0882-7974.16.2.187

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial Invariance within Longitudinal Structural Equation Models: Measuring the Same Construct across Time. *Child Development Perspectives*, *4*(1), 10–18. doi:10.1111/j.1750-8606.2009.00110.x.Factorial

Winkleby, M. A., Jatulis, D. E., Frank, E., & Fortmann, S. P. (1992). Socioeconomic status and health: How education, income, and occupation contribute to risk factors for cardiovascular disease. *American Journal of Public Health*, *82*(6), 816–820. doi:10.2105/AJPH.82.6.816

Wisniewski, S. R., Rush, A. J., Nierenberg, A. A., & Gaynes, B. N. (2009). Can Phase III Trial Results of Antidepressant Medications Be Generalized to Clinical Practice? A STAR*D Report. *American Journal of Psychiatry*, *166*(5), 599–607. doi:10.1176/appi.ajp.2008.08071027

Zachar, P. (2002). The Practical Kinds Model as a Pragmatist Theory of Classification.

Philosophy, Psychiatry, & Psychology, *9*(3), 219–227. doi:10.1353/ppp.2003.0051

Zimmerman, M., Ellison, W., Young, D., Chelminski, I., & Dalrymple, K. (2014). How many different ways do patients meet the diagnostic criteria for major depressive disorder? *Comprehensive Psychiatry*. doi:10.1016/j.comppsych.2014.09.007

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? *The Journal of Clinical Psychiatry*, *73*(10), 1287–91. doi:10.4088/JCP.12m07775