Journal of Affective Disorders 222 (2017) 204–210

Contents lists available at ScienceDirect

Journal of Affective Disorders

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

Research paper

Symptom severity and disability in psychiatric disorders: The U.S. Collaborative Psychiatric Epidemiology Survey

R. García-Velázquez[a,*], M. Jokela[a], T.H. Rosenström[a,b]

[a] Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Finland
[b] Department of Mental Disorders, Norwegian Institute of Public Health, Oslo, Norway

ARTICLE INFO

Keywords:
Diagnostic definitions
Disability
Validity
Affective disorders
Anxiety disorders
Item Response Theory

ABSTRACT

Background: While most psychiatric diagnoses are based on simple counts of symptoms, some symptoms may be sign of a more severe mental syndrome than others. This calls for validated estimates of the relative severity specific symptoms imply within a disorder. We focused on four diagnostic disorders: Manic Episode (ME), Major Depressive Episode (MDE), Post-traumatic Stress Disorder (PTSD) and Generalized Anxiety Disorder (GAD). Symptom-specific severity parameters were estimated, and validated by examining their association with levels of self-reported disability in daily activities over and above the number of symptoms.

Methods: Data from the cohort study of the U.S. Collaborative Psychiatric Epidemiology Surveys (CPES) was used, which comprises the National Comorbidity Survey Replication, National Survey of American Life, and the National Latino and Asian American Study. The four analytic datasets included respondents who endorsed disorder-specific pre-screening symptoms according to the World Mental Health Survey Initiative's version of the Composite International Diagnostic Interview. Disability was measured using the WHO Disability Assessment Schedule. Item Response Theory and Tobit models were implemented.

Results: For ME, PTSD, and GAD (not MDE) symptom severity based on psychometric Item Response Theory predicted disability outcomes after adjusting for symptom count. For PTSD, symptom count was not associated with disability.

Limitations: The analytic sample for each psychiatric disorder was based on a pre-selection stemming from index criteria (e.g. sadness or pleasure loss for MDE), which implies that our results are only generalizable to those individuals at risk rather than for the entire population. Additionally, we acknowledge that the use of unidimensional models is only one of the several options to model psychopathological constructs.

Conclusions: The same number of symptoms may be related to different levels of disability, depending on the specific symptoms from which the person suffers. Diagnostic procedures and treatment decisions may benefit from such additional information without extra costs.

1. Introduction

Diagnostics of psychiatric disorders are based on the number of symptoms as defined in diagnostic manuals. In most instances, the diagnosis does not discriminate between the specific symptoms that the person suffers from. That is, two people with the same diagnosis may have only partially overlapping sets of symptoms. However, the different symptom combinations may contain information on the clinical severity of the person's disorder, over and above the simple number of symptoms (Fried and Nesse, 2015a). Such knowledge could be used for more informative diagnosis of psychiatric disorders.

According to classic psychometrics, *construct validity* pertains to the effective measurement of the intended theoretical construct (here, a specific mental disorder) and to appropriate inferences based on it (Cronbach and Meehl, 1955; Loevinger, 1957). The *structural* component of construct validity implies that the symptoms indexing the construct occur and coexist consistently as its representatives. The *external* component of construct validity implies that the measure is associated with relevant external criteria not directly used to measure the construct, that is, with independently measured outcomes.

Mental disorders have been conceptualized as *harmful dysfunctions* (Wakefield, 1992), and for most psychiatric diagnoses of the DSM, the disorder must be related to *clinically significant disturbance, distress, or disability* (American Psychiatric Association, 2013). Thus, the purpose of a valid psychiatric diagnosis is to identify functionally impaired individuals as opposed to non-impaired individuals. Such a purpose also

* Correspondence to: Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Haartmaninkatu 3, Room E215, P.O. Box 9, 00014, Finland.
E-mail address: regina.garciavelazquez@helsinki.fi (R. García-Velázquez).

http://dx.doi.org/10.1016/j.jad.2017.07.015
Received 2 May 2017; Received in revised form 1 July 2017; Accepted 6 July 2017
Available online 08 July 2017
0165-0327/

establishes impairment as a fundamental criterion in external validity studies. If specific symptoms carry information about the clinical impairment associated with the disorder over and above the simple count of symptoms, it would be possible to use this information to sharpen diagnosis and more effectively orient therapeutic targets. Symptom-level analyses of psychiatric disorders suggest that specific symptoms may be differently related to psychosocial functioning (Fried and Nesse, 2014, 2015a; Fried et al., 2015; Tweed, 1993; Wakefield and Schmitz, 2017a, 2017b). However, the performance of specific diagnostic criteria with respect to clinical measures of disability has not been examined in detail across the most disabling psychiatric disorders, such as Major Depression or Post-traumatic Stress Disorder.

Continuous dimensions, rather than categories, seem to best explain the structure of psychiatric symptoms, suggesting that disorders reflect underlying continuous variables (Edens et al., 2006; Haslam et al., 2012; Marcus et al., 2006, 2008). Models of Item Response Theory (IRT; de Ayala, 2013; Reckase, 2009) link each symptom to a threshold on an underlying disorder continuum based on the distribution of symptom occurrence. In modeling psychiatric disorders, the *psychometric severity* parameter (known as *difficulty* parameter in ability testing, or *b* parameter in general) determines the level of the underlying disorder continuum that is required for an individual to endorse a symptom with a 50% probability (Reise and Waller, 2009; Thomas, 2011). After determining the psychometric severity of specific symptoms, it is possible to examine whether these severity estimates are related to external indicators of functional disability, including difficulties in self-care, life activities or cognition (Üstün, 2010). Fig. 1 displays the rationale of our approach.

Using data from three large studies of psychiatric epidemiology (Alegria et al., 2015; Pennell et al., 2004), we estimated psychometric symptom severity using IRT models and examined their associations with the external validity criterion of self-reported disability, controlling for symptom count. We focused on four disorders – Manic Episode (ME), Major Depressive Episode (MDE), Post-traumatic Stress Disorder (PTSD) and Generalized Anxiety Disorder (GAD)—which constitute the four most impairing disorders among those examined by Druss et al. (2009) as rated by the Sheehan Disability Scale.

2. Methods

2.1. Data

The Collaborative Psychiatric Epidemiological Surveys (CPES) comprised three multi-stage area probability samples conducted between 2001 and 2003: the National Comorbidity Survey Replication (NCS-R, N = 9282), the National Survey of American Life (NSAL, N = 6082), and the National Latino and Asian American Study (NLAAS, N = 4649).

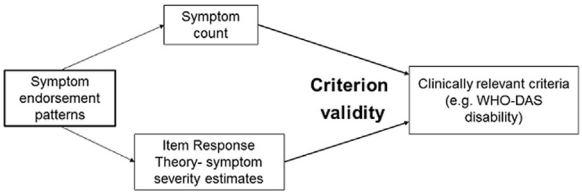

Fig. 1. Validation of Item Response Theory-derived symptom severity as referred to a criterion. Footnote: Once a set of symptoms has been defined (e.g. according to DSM-5), symptom endorsement data can be used to (a) derive symptom counts, which generally determine whether a diagnosis is fulfilled (e.g. Generalized Anxiety Disorder requires compulsory criterion plus three associated symptoms), and (b) obtain information on symptom characteristics (i.e. Item Response Theory-based severity). Symptom count and severity estimates can be studied in association with a clinically relevant variable such as disability (criterion validity). Using data at the symptom-level maximizes the use of information and brings insights on single symptom functioning. The utility of Item Response Theory and the validity of its assumptions is probed by the ability of latent severity estimates to predict the criterion variable over and above symptom count.

The data was available to us via the Inter-university Consortium for Political and Social Research service (Alegria et al., 2015). The joint sample is representative of adults (i.e., age 18 years or more) residing in households in coterminous United States, Alaska and Hawaii, excluding institutionalized persons and those living on military bases (NCS-R and NSAL also excluded non-English speakers). Comparable methodologies were used, including using trained lay interviewers to conduct interviews primarily in person. The average response rate of the CPES is 72.7%. Details of each survey can be found elsewhere (Heeringa et al., 2004; Pennell et al., 2004). The final CPES sample includes 20,013 individuals (8550 men and 11,463 women).

From the overall dataset, a different subsample was extracted for each disorder (GAD, ME, MDE, PTSD) containing individuals who answered to the specific diagnostic interview after having endorsed prescreening symptoms. Our samples were a total of N = 4214 for ME, N = 4152 for MDE, N = 3128 for PTSD, and N = 3610 for GAD.

2.2. Measures

2.2.1. Psychiatric symptoms

The respondents were interviewed according to the World Mental Health Survey Initiative's version of the Composite International Diagnostic Interview (WMH-CIDI), which is a modified version of the original WHO-CIDI (Kessler and Üstün, 2004). Both WMH-CIDI and the other CPES questions were administered using a computer-assisted interview. The presence of a symptom was determined exactly as in the DSM-IV part of the ICPSR documentation for the diagnostic algorithms (Alegria et al., 2015).

2.2.2. Functional impairment

The World Health Organization Disability Assessment Schedule (WHODAS) was used to measure disability as defined in the International Classification of Function, Disease and Health. Scores are a product of frequency and severity of problems (none, mild, moderate, severe) that respondents reported experiencing in the past 30 days, and are normalized to values ranging from 0 to 100, where higher numbers indicate worse functioning. We used a single overall score based on the average disability over the domains of functioning: cognition, mobility, self-care, social interaction, role functioning, participation (Üstün, 2010).

As another indicator of validity based on clinical relevance, we used the item "Beginning yesterday and going back 30 days, how many days out of the past 30 were you totally unable to work or carry out your normal activities because of problems with either your physical health, your mental health, or your use of alcohol or drugs?" as a variable called "disability days" from now on. Due to highly similar findings, results using disability days are shown only in the Supplement. Both outcomes were assessed at a general level, and thus were not disorder-specific.

2.3. Statistical analysis

2.3.1. Item Response Theory for estimating psychometric severity

For each of the four disorders, a series of unidimensional IRT models were estimated. This procedure was chosen because the disorders are typically treated as single diagnostic syndromes, implying unidimensionality. Each set of diagnostic criteria was analyzed for *essential unidimensionality* to prevent interpretation of overly biased models (Reise and Rodriguez, 2016). Since diagnostic definitions do not make a distinction in terms of background variables, we chose to keep the paradigm as comparable as possible and hence estimated a general severity parameter per criterion (i.e. no gender- or age-stratified analyses were conducted).

The diagnostic criteria composed of several single symptoms were analyzed by splitting them (for instance MDE A.4 *change in sleep,* which contains *insomnia or hypersomnia*). Those criteria that were necessary

for the other criteria to be assessed in the diagnostic algorithm were not included because they defined the sample and therefore had no variation. MDE criteria present an exceptional case due to strong negative correlations among paired poles of symptoms (weight gain-loss, hypersomia-insomnia, motor retardation-agitation), and thus a two-tier bifactor model was fit (Cai, 2010a), in which only the parameters corresponding to the general factor (i.e. those capturing the common variation) were used.

All IRT models were estimated using the Metropolis-Hastings Robbins-Monro algorithm (Cai, 2010b). Models were initially estimated in the slope-intercept parameterization due to computational reasons, and the fact that it is the only parameterization in multidimensional IRT (Chalmers, 2012). Then, for the sake of an easier interpretation, the parameters were transformed into the traditional IRT parameterization (slope-threshold). All the necessary information about IRT models and converting parameters from slope-intercept parameterization to the traditional form is included in Supplement 1, and is further elaborated by de Ayala (2013) and Reckase (2009).

The most useful models were selected based on Bayesian Information Criterion (BIC) and likelihood ratio test (LRT). BIC decreases as model fit improves, and penalizes model complexity. We also inspected precision in terms of parameter standard error. In addition to model comparison, goodness of fit (GOF) of the models we retained was assessed with the following indices: M_2 (measure of fit distributed according to χ^2, specifically developed for IRT models; Maydeu-Olivares, 2013), RMSEA (based on M_2; acceptable values below .08, close fit $\leq$.05), SRMSR (acceptable $\leq$.1, good $\leq$.08), and CFI (acceptable $\geq$ .90, good $\geq$ .95) according to general thresholds (Kenny, 2015; Kline, 2015). Notice, however, that caution has been called for regarding the use of the above rules of thumb when interpreting these GOF indices in the context in IRT (Cai and Hansen, 2013; Kessler and Üstün, 2004)

Because the present work relies on the assumption that different symptoms imply different levels of severity, we verified that all the *d* parameters (slope-intercept form) could not be constrained as equal using LRT. Next, *b* parameters (traditional form) and their confidence intervals were estimated implementing the Delta Method (see Supplement 2).

2.3.2. Tobit regression models for predicting disability

The validity of psychometric severity estimates was tested based on the following hypothesis: assuming the symptoms map on a 'true' continuum of severity, one would expect that for two people with equal number of endorsed symptoms, the one with more "severe" symptoms would display more disability. Thus, regression models were fit in which WHODAS score was predicted by the maximum severity of the symptoms endorsed, controlling for number of symptoms, age, gender and ethnic background.

Because of the high frequency of floor values of the WHODAS variable we used censored regression model, a generalization of the Tobit model (Tobin, 1958). It assumes normal distribution underlying the *disability* outcome variable, modeling it as being limited to the range of WHODAS (i.e. disability ranges from 0 to 100). We chose heteroscedastic modeling due to WHODAS variability according to symptom count. Here, β_{maxsev} denotes the regression coefficient from regressing disability onto the maximum severity of the criteria endorsed by the participant, adjusted for symptom count. It informs of the predictive value of the psychometric severity after the symptom count has been ruled out. Two scenarios are possible: (1): β_{maxsev} differs statistically from 0, indicating that IRT-estimated severities predict disability independently of the symptom count, and hence showing evidence of criterion validity; or (2) β_{maxsev} is not significant, indicating that the assumed latent dimension has non-significant predictive value with respect to disability.

The maximum endorsed severity was chosen as a proxy for disorder severity for several reasons. First, it is not diluted due to endorsing milder symptoms, as could be the case when averaging severity (Fayers and Hand, 2002). As IRT assumes monotonically increasing relation between the latent variable and the probability of endorsing an item, individuals are likely to endorse multiple low-severity symptoms below their corresponding attribute level (θ, see Supplement 1). Second, it is independent from symptom count, unlike the IRT estimated θ, which can be problematic to use in variance decomposition (Berg et al., 2007). Third, IRT symptom estimates are not as affected by non-normality of the trait as θ estimates (Sass et al., 2008). Fourth, the maximum severity index is a simple measure of the symptom-specific burden a person might experience and captures the possibility that symptoms are not interchangeable indicators of a given construct. If informative about disability, "worst symptom" (i.e. most severe) offers a concise clinical variable supplementing the usual symptoms counts.

2.3.3. Simulation as a reference

A simulation study was conducted in order to ensure the sensitivity of our procedure. We generated datasets where trait level (i.e. θ) was a true predictor of disability, and estimated the statistical power for a *correct positive* result under the concrete parameters of our data (same sample size and symptom parameters). The simulation results serve as a reference on how our procedure would behave under known conditions. For detailed information and results, see Supplement 3.

All statistical analyses were conducted in the R environment version 3.3.1 (R Core Team, 2017) and the packages *psych* 1.6.8 (Revelle, 2015), *mirt* 1.21 (Chalmers, 2012), and *crch* 1.0-0 (Messner et al., 2014). The multi-stage sampling weights were not used.

3. Results

Following the criteria in the Methods section, two-parameter logistic (2PL) IRT models for all disorders were retained. The GOF for the retained models is presented in Table 1. All models satisfy adequately or well fit standards, with the exception of Manic Episode, for which GOF was found to be overall acceptable with marginally inadequate RMSEA. Sample characteristics, and symptom endorsement and parameter estimates of the retained models are shown in Supplement 4 and 5. The GOF of all the models we tested can be found in Supplement 6. Average correlation between WHODAS and symptom count was r = .23. WHODAS and maximum severity correlated on average r = .15.

For all the disorders, the free-intercept models fit statistically better than the constrained counterparts, indicating that the distribution of the symptoms entails different levels of psychometric severity ($p < .001$, and lower BIC values; Supplement 7). Because severity varies across symptoms, it makes sense to analyze its association with external indicators of disability.

Left column of Fig. 2 displays average WHODAS score plotted against the IRT severity estimate of the endorsed symptom. The right column shows the association between symptom count and average WHODAS (see Supplement 8 for the disability days instead of WHODAS; the two disability indices correlated at .77). The MDE items

Table 1
Model fit for the selected Item Response Theory models.

Model	M_2	df	p	RMSEA (95% CI)	SRMSR	CFI
Manic Episode, 2PL	3247.675	104	< .001	.085 [.082, .088]	.101	.932
Major Depressive Episode, bifactor 2PL	1166.310	84	< .001	.056 [.052, .059]	.048	.937
Post-traumatic Stress Disorder, 2PL	1949.890	119	< .001	.07 [.067, .073]	.057	.971
Generalized Anxiety Disorder, 2PL	200.660	20	< .001	.05 [.044, .056]	.039	.934

2PL: two-parameter logistic Item Response Theory model. 95% CI: 95% Confidence Interval.

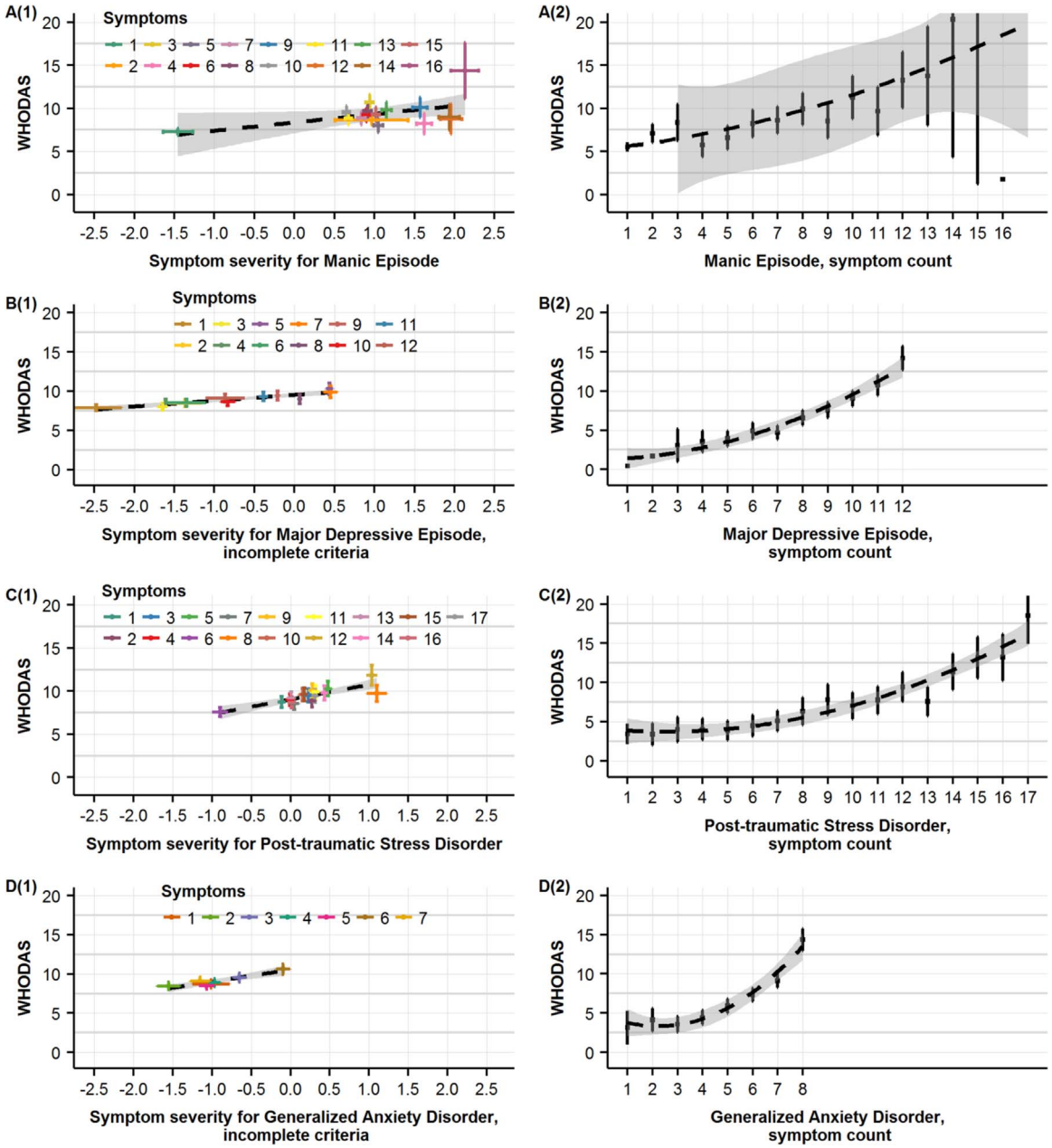

Fig. 2. Criteria of the DSM diagnostic definitions and associated WHODAS-disability score. Footnote: Whiskers denote 95% confidence intervals. The values in the legends correspond to diagnostic criteria which are identified and numbered in Supplement 5. A: Manic Episode, B: Major Depressive Episode (excluding weight gain, hypersomnia, and motor agitation), C: Post-traumatic Stress Disorder, D: Generalized Anxiety Disorder (excluding uncontrollable worry). Left column (1): Item Response Theory severity estimates on the x-axis. Smaller severity estimates correspond with higher endorsement frequencies after having endorsed the screening items for the disorder, and vice versa. Right column (2): Raw symptom count on the x-axis. The highest symptom counts had a small sample size and have been omitted in the case of Manic Episode.

weight gain, *hypersomnia*, and *motor agitation* are not presented because most of their variance was explained by their corresponding specific factors and their loadings onto the general depression factor were extremely low (λ_G = .00, − .13, and .05), meaning that they were unrelated to the modeled latent MDE dimension. The GAD symptom *uncontrollable worry* had extremely low severity (b = − 6.31) and hence was not plotted. Severity estimates in IRT indicate the extent of latent trait (i.e. underlying disorder) required in order to endorse an item with .50 probability.

Next, a series of Tobit regression models were fit for each mental disorder separately (Table 2). Quadratic terms were included for symptom count as they were visually evident in Fig. 2, and therefore the symptom count variable was centered. Maximum severity predicted disability for ME, PTSD, and GAD. It showed a positive sign meaning that the higher severity score, the higher disability on average. Symptom count predicted disability either in a linear or quadratic trend for ME, MDE, and GAD (but not for PTSD).

4. Discussion

This is, to our knowledge, the first study to use IRT models to determine the *psychometric severity* of specific psychiatric diagnostic criteria and jointly examine how these estimates are associated with clinically relevant external criteria of disability. We observed that symptoms did vary in severity as determined by their IRT parameters, and that symptom severity was associated with external indicators of psychosocial and functional impairment in a representative community sample of U.S. This suggests that the IRT severity parameters are not

Table 2
Tobit regression coefficients for predicting disability score based on symptom count and maximum severity.

Disorder	Term	Estimate	Std. error	Z	Pr (> |Z|)
Manic Episode	max. severity	1.253	.369	3.39	< .001***
	symptom count	− .158	.386	− .408	.683
	symptom count^2	.075	.031	2.42	.016*
Major Depressive Episode	max. severity	− .565	.838	− .67	.501
	symptom count	1.682	.856	1.96	< .049*
	symptom count^2	.003	.052	.06	.954
Post-traumatic Stress Disorder	max. severity	2.039	1.010	2.02	. 044*
	symptom count	.523	.419	1.25	.212
	symptom count^2	.0400	.022	1.85	.064
Generalized Anxiety Disorder	max. severity	2.529	.658	3.84	< .001***
	symptom count	− 2.033	1.120	− 1.82	< .001***
	symptom count^2	.386	.104	3.72	< .070

The estimates are regression coefficients of four separate heteroscedasticity-robust Tobit regression models. The models were adjusted for gender, age and ethnic background. Significance codes: '***' < .001, '**' ≤ .01, '*' ≤ .05.

purely structural psychometric indicators but may have clinical relevance in determining disability associated with specific diagnostic symptoms.

This work contributes to a growing number of studies suggesting that specific symptoms function differently within diagnostic categories (Fried, 2015; Fried and Nesse, 2014, 2015a, 2015b; Fried et al., 2015; Tweed, 1993; Wakefield and Schmitz, 2017b). Our results further imply that different symptoms of the same psychiatric disorder cannot be taken as interchangeable indicators of the underlying disability associated with the diagnosis, and thereby support the use of symptom-level information in psychopathology. While the heterogeneity of many psychiatric disorders is widely acknowledged, very few studies have suggested ways to quantify and incorporate such heterogeneity into diagnoses. Other studies focusing on disability have, instead, contributed to the nosological debate by suggesting empirically-driven changes to the set of symptoms defining MDE (Rosenström and Jokela, 2017), or by describing novel symptom profiles that combine anxiety and depression syndromes (Wanders et al., 2016). One contribution of this work is to illustrate that the IRT-estimated severity could be used as a parallel diagnostic feature that helps clinicians working strictly within DSM-5 framework in prioritizing interventions to most disabled individuals and most disabling symptoms. For instance, *feelings of worthlessness* and *psychomotor retardation* have been considered as criteria for *complicated* depression (Wakefield et al., 2007). *Feelings of worthlessness* also predict concurrent and post-remission suicide attempts (Bolton et al., 2008; Wakefield and Schmitz, 2015). These were the two most severe symptoms in our study within the MDE criteria set (see Supplement 5).

Symptoms with mild severity, in turn, could be revised to achieve more effective indicators of dysfunction (First and Wakefield, 2013). For example, it has been suggested that including *sadness* might inflate diagnostic decisions towards false positives of depression (Horwitz and Wakefield, 2007; Rosenström and Jokela, 2017; Wakefield, 2015; Wakefield and Demazeux, 2016; Wakefield and First, 2013). The discussion about tuning GAD requirements of worry being *uncontrollable* and *excessive*, so that they indeed capture pathologic worry, is another example (Andrews et al., 2010; Ruscio et al., 2005). Such criteria were also among the least psychometrically severe symptoms in our study. It has also been found that pathological worry registered a very low inter-rater agreement, and that other criteria predict better GAD persistence on a six-month follow-up (Wittchen et al., 1995). In this sense, the low severity is coherent with previous empirical findings that question the ambiguous nature of these symptoms as they are defined in DSM.

While IRT may have its limitations in psychopathology research (Borsboom et al., 2016), in our present approach it had merits in reflecting psychosocial disability over and above the symptom counts for GAD, PTSD, and ME. MDE registered rather limited statistical power according to the results of our simulation (Supplement 3), and hence effects were unlikely to show. Overall, the congruent trend between severity estimates and external criteria of impairment (WHODAS and disability days, in Supplement 8), suggests that IRT can provide additional and clinically valid information over and above the standard approaches. Furthermore, in the PTSD model symptom count did not predict disability while maximum severity did.

A second characteristic that is provided by IRT is symptom discrimination. Additional to severity, gathering knowledge on discrimination would also tell more of a case than the mere amount of symptoms. For instance, a symptom with high discrimination may be useful for predicting a case with higher impairment, if the symptom discriminates well among high and medium-low trait levels and its severity is well established. This knowledge may assist clinicians and it is material for future studies.

In addition to their applied value, our findings may contribute to the ongoing debate on psychiatric nosology (Borsboom and Cramer, 2013; Fried et al., 2016; Krueger et al., 2016; Schmittmann et al., 2013; Stein et al., 2013), since the severity index of IRT is mathematically related to "threshold" parameter of (Ising) network models found in the recently introduced "network theory of psychopathology" (Borsboom, 2017; Epskamp et al., 2016; Marsman et al., 2015). Symptoms with high IRT severity correspond with a low Ising *threshold parameter*, and vice versa (Cramer et al., 2016; Marsman et al., 2015). The suitability of IRT-derived severity parameters is also straightforward in the framework of the Hierarchical Taxonomy of Psychopathology model (Kotov et al., 2017). For instance, it would be possible to profit from multidimensional IRT models by examining how severity and discrimination parameters of the different symptoms perform within comorbidity models (e.g. anxiety and depression modeled as internalizing spectrum into a two-trait MIRT).

5. Limitations

We recognize certain limitations in this work. First, the interview design was based on screening items by which participants were filtered into further specific sections, which may have biased the estimates towards a milder severity. This entails that our results are suited for those at risk rather than for the entire population. Because pre-selection based on index problems is so prevalent in epidemiology and clinical practice, it would be important to systematically compare psychometric structural estimates in the selected sub-groups versus the total population. Especially sub-groups selected based on sum-scores tend to have different psychometric structure in comparison to the total population (Muthén, 1989). With regard to generalizability, we acknowledge that the use of a dataset such as CPES is a strength, which at the same time comes with the cost of being specific of the U.S. population. Therefore, replication remains an issue for further work.

Second, unidimensional models are not necessarily the most

suitable option for all mental disorders. For instance, recent studies on PTSD have reported a seven-dimensional structure (Pietrzak et al., 2015) or considered PTSD to be formed by a network of inter-related symptoms rather than symptoms reflecting an underlying uniform disease (McNally et al., 2015). In this sense, the marginal goodness of fit registered for ME could be indeed due to underlying dimensionality. However, unidimensional models for multidimensional data are relatively robust when multidimensionality is due to multiple latent dimensions that are moderately correlated, or if there is a strong general factor (Drasgow and Parsons, 1983; Harrison, 1986; Ip, 2010; Reise et al., 2014). Additionally, our data did conform to *essential unidimensionality* conditions (Reise and Rodriguez, 2016). Yet, our purpose was *not* to maximize goodness of fit irrespective of external validity, but to find models that both describe the data and allow statistically reliable estimates of severity to be compared with external validators.

In conclusion, results from IRT models suggest that it is possible to derive estimates of psychometric severity for specific psychiatric diagnostic symptoms, and that these estimates are systematically related to disability over and above the mere number of criteria. It may be possible to integrate information on the severity of specific symptoms into clinical diagnostic tools, to better identify individuals whose symptom combination is susceptible of a particularly disabling psychiatric condition.

Role of the funding source

This work was supported by the Emil Aaltonen Foundation.

Appendix A. Supplementary material

Supplementary documentation associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jad.2017.07.015.

References

Alegria, M., Jackson, J.S., Kessler, R.C., Takeuchi, D., 2015. Collaborative Psychiatric Epidemiology Surveys (CPES), 2001–2003 [United States]. Inter-university Consortium for Political, Social Research (ICPSR) [distributor], pp. 12–19.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual Of Mental Disorders (DSM-5). American Psychiatric Pub.

Andrews, G., Hobbs, M.J., Borkovec, T.D., Beesdo, K., Craske, M.G., Heimberg, R.G., Rapee, R.M., Ruscio, A.M., Stanley, M.A., 2010. Generalized worry disorder: a review of DSM-IV generalized anxiety disorder and options for DSM-V. Depress. Anxiety 27, 134–147.

Bolton, J.M., Belik, J.M.B.S.-L., Enns, M.W., Cox, B.J., Sareen, J., 2008. Exploring the correlates of suicide attempts among individuals with major depressive disorder: findings from the National Epidemiologic Survey on Alcohol and Related Conditions. J. Clin. Psychiatry 69, 1139–1149.

Berg, van den, S.M., Glas, C.A., Boomsma, D.I., 2007. Variance decomposition using an IRT measurement model. Springer Behav. Genet. 37. pp. 604–616.

Borsboom, D., 2017. A network theory of mental disorders. World Psychiatry 16, 5–13.

Borsboom, D., Cramer, A.O., 2013. Network analysis: an integrative approach to the structure of psychopathology. Annu. Rev. Clin. Psychol. 9, 91–121.

Borsboom, D., Rhemtulla, M., Cramer, A., Maas, H. van der, Scheffer, M., Dolan, C., 2016. Kinds versus continua: a review of psychometric approaches to uncover the structure of psychiatric constructs. Psychol. Med. 46, 1567.

Cai, L., 2010a. A two-tier full-information item factor analysis model with applications. Psychometrika 75, 581–612.

Cai, L., 2010b. Metropolis-hastings robbins-monro algorithm for confirmatory item factor analysis. J. Educ. Behav. Stat. 35, 307–335.

Cai, L., Hansen, M., 2013. Limited-information goodness-of-fit testing of hierarchical item factor models. Br. J. Math. Stat. Psychol. 66, 245–276.

Chalmers, R.P., 2012. mirt: a multidimensional item response theory package for the R environment. J. Stat. Softw. 48, 1–29.

Cramer, A.O., van Borkulo, C.D., Giltay, E.J., van der Maas, H.L., Kendler, K.S., Scheffer, M., Borsboom, D., 2016. Major depression as a complex dynamic system. PloS One 11 (12), e0167490.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychol. Bull. 52, 281.

de Ayala, R.J., 2013. The theory and practice of Item Response Theory. Guilford Publications.

Drasgow, F., Parsons, C.K., 1983. Application of unidimensional item response theory models to multidimensional data. Appl. Psychol. Meas. 7, 189–199.

Druss, B.G., Hwang, I., Petukhova, M., Sampson, N.A., Wang, P.S., Kessler, R.C., 2009. Impairment in role functioning in mental and chronic medical disorders in the United States: results from the National Comorbidity Survey Replication. Mol. Psychiatry 14, 728–737.

Edens, J.F., Marcus, D.K., Lilienfeld, S.O., Poythress, N.G., 2006. Psychopathic, not psychopath: taxometric evidence for the dimensional structure of psychopathy. J. Abnorm. Psychol. 115, 131–144.

Epskamp, S., Maris, G.K.J., Waldorp, L.J., Borsboom, D., 2016. Network Psychometrics. ⟨https://arxiv.org/pdf/1609.02818v1.pdf⟩ (Accessed 25 March 2017).

Fayers, P.M., Hand, D.J., 2002. Causal variables, indicator variables and measurement scales: an example from quality of life. J. R. Stat. Soc.: Ser. A (Stat. Soc.) 165, 233–253.

First, M.B., Wakefield, J.C., 2013. Diagnostic criteria as dysfunction indicators: bridging the chasm between the definition of mental disorder and diagnostic criteria for specific disorders. Can. J. Psychiatry 58, 663–669.

Fried, E.I., 2015. Problematic assumptions have slowed down depression research: why symptoms, not syndromes are the way forward. Front. Psychol. 6, 309.

Fried, E.I., Borkulo, C.D. van, Cramer, A., Boschloo, L., Schoevers, R.A., Borsboom, D., 2016. Mental disorders as networks of problems: a review of recent insights. Soc. Psychiatry Psychiatr. Epidemiol. 52, 1–32.

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 sum-scores don't add up: why analyzing specific depression symptoms is essential. BMC Med. 13, 72.

Fried, E.I., Nesse, R.M., 2015b. 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., Guille, C., Sen, S., 2015. The differential influence of life stress on individual symptoms of depression. Acta Psychiatr. Scand. 131, 465–471.

Harrison, D.A., 1986. Robustness of IRT parameter estimation to violations of the unidimensionality assumption. J. Educ. Behav. Stat. 11, 91–115.

Haslam, N., Holland, E., Kuppens, P., 2012. Categories versus dimensions in personality and psychopathology: a quantitative review of taxometric research. Psychol. Med. 42, 903–920.

Heeringa, S.G., Wagner, J., Torres, M., Duan, N., Adams, T., Berglund, P., 2004. Sample designs and sampling methods for the Collaborative Psychiatric Epidemiology Studies (CPES). Int. J. Methods Psychiatr. Res. 13, 221–240.

Horwitz, A.V., Wakefield, J.C., 2007. The Loss of Sadness: How Psychiatry Transformed Normal Sorrow Into Depressive Disorder. Oxford University Press.

Ip, E.H., 2010. Empirically indistinguishable multidimensional IRT and locally dependent unidimensional Item Response Models. Br. J. Math. Stat. Psychol. 63, 395–416.

Kenny, D.A., 2015. Measuring Model Fit. ⟨http://http://davidakenny.net/cm/fit.htm⟩ (Accessed 15 December 2016).

Kessler, R.C., Üstün, T.B., 2004. The World Mental Health (WMH) survey initiative version of the World Health Organization (WHM) Composite International Diagnostic Interview (CIDI). Int. J. Methods Psychiatr. Res. 13, 93–121.

Kline, R.B., 2015. Principles and Practice of Structural Equation Modeling. Guilford Publications.

Krueger, R.F., Tackett, J.L., MacDonald, A., 2016. Toward validation of a structural approach to conceptualizing psychopathology: a special section of the journal of abnormal psychology. J. Abnorm. Psychol. 125, 1023–1026.

Kotov, R., Krueger, R.F., Watson, D., Achenbach, T.M., Althoff, R.R., Bagby, R.M., Brown, T.A., Carpenter, W.T., Caspi, A., Clark, L.A., Eaton, N.R., Forbes, M.K., Forbush, K.T., Goldberg, D., Hasin, D., Hyman, S.E., Ivanova, M.Y., Lynam, D.R., Markon, K., Miller, J.D., Moffitt, T.E., Morey, L.C., Mullins-Sweatt, S.N., Ormel, J., Patrick, C.J., Regier, D.A., Rescorla, L., Ruggero, C.J., Samuel, D.B., Sellbom, M., Simms, L.J., Skodol, A.E., Slade, T., South, S.C., Tackett, J.L., Waldman, I.D., Waszczuk, M.A., Widiger, T.A., Wright, A.G.C., Zimmerman, M., 2017. The Hierarchical Taxonomy of Psychopathology (HiTOP): a dimensional alternative to traditional Nosologies. J. Abnorm. Psychol (Advance online publication).

Loevinger, J., 1957. Objective tests as instruments of psychological theory. Psychol. Rep. 3, 635–694.

Marcus, D.K., Lilienfeld, S.O., Edens, J.F., Poythress, N.G., 2006. Is antisocial personality disorder continuous or categorical? A taxometric analysis. Psychol. Med. 36, 1571–1581.

Marcus, D.K., Ruscio, J., Lilienfeld, S.O., Hughes, K.T., 2008. Converging evidence for the latent structure of antisocial personality disorder: consistency of taxometric and latent class analyses. Crim. Justice Behav. 35, 284–293.

Marsman, M., Maris, G., Bechger, T., Glas, C., 2015. Bayesian inference for low-rank Ising networks. Sci. Rep. 5, 9050.

Maydeu-Olivares, A., 2013. Goodness-of-fit assessment of item response theory models. Meas.: Interdiscip. Res. Perspect. 11, 71–101.

McNally, R.J., Robinaugh, D.J., Wu, G.W., Wang, L., Deserno, M.K., Borsboom, D., 2015. Mental disorders as causal systems a network approach to posttraumatic stress disorder. Clin. Psychol. Sci. 3, 836–849.

Messner, J.W., Zeileis, A., Broecker, J., Mayr, G.J., 2014. Probabilistic wind power forecasts with an inverse power curve transformation and censored regression. Wind Energy 17, 1753–1766.

Muthén, B.O., 1989. Factor structure in groups selected on observed scores. Br. J. Math. Stat. Psychol. 42, 81–90.

Pennell, B.-E., Bowers, A., Carr, D., Chardoul, S., Cheung, G.-Q., Dinkelmann, K., Gebler, N., Hansen, S.E., Pennell, S., Torres, M., 2004. The development and implementation of the National Comorbidity Survey Replication, the National Survey of American Life, and the National Latino And Asian American Survey. Int. J. Methods Psychiatr. Res. 13, 241–269.

Pietrzak, R.H., Tsai, J., Armour, C., Mota, N., Harpaz-Rotem, I., Southwick, S.M., 2015. Functional significance of a novel 7-factor model of DSM-5 PTSD symptoms: results from the National Health and Resilience in Veterans Study. J. Affect. Disord. 174,

522–526.
R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Reckase, M., 2009. Multidimensional Item Response Theory 150 Springer.
Reise, S.P., Cook, K.F., Moore, T.M., 2014. Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In: Reise, S., Revicki, D. (Eds.), Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment. Routledge, New York, pp. 13–40.
Reise, S., Rodriguez, A., 2016. Item response theory and the measurement of psychiatric constructs: some empirical and conceptual issues and challenges. Psychol. Med. 46, 2025–2039.
Reise, S.P., Waller, N.G., 2009. Item response theory and clinical measurement. Annu. Rev. Clin. Psychol. 5, 27–48.
Revelle, W., 2015. Psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois.
Rosenström, T., Jokela, M., 2017. Reconsidering the definition of major depression based on collaborative psychiatric epidemiology surveys. J. Affect. Disord. 207, 38–46.
Ruscio, A.M., Lane, M., Roy-Byrne, P., Stang, P.E., Stein, D.J., Wittchen, H.-U., Kessler, R.C., 2005. Should excessive worry be required for a diagnosis of Generalized Anxiety Disorder? Results from the US National Comorbidity Survey Replication. Psychol. Med. 35, 1761–1772.
Sass, D., Schmitt, T., Walker, C., 2008. Estimating non-normal latent trait distributions within Item Response Theory using true and estimated item parameters. Appl. Meas. Educ. 21, 65–88.
Schmittmann, V.D., Cramer, A.O., Waldorp, L.J., Epskamp, S., Kievit, R.A., Borsboom, D., 2013. Deconstructing the construct: a network perspective on psychological phenomena. New Ideas Psychol. 31, 43–53.
Stein, D.J., Lund, C., Nesse, R.M., 2013. Classification systems in psychiatry: diagnosis and global mental health in the era of DSM-5 and ICD-11. Curr. Opin. Psychiatry 26, 493–497.
Thomas, M.L., 2011. The value of Item Response Theory in clinical assessment: a review. Assessment 18, 291–307.
Tobin, J., 1958. Estimation of relationship for limited dependent variables. Econometrica 26, 24–36.
Tweed, D.L., 1993. Depression-related impairment: estimating concurrent and lingering effects. Psychol. Med. 23, 373–386.
Üstün, T.B., 2010. Measuring Health and Disability: Manual for WHO Disability Assessment Schedule WHODAS 2.0. World Health Organization.
Wakefield, J.C., 1992. The concept of mental disorder: on the boundary between biological facts and social values. Am. Psychol. 47, 373–388.
Wakefield, J.C., 2015. DSM-5, psychiatric epidemiology and the false positives problem. Epidemiol. Psychiatr. Sci. 24, 188–196.
Wakefield, J.C., Demazeux, S., 2016. Sadness or Depression? International Perspectives on the Depression Epidemic and Its Meaning. Springerpp. 2016.
Wakefield, J.C., First, M.B., 2013. Clarifying the boundary between normality and disorder: a fundamental conceptual challenge for psychiatry. Can. J. Psychiatry 58, 603–605.
Wakefield, J.C., Schmitz, M.F., 2015. Feelings of worthlessness during a single complicated major depressive episode predict postremission suicide attempt. Acta Psychiatr. Scand. 133, 257–265.
Wakefield, J.C., Schmitz, M.F., 2017a. Symptom quality versus quantity in judging prognosis: using NESARC predictive validators to locate uncomplicated major depression on the number-of-symptoms severity continuum. J. Affect. Disord. 208, 325–329.
Wakefield, J.C., Schmitz, M.F., 2017b. Severity of complicated versus uncomplicated subthreshold depression: new evidence on the “Monotonicity Thesis” from the national comorbidity survey. J. Affect. Disord. 212, 101–109.
Wakefield, J.C., Schmitz, M.F., First, M.B., Horwitz, A.V., 2007. Extending the bereavement exclusion for major depression to other losses: evidence from the National Comorbidity Survey. Arch. Gen. Psychiatry 64, 433–440.
Wanders, R., Van Loo, H., Vermunt, J., Meijer, R., Hartman, C., Schoevers, R., Wardenaar, K., De Jonge, P., 2016. Casting wider nets for anxiety and depression: disability-driven cross-diagnostic subtypes in a large cohort. Psychol. Med. 46, 3371.
Wittchen, H.-U., Kessler, R.C., Zhao, S., Abelson, J., 1995. Reliability and clinical validity of UM-CIDI DSM-III-R Generalized Anxiety Disorder. J. Psychiatr. Res. 29, 95.