Accepted refereed manuscript of: Teymoori A, Real R, Gorbunova A, Haghish EF, Andelic N, Wilson L, Asendorf T, Menon D & von Steinbüchel N (2020) Measurement invariance of assessments of depression (PHQ-9) and anxiety (GAD-7) across sex, strata and linguistic backgrounds in a European-wide sample of patients after Traumatic Brain Injury. *Journal of Affective Disorders*, 262, pp. 278-285. DOI: https://doi.org/10.1016/j.jad.2019.10.035

Measurement invariance of assessments of depression (PHQ-9) and anxiety (GAD-7) across sex, strata and linguistic backgrounds in a European-wide sample of patients after Traumatic Brain Injury

Ali Teymoori[1], *Ruben Real*[1], *Anastasia Gorbunova*[1], *Haghish E. F.*[1], *Nada Andelic*[2, 3], *Lindsay Wilson*[4], *Thomas Asendorf* [1,5], *David Menon*[6], *Nicole v. Steinbüchel*[1]

[1] The Institute of Medical Psychology and Medical Sociology, Medical Center, Georg August University of Göttingen, Germany
[2] Department of Physical Medicine and Rehabilitation, Oslo University Hospital, Oslo, Norway
[3] Institute of Health and Society, Research Centre for Habilitation and Rehabilitation Models and Services (CHARM), Faculty of Medicine, University of Oslo, Oslo, Norway
[4] Department of Psychology, University of Stirling, Stirling, the UK
[5] Institute of Medical Statistics, Medical Center, Georg August University of Göttingen, Germany
[6] Division of Anaesthesia, University of Cambridge/Addenbrooke's Hospital, Cambridge, UK

Abstract

Background. The Patient Health Questionnaire-9 (PHQ-9) and the Generalized Anxiety Disorder (GAD-7) are two widely used instruments to screen patients for depression and anxiety. Although many studies have investigated the validity of these two measurement instruments for medical settings, few studies have focused on their invariance across groups with different demographic and linguistic background. Comparable psychometric properties

across different demographic and linguistic groups are necessary for multiple group comparison and international research on depression and anxiety.

Objectives and Method. The main aim of this study is to examine measurement invariance for the PHQ-9 and GAD-7 in traumatic brain injury (TBI) medical setting by: a) the sex of the participants, b) recruitment stratum, and c) linguistic background. This study is based on non-randomized observational data six months after TBI that were collected in 18 countries from 2014 to 2017 in the CENTER-TBI study (Collaborative European NeuroTrauma Effectiveness Research after TBI). We used multiple methods to detect Differential Item Functioning (DIF) including Item Response Theory (IRT), logistic regression (LR), and the Mantel-Haenszel (MH) method.

Results. The total number of participants at the center-TBI study was 4509. We analyzed those who had 16 years of age or above, which were 4360 participants. 473 of the patients were deceased at the 6-month post-injury, majority of whom were from ICU stratum (83%). Out of the remaining 3886 participants, 2137 participants completed the data for psychological outcome including PHQ-9 and GAD-7. The participants were 738 (34.5%) women and 1399 (65.5%) men, encompassing patients primary admitted to the Intensive Care Unit (ICU, 885 [41.4%] at the time of enrollment), patients admitted to hospital ward (Admission stratum, 805 patients [37.7%]), and patients evaluated in the Emergency Room and discharged (ER, 447 [20.9%]). Results supported the invariance of PHQ-9 and GAD-7 across sex, patient strata and linguistic background. For different strata three PHQ-9 items and one GAD-7 item and for different linguistic groups only two GAD-7 items were flagged as showing differences in two out of four DIF tests. However, the magnitude of the DIF effect was negligible and did not seem to affect the latent mean of the scales.

Conclusion. The PHQ-9 and GAD-7 scales are invariant across sex, strata, and linguistic groups. The findings demonstrate adequate psychometric properties for PHQ-9 and GAD-7,

allowing direct comparison of depression and anxiety in multilingual studies after TBI as well as across sex and strata.

Keywords: Depression, anxiety, measurement invariance, Differential Item Functioning (DIF)

Background

Traumatic Brain Injury (TBI) is characterized by alterations of brain functions including loss of consciousness and/or memory, neurological deficit such as loss of balance or vision, and alteration of mental state at the time of injury such confusion and disorientation (Maas et al., 2017). TBI is generally categorized as severe, moderate and mild (Maas et al., 2017). More than 50 million people worldwide experience TBI each year and, according to various estimations, nearly half of the world's population will suffer from some form of TBI at least once over their lifetime (Maas et al., 2017).

Depression and anxiety are the most commonly experienced mental health disorders among patients after TBI (Moore, Terryberry-Spohr, & Hope, 2006; Perry et al., 2016). When untreated, depression and anxiety not only impede the patient's recovery from TBI but also leave a lasting impairment in their post-TBI quality of life (Mooney & Speed, 2001), such as post-concussion symptoms, deterioration of executive functions (Fann et al., 2005; Rapoport, Kiss, & Feinstein, 2006), and poorer social functioning. These cognitive and psychological problems are matched by a range of structural changes, including lower prefrontal gray matter volumes (Jorge et al., 2004). After TBI the prevalence of depression is between 15% to 27% (Fann et al., 2005; Seel et al., 2003) and the prevalence of anxiety is between 23-29% (Bryant et al., 2010; Mooney & Speed, 2001; Moore et al., 2006), which these rates are higher than their prevalence rate in the general population (Moore et al., 2006). Bryant et al. (2010) found that

general anxiety disorder is experienced more frequently than other types of anxiety among patients after mild TBI. Given this prevalence of depression and anxiety among patients with TBI and their impact on recovery process, accurate, comparable, valid and reliable assessment of depression and anxiety symptoms is gaining importance in multinational studies as well as in primary care.

Screening instruments have enabled the detection of depression and anxiety, and thereby help identify patients who would benefit from treatment. Two widely used instruments for screening patients with depression and anxiety are the nine-item Patient Health Questionnaire (PHQ-9: Kroenke & Spitzer, 2002; Kroenke, Spitzer, & Williams, 2001) and the seven-item scale for General anxiety Disorder (GAD-7: Spitzer, Kroenke, Williams, & Löwe, 2006). The PHQ-9 items are based on the DSM-IV criteria and demonstrate relatively high sensitivity and specificity to detect possible depression (Kroenke & Spitzer, 2002). Similarly, the GAD-7 items are based on DSM-IV criteria and have shown high specificity and sensitivity in identifying possible anxiety disorders (Kroenke, Spitzer, Williams, Monahan, & Löwe, 2007; Spitzer et al., 2006). These instruments have been successfully used in various medical contexts (see Kroenke et al., 2010) including TBI (e.g., Fann et al., 2005; Fogelberg, Hoffman, Dikmen, Temkin, & Bell, 2012).

Despite much research on their validity and optimal cut-off points (for a review, see Kroenke et al., 2010), few studies have investigated the PHQ-9 and GAD-7 invariance across different groups. Currently only a handful of studies have rigorously evaluated the invariance of the PHQ-9 and GAD-7 across patients' socio-demographic backgrounds, and even fewer have assessed their cross-linguistic invariance (Arthurs et al., 2012; Galenkamp, Stronks, Snijder, & Derks, 2017). The PHQ-9 and GAD-7 have been shown to have comparable psychometric characteristics in men and women in the general population and in a range of medical settings (Löwe et al., 2008; Rutter & Brown, 2017). Since sex differences in depression

and anxiety scores have been observed among patients with TBI (Bay, Sikorskii, & Saint-Arnault, 2009.; Van Reekum, Bolago, Finlayson, Garner, & Links, 1996), it is important to examine whether the test items function similarly between men and women, and ensure that the finding of sex differences in the incidence of these conditions in TBI is not due to measurement error. In addition, some of the items in PHQ-9 and GAD-7 scales deal with somatic aspect of depression and anxiety symptoms, assessment of which could be modulated by the presence and severity of both TBI and extracranial injuries. Given that the spectrum of TBI includes patients in different care strata, with different severities and types of both cranial and extracranial injuries it is also important that the tools we use to assess psychological health outcomes are not confounded by these factors, hence the necessity of strata measurement invariance. Finally, the cross-cultural measurement invariance of PHQ-9 and GAD-7 has focused predominantly on the racial/ethnic groups in a country (e.g., African American vs Asian American, see Keum, Miller, & Inkelas, 2018), rather than inter-linguistic invariance. The PHQ-9 and GAD-7 scales have been translated to, and validated in, many languages (for a review see, Gilbody et al., 2007; Plummer et al., 2016), predicating the necessity of providing evidence of comparable psychometric properties across different linguistics groups. Demonstration of such multilingual measurement invariance is essential to allow study and comparison of depression and anxiety across different countries.

This manuscript seeks to to provide evidence of comparable psychometric properties of PHQ-9 and GAD-7 across sex, strata and linguistic background in the multilingual CENTER-TBI study, thus providing insights that can be applied to the broader medical context of TBI and other neurological diseases.

Method

Participants

The study uses the CENTER-TBI (Core 2.0) data which is a multinational European data obtained from 59 different medical and research centers across 18 countries. The CENTER-TBI study is based on prospective longitudinal non-randomized observational data that initially recruited 4509 patients with a clinical diagnosis of TBI. Inclusion criteria consist of patients recruited within 24 hours after their TBI, diagnosis of TBI, clinical indication for a CT-scan, and informed consent (Maas et al., 2015).

Patients with severe preexisting neurological disorders which might have confounded neurological outcome assessment (such as cerebrovascular accident, transient ischemic attacks, epilepsy, etc) were excluded from the study. We also included participants that have 16 years of age or above, which were 4360 participants. 473 of the patients were deceased at the 6-month post-injury, majority of whom were from ICU stratum (83%). Out of the remaining 3886 participants, 2137 participants completed the data for psychological outcome including PHQ-9 and GAD-7. The participants were 738 (34.5%) women and 1399 (65.5%) men, encompassing patients primary admitted to the Intensive Care Unit (ICU, 885 [41.4%] at the time of enrollment), patients admitted to hospital ward (Admission stratum, 805 patients [37.7%]), and patients evaluated in the Emergency Room and discharged (ER, 447 [20.9%]).

Ethical approval

The CENTER-TBI study has been conducted in conformance with all relevant local national ethical guideline and regulatory requirements for recruiting human subjects, as well as with relevant data protection, privacy regulations and informed consent. The study obtained ethical clearance from both, EU and the relevant institutions across all countries that were involved in the project (for a list of sites, ethical committees, and ethical approval details, see https://www.center-tbi.eu/project/ethical-approval).

Instruments

Socio-demographic information was assessed at the time of inclusion into the study to examine participants' sex, age, family status (single, partnership, married, divorced), and socio-economic background (e.g., education level, employment status).

The *Glasgow Coma Scale* (GCS) assesses coma and impaired consciousness after TBI (Teasdale et al., 2014). The GCS scores were obtained at several time points within 24 hours post-injury such as pre-hospital, first arrival at hospital, and post-stabilization. Following the IMPACT methodology (Marmarou et al., 2007), GCS scores are based on the post-stabilization period, and when the score was not available at the post-stabilization stage, the previous non-missing scores were used. The GCS categorizes injury into severe (3–8), moderate (9–12) and mild (13–15).

The *Glasgow Outcome Scale, Extended (GOSE)* assesses functional disabilities after TBI (Wilson, Pettigrew, & Teasdale, 1998). GOSE classifies functional outcomes into eight categories from 1 to 8: dead (1), vegetative state (2), lower severe disability (3), upper severe disability (4), lower moderate disability (5), upper moderate disability (6), lower good recovery (7) and upper good recovery (8).

The *PHQ-9* measures the frequency of symptoms of depression using nine items on a 4-point Likert-scale ranging from 0 (*not at all*) to 3 (*nearly every day*). A total score ranging from 0 to 27 is obtained by summing all items; ordinary mean substitution is used for missing items if less than one third (less than three items) are missing. Based on the total score of PHQ-9, the depression symptoms severity are categorized into minimal (0-4), mild (5-9), moderate (10-14), moderately severe (15-19), and severe (20-27) (Kroenke et al., 2001).

The *GAD-7* is a brief self-report scale for symptoms of General Anxiety Disorder (GAD, Spitzer et al., 2006). Seven items asses the frequency of symptoms of anxiety with a 4-point Likert-scale ranging from 0 (*not at all*) to 3 (*nearly every day*). A total score (min 0, max 21) is obtained by summing across all items; ordinary mean substitution is used for missing items providing less than one third (less than two items) are missing. The total score is categorized into minimal (0-4), mild (5-9), moderate (10-14), and severe (15-21) anxiety symptoms (Spitzer et al., 2006).

We used the translated versions of the PHQ-9 and GAD-7 questionnaires that were already available in the respective languages (see https://www.phqscreeners.com/select-screener/36).

Statistical analysis

Internal consistency was analyzed using Cronbach's alpha and Guttmann's coefficient. The item level descriptive statistics of the PHQ-9 and GAD-7 items as well as the mean analysis showed skewed responses in many items (Appendix 1), creating a floor effect and violating the normality assumption. In addition, in response to depression and anxiety items, the majority of participants answered 0 and 1 response categories, the lower end of the item responses, indicating that they were not at all or just a few days of the week bothered by the symptoms of the depression and anxiety. For instance, in response to a question as to on how many days of the last two weeks individuals were bothered by thoughts about death (item 9 of PHQ-9), 1844 participants were not bothered at all (0), 210 for just several days (1), 32 for more than half the days (2), and 35 participants nearly every day (3). Moreover, using the original response format, the initial CFA analysis shows dissatisfactory model fit indices even with the use of ordinal estimator, the weighted least square mean and variance adjusted (WLSMV). Consequently, we

dichotomized the items, with two values: 0 (no depression/no anxiety) and 1 (some degree of depression/anxiety, collapsing the original scores from one to four into one).

Before conducting the DIF-test, we examined the unidimensionality of PHQ-9 and GAD-7 scales as a precondition for doing IRT-based measurement invariance, since most DIF tests cannot account for the relations between subdimensions (Reise, Widaman, & Pugh, 1993; Tay, Meade, & Cao, 2015). We inspected the scree-plot of the successive eigenvalues and the Kaiser-Guttman criterion to examine the optimal number of factors. To further test the unidimensionality of PHQ-9 and GAD-7, Confirmatory Factor Analysis (CFA) was used with WLSMV estimator and theta parameterization for ordinal variables. To assess the model fit indices, we used the result of chi square (χ^2) and the alternative model fit indices, including the root mean square error of approximation (RMSEA, acceptable model fit if < 0.05) and its 95% confidence interval (CI), the Standardized Root Mean Square Residual (SRMR, acceptable model fit if < 0.08), and the estimate of more than 0.95 for the Comparative Fit Index (CFA), the Tucker-Lewis Index (TLI), the Normed Fit Index (NFI), and incremental fit index (IFI) (Byrne, 2016; Hu & Bentler, 1999).

We examined measurement invariance with methods for detecting Differential Item Functioning (DIF), to essess whether the items functioned similarly across different sexes, strata and linguistic groups (Putnick & Bornstein, 2016). DIF occurs when the relation between the latent variable and item responses differ on item parameters, such as item difficulty, across groups. The existence of DIF would indicate that group differences might not be due to actual differences of groups in the variable under investigation, rather due to other factors such as measurement artifacts or external contextual factors (Zumbo, 2007).

Following Hambleton (2006), we applied multiple methods of DIF detection including Item Response Theory (IRT), logistic regression (LR), and the Mantel-Haenszel (MH) methods (see Hambleton, 2006; Zumbo, 2007) to ensure potential DIF would not go undetected. We

performed Lord's chi-square approach for IRT-based DIF and two approaches for logistic regression including likelihood ratio approach and the Wald test (Magis, Béland, Tuerlinckx, & De Boeck, 2010). We tested both uniform (MH) and non-uniform DIF (LR and Lord-χ^2). The uniform DIF examines whether the items are invariant relative to the reference group. In the non-uniform DIF, a priority is not given to the reference group and an interaction term between group membership and individual ability to answer the items are taken into account (Magis et al., 2010; Tay et al., 2015). Moreover, because the existence of one or more DIF items might influence the result of the test for the rest of the items, we applied an item deletion method known as item purification, in which the test is conducted again with iterative elimination of the DIF items (Clauser & Mazor, 1998). After the detection of DIF items, the analysis was repeated with deletion of the DIF items to estimate the test parameters such as the test score. The test was then repeated to ensure detection of potential DIF in the remaining items, if present (Clauser & Mazor, 1998). The process stops when the two successive iteration yields the same results. Finally, the conventional setting for DIF is pairwise comparison of a reference group with a focal group. Some recent methods can accommodate multiple group comparison and was specifically developed to compare all groups simultaneously, such as generalized Lord test and generalized logistic regression (for a detailed explanation about each method, see the supplementary materials).

We evaluated our study sample for sex and strata measurement invariance. However, due to the sample size restriction for DIF tests, we chose the linguistic groups with more than 200 respondents after the case-wise deletion of the missing values. Data questionnaires were conducted in the respective countries' native languages (see Table 1). We combined the data from Belgian and Switzerland institutes with the respective languages obtained from other countries such that the data from centers in the Flemish speaking part of Belgium was analyzed together with the data from the Netherlands and the data from the French speaking part of

Belgium and Switzerland analyzed together with the data from France. Consequently, we chose, six linguistic groups that had more than 200 participants including participants with Dutch (from the Netherlands and the Flemish speaking part of Belgium), English, Italian, Spanish, Finish, and Norwegian linguistic background.

All analyses were conducted using R version 3.5.1, and the packages "lavaan" (Rosseel, 2012), "psych" (Revelle, 2017), and "difR" (Magis, Beland, & Raiche, 2018). We exported the data from the Neurobot platform of the CENTER-TBI (https://center-tbi.incf.org/) and used the "CENTER Core 2.0" dataset which were the latest curated version of the data available at the time of current paper's data analysis in July 2019.

Results

Respondents characteristics

The basic demographic and medical characterization such as sex, age, patients stratum, GCS, and GOSE (see Table 1). The mean age of the our sample was 49.19 (*SD* = 19.30) with almost a third being 65 years of age or older (521, 24.4%). Women (*M* = 52.49, *SD* = 19.73) had higher mean age than men (*M* = 47.45, *SD* = 18.85; *t(1441)* = 5.70, $p < .001$). The level of education was relatively high: 34.1% (651) secondary or high school education, 21.1% (402) post-high school education such as technical college or professional training, and 27.9% (533) university education. Only 1.2% (23) of participants had no education. More than a quarter of our study sample were Dutch-speaking (29.6%) and the other participants were from diverse linguistic background such as English (10.4%), Italian (13.6%), Spanish (12.8%), Finish (10.4%), Norwegian (13.1%). According to GOSE score, more than half of the participants (62.8%) showed good recovery after TBI, a quarter had moderate disability (25.7%) and 10.3% had severe diability at 6 months after the TBI.

The mean scores for depression (PHQ-9, *M* = 5.07, *SD* = 5.35) and anxiety symptoms (GAD-7, *M* = 3.63, *SD* = 4.54) were rather low given that the maximum score for PHQ-9 and GAD-7 were 27 and 21, respectively (Table 1). Categorizing the patients based on the severity of depression, more than two third of the participants were classified as having no (1233 [58.0%], 1484 [69.9%]) or mild (510 [24.0%], 402 [18.9%]) depression and anxiety, respectively (Table 2). For depression, 18% of the participants suffered from moderate (225 [10.6%]), moderately severe (103 [4.8%]), or severe (54 [2.5%]) symptoms. Men were more likely to be in the minimal level of depression severity (69.3%) than women and in turn women were more likely to be in moderate and moderately severe depression categories such that only 34.6% of the PHQ-9 respondents were women, but 41.8% of moderate and 43.7% of the moderately severe participants were women. For anxiety, 11% of participants showed moderate (148 [7.0%]) or severe (88 [4.1%]) symptoms and, similar to PHQ-9 categories distribution, women had higher proportion of these moderate and severe anxiety categories than men. The demographics of the patients with different severity of depression and anxiety are shown in Table 2.

Table 1

Descriptive characteristics of the study sample at 6-month data collection (n = 2137)

Variable		Overall	Stratified	
N		2137	Female [738]	Male [1399]
Language [n, (%)]				
	Dutch	587 (29.6)	226 (38.5)	361 (61.5)
	English	207 (10.4)	70 (33.8)	137 (66.2)
	Finish	207 (10.4)	83 (40.1)	124 (59.9)
	French	117 (5.9)	27 (23.1)	90 (76.9)
	German	81 (4.1)	26 (32.1)	44 (54.3)
	Italian	269 (13.6)	84 (31.2)	185 (68.8)
	Norwegian	260 (13.1)	81 (31.2)	179 (68.8)
	Spanish	254 (12.8)	76 (29.9)	178 (70.1)
Age groups [n, (%)]				
	[16-24]	325 (15.2)	95 (29.2)	230 (70.8)
	[25-34]	262 (12.3)	78 (29.8)	184 (70.2)
	[35-44]	274 (12.8)	67 (24.5)	207 (75.5)
	[45-54]	357 (16.8)	130 (36.4)	228 (63.9)
	[55-64]	397 (18.6)	146 (36.8)	251 (63.2)
	[>=65]	521 (24.4)	222 (42.6)	299 (57.4)
Patient type [n, (%)]				
	ER	447 (20.9)	198 (44.3)	249 (55.7)
	Admission	805 (37.7)	297 (36.9)	508 (63.1)
	ICU	885 (41.4)	243 (27.5)	642 (72.5)

GOSE [n, (%)]				
	Severe dis. [2-4]	221 (10.3)	81 (36.7)	140 (63.3)
	Moderate dis. [5-6]	549 (25.7)	185 (33.7)	364 (66.3)
	Good recovery [7-8]	134 (62.8)	459 (342.5)	882 (658.2)
	NA	26 (1.2)	13 (50)	13 (50)
GCS [n, (%)]				
	Severe	340 (16.4)	90 (26.5)	250 (73.5)
	Moderate	162 (7.8)	55 (34)	107 (66)
	Mild	1571 (75.8)	575 (36.6)	996 (63.4)
Employment category [n, (%)]				
	Employed. F.T.	909 (46.0)	225 (24.8)	684 (75.2)
	Employed. P.T.	220 (11.1)	119 (54.1)	101 (45.9)
	Retired	492 (24.9)	207 (42.1)	285 (57.9)
	Sick leave	15 (0.8)	7 (46.7)	8 (53.3)
	Student	198 (10.0)	65 (32.8)	133 (67.2)
	Unemployed	144 (7.3)	38 (26.4)	106 (73.6)
Relationship status [n, (%)]				
	Partnered	185 (9.2)	61 (33)	124 (67)
	Married	912 (45.3)	296 (32.5)	616 (67.5)
	Never married	615 (30.5)	177 (28.8)	438 (71.2)
	Div., sep., wido.	303 (15.0)	155 (51.2)	148 (48.8)
Education level [n, (%)]				
	None	23 (1.2)	12 (52.2)	11 (47.8)
	Currently studying	58 (3.0)	22 (37.9)	36 (62.1)
	Primary school	240 (12.6)	86 (35.8)	154 (64.2)
	Secondary/high school	651 (34.1)	213 (32.7)	438 (67.3)
	Post high school	402 (21.1)	109 (27.1)	293 (72.9)
	University/college	533 (27.9)	209 (39.2)	324 (60.8)
[Mean, (SD)]				
	Depression (PHQ-9)	5.07 (5.35)	5.73 (5.48)	4.72 (5.25)
	Anxiety (GAD-7)	3.63 (4.54)	4.19 (4.75)	3.34 (4.41)

Note: ER: emergency room, ICU: intensive care unit, Severe dis.: severe disability, Moderate dis.: moderate disability, Employed. F.T.: employed full-time, Employed. P.T.: employed part-time, Div., sep., wido.: divorced, separated or widowed.

Reliability

Both the Cronbach alpha (α = 0.87, 0.91) and Gutman lambda two (λ_6 = 0.88, 0.91) showed very good internal consistency for PHQ-9 and GAD-7, respectively. The correlation between PHQ-9 and GAD-7 was also high, $r = 0.80$, $p < 0.001$. A more in-depth analysis of reliability, however, indicated that the reliability coefficient was not uniform and not consistent across the entire range of the latent score. Factor analysis of the IRT parameters of item difficulty and item discrimination suggested that the tests performed better for distinguishing those who have higher scores, and provided much less information about those who had less severe depression

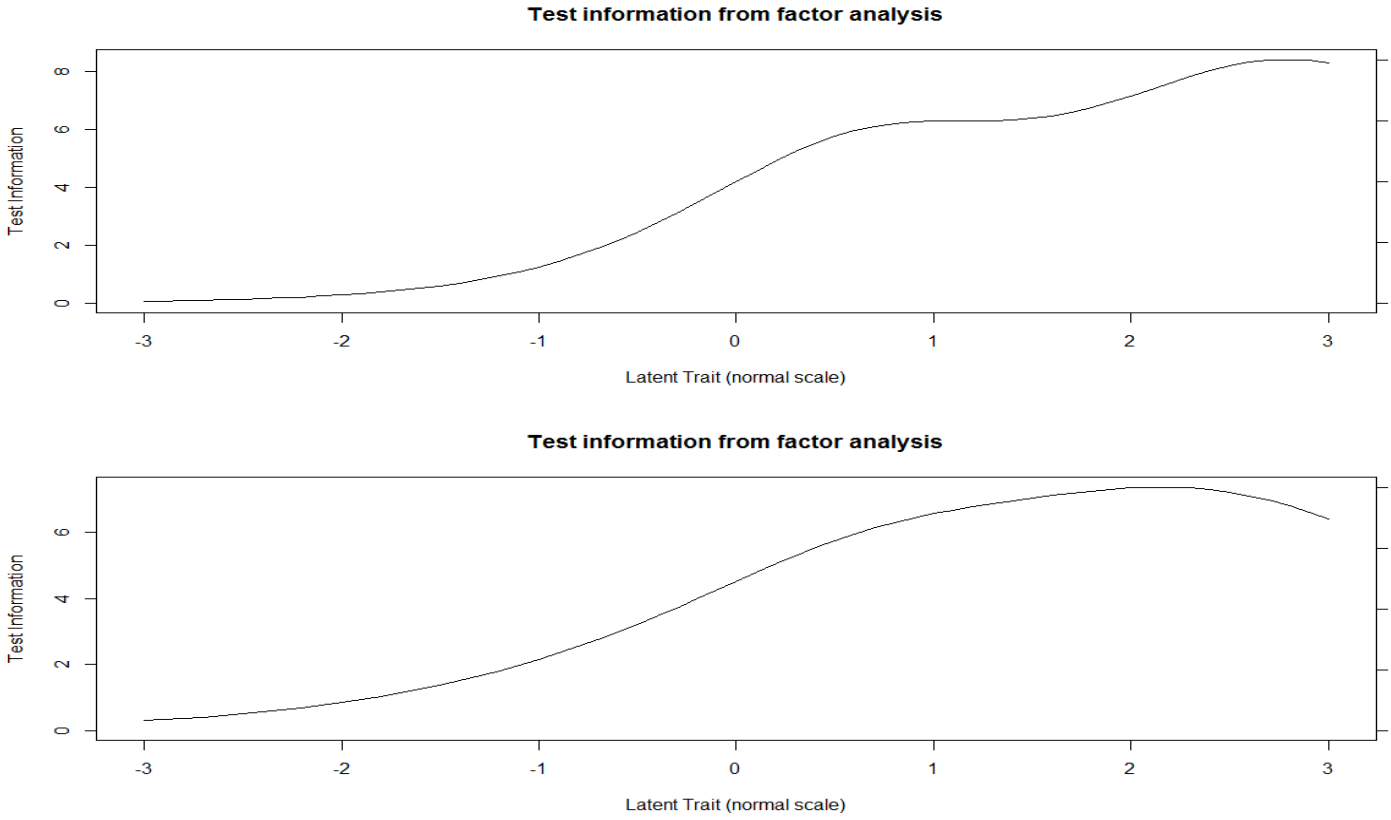
and/or anxiety (see Figure 1, for more information regarding the extraction of item difficulty and item discrimination parameters' estimates from factor analysis, see Revelle, 2017).

Table 2

Prevalence and demographics of patients with different depression and anxiety severity

		Depression (N = 2127, Men: 1391 [65.4%], Women: 736 [34.6%])				
Variables		Level of Severity				
		Minimal (0-4)	**Mild (5-9)**	**Moderate (10-14)**	**Moderately Severe (15-19)**	**Severe (20-27)**
N (%)		1233 (58.0)	510 (24.0)	225 (10.6)	103 (4.8)	54 (2.5)
Sex: Male (%)		854 (69.3)	313 (61.4)	131 (58.2)	58 (56.3)	34 (63.0)
Patient Strata (%)						
	ER.	285 (23.1)	86 (16.9)	43 (19.1)	22 (21.4)	9 (16.7)
	Admission	513 (41.6)	166 (32.5)	72 (32.0)	36 (35.0)	17 (31.5)
	ICU	435 (35.3)	258 (50.6)	110 (48.9)	45 (43.7)	28 (51.9)
Injury Severity						
	Mild	939 (76.2)	364 (71.4)	158 (70.2)	68 (66.0)	38 (70.4)
	Moderate	87 (7.1)	37 (7.3)	24 (10.7)	8 (7.8)	4 (7.4)
	Severe	170 (13.8)	94 (18.4)	39 (17.3)	23 (22.3)	10 (18.5)
	NA	37 (3.0)	15 (2.9)	4 (1.8)	4 (3.9)	2 (3.7)
		Anxiety (N = 2124, Men: 1392 [65.5%], Women: 732 [34.5%]				
Variables		Level of Severity				
		Minimal (0-4)	**Mild (5-9)**	**Moderate (10-14)**	**Severe (15-21)**	
N (%)		1484 (69.9)	402 (18.9)	148 (7.0)	88 (4.1)	
Sex: Male (%)		1011 (68.1)	245 (60.9)	84 (56.8)	52 (59.1)	
Patient Strata (%)						
	ER	313 (21.1)	84 (20.9)	29 (19.6)	17 (19.3)	
	Admission	588 (39.6)	141 (35.1)	49 (33.1)	24 (27.3)	
	ICU	583 (39.3)	177 (44.0)	70 (47.3)	47 (53.4)	
Injury Severity						
	Mild	1102 (74.3)	293 (72.9)	106 (71.6)	60 (68.2)	
	Moderate	114 (7.7)	26 (6.5)	11 (7.4)	10 (11.4)	
	Severe	221 (14.9)	72 (17.9)	29 (19.6)	15 (17.0)	
	NA	47 (3.2)	11 (2.7)	2 (1.4)	3 (3.4)	

Figure 1. Test Information Function (TIF) from Factor analysis of PHQ-9 (upper panel) and GAD-7 (lower panel) item parameters



The unidimensionality of the PHQ-9 and GAD-7

As explained earlier, we dichotomized responses to 0 (no depression/anxiety) and 1 (some depression/anxiety) by collapsing the original score of 1-4 to 1. We used parallel analysis of the scree plots and the Kaiser-Guttman criterion to determine the number of factors. We used both principal component and principal factor analyses with a tetrachoric correlation matrix of the residuals given the dichotomized nature of the rescaled items. Only one factor seemed to fit the data since a) only one factor had an eigenvalue more than 1, b) there was a sharp break in the scree plot between the first and second factor, and c) the first factor explained most of the variance (see Appendix 2 for parallel analysis of the scree plots for PHQ-9 and GAD-7).

We conducted CFA to further evaluate the unidimensionality of the latent structure of the PHQ-9 and GAD-7. For PHQ-9, the chi-square test was significant (χ^2 (27) = 91.6, p < .001)

which was expected given the large sample size. We used alternative indices as indication of model fit, and obtained good model fit indices, RMSEA = 0.03 (95 % CI = 0.03, 0.04), SRMR = .04, CFI = 0.997, TLI = 0.995, NFI = 0.995, IFI = 0.997. For GAD-7, despite the significant chi-square (χ^2 (14) = 71.0, p < .001), alternative fit indices suggest good model fit, RMSEA = 0.04 (95 % CI = 0.03, 0.05), SRMR = .03, CFI = 0.998, TLI = 0.997, NFI = 0.997, IFI = 0.998 (Hu & Bentler, 1999), confirming the unidimensionality of our measurements.

Invariance

We used several statistical DIF tests including Lord $\chi 2$, Logistic Regression (LR) with two likelihood ratio and Wald criteria, and Mantel-Haenszel (MH) method. We considered items as noninvariant if they show significant DIF in more than half of the tests (more than 2 out of 4 tests).

Regarding the sex invariance, one item of PHQ-9 (item 1: little interest or pleasure in doing thing) and two items of GAD-7 (item 5: Being so restless that is hard to sit still; item 6: Becoming easily annoyed or irritable) showed significant DIF, but the effect size was small, with negligible DIF effect in three out of four tests, implying that items were invariant between men and women (see Appendix 3). The item characteristics curve (ICC) based on the IRT-Lord test also demonstrated that the item difficulty parameter did not noticeably differ between men and women (Appendix 4).

To examine invariance between patient's strata, the DIF compared the item parameters between three strata of ICU, admission, and ER. This analysis showed that three PHQ-9 items (item 5: poor appetite or overeating; item 8: moving or speaking slowly or being restless; item 9: death or injury thoughts) and one GAD-7 item (item 6: become easily irritated or annoyed)

were flagged as noninvariant in two out of four tests; however, the effect sizes were negligible (Table 3, for a 2PL-ICC for different strata, see Appendix 5).

Finally, we conducted a multigroup comparison of groups with different linguistic background for DIF (Table 4). As can be seen in Table 4, the PHQ-9 items are invariant across the six linguistic groups. Only two of the GAD-7 items (item 3: Worrying too much about different things; item 5: Being so restless that is hard to sit still) were flagged for DIF in two out of the four tests, but effect size in LR tests were negligible (see Appendix 6, for a 2PL-ICC of different linguistic groups).

Table 3
DIF Analyses for PHQ-9 and GAD-7 across patients' strata

	IRT-based methods	Non-IRT-based methods									N. DIF
		Logistic Regression (LR)								M-H	
	GLord $\chi2$	LR-LRT				LR-Wald				M-H χ^2	
	GLord $\chi2$	LR	ΔR2	ZT	JG	LR	ΔR2	ZT	JG	M-H χ^2	
Depression											
1	29.83***	18.16***	<0.01	A	A	17.06**	<0.01	A	A	1.55	1/4
2	17.15*	4.60	<0.01	A	A	4.59	<0.01	A	A	1.34	1/4
3	9.06	9.60	<0.01	A	A	9.21	<0.01	A	A	22.96***	1/4
4	8.74	19.02***	<0.01	A	A	18.71***	<0.01	A	A	15.00**	1/4
5	21.36**	5.33	<0.01	A	A	5.33	<0.01	A	A	13.67**	2/4
6	11.99	0.44	<0.01	A	A	0.44	<0.01	A	A	0.00	0/4
7	4.65	0.40	<0.01	A	A	0.40	<0.01	A	A	0.13	0/4
8	16.58*	1.23	<0.01	A	A	1.21	<0.01	A	A	11.04*	2/4
9	22.22**	9.30	<0.01	A	A	8.68	<0.01	A	A	10.10*	2/4
Anxiety											
1	5.60	12.80	<0.01	A	A	11.31	<0.01	A	A	1.92	0/4
2	3.74	7.27	<0.01	A	A	7.00	<0.01	A	A	5.33	0/4
3	5.26	9.90	<0.01	A	A	9.72	<0.01	A	A	11.18*	1/4
4	6.13	19.08**	<0.01	A	A	15.71*	<0.01	A	A	7.60	0/4
5	3.44	7.20	<0.01	A	A	6.96	<0.01	A	A	4.77	0/4
6	22.33**	25.93***	<0.01	A	A	27.29***	<0.01	A	A	21.13***	2/4
7	6.51	5.89	<0.01	A	A	5.37	<0.01	A	A	6.19	0/4

DIF = Differential Item Functioning; IRT = Item Response Theory; LR = Logistic Regression detecting both uniform and non-uniform differential item functioning; ΔR2 = change in R^2 of the nested models (Nagelkerke,

1991); ZT = The Zumbo & Thomas (ZT) effect size for ΔR2 ("A" = negligible effect; "B" = moderate; "C" = large); JG = the Jodoin & Gierl (JG) effect size for ΔR2 ("A" = negligible effect; "B" = moderate; "C" = large); GLord = Generalized Lord's chi-square method; M-H χ^2 = Mantel-Haenszel chi-square;
Signifance codes: *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$

Table 4
DIF analyses for PHQ-9 and GAD-7 between six linguistic groups

Items	IRT-based methods	Non-IRT-Based methods									N. DIF
		Logistic Regression								M-H	
	GLord χ^2	LR-LRT				LR-Wald				M-H χ^2	
	GLord χ^2	LR	ΔR2	ZT	JG	LR	ΔR2	ZT	JG	M-H χ^2	
Depression											
1	29.43**	14.76	<0.01	A	A	14.79	<0.01	A	A	10.88	1/4
2	28.33*	16.77	<0.01	A	A	13.48	<0.01	A	A	7.96	1/4
3	19.35	34.72**	<0.01	A	A	32.80**	<0.01	A	A	8.72	0/4
4	4.87	15.72	<0.01	A	A	11.14	<0.01	A	A	9.60	0/4
5	28.68*	17.29	<0.01	A	A	18.17	<0.01	A	A	6.33	1/4
6	26.34*	14.79	<0.01	A	A	13.85	<0.01	A	A	10.34	1/4
7	21.52	24.05	<0.01	A	A	23.50	<0.01	A	A	12.45	0/4
8	5.50	1.39	<0.01	A	A	1.40	<0.01	A	A	1.41	0/4
9	7.90	1.39	<0.01	A	A	1.40	<0.01	A	A	1.41	0/4
Axiety											
1	22.16	12.85	<0.01	A	A	7.99	<0.01	A	A	34.56***	1/4
2	10.90	2.70	<0.01	A	A	3.67	<0.01	A	A	13.96	0/4
3	49.67***	11.51	<0.01	A	A	11.18	<0.01	A	A	55.76***	2/4
4	15.32	2.30	<0.01	A	A	1.35	<0.01	A	A	4.69	0/4
5	29.54**	7.85	<0.01	A	A	6.94	<0.01	A	A	26.99***	2/4
6	13.26	2.99	<0.01	A	A	4.76	<0.01	A	A	7.01	0/4
7	7.69	10.09	<0.01	A	A	9.18	<0.01	A	A	6.76	0/4

Discussion

The prevalence of depression and anxiety that we detected in our TBI population using PHQ-9 and GAD-7 were similar to most studies in literature (Fann et al., 2005; van der Horn, Spikman, Jacobs, & van der Naalt, 2013). Less than a quarter of our sample had moderate or more severe symptoms of depression and anxiety. The prevalence of moderate or severe

depression and anxiety were higher in women in comparison to men, and in patients recruited from ICU stratum in comparison to patients recruited from hospital wards and emergency rooms.

We examined the equivalence of the scales across sex, patient strata, and linguistic background. We first confirmed the unidimensionality of the PHQ-9 and GAD-7 scales with the use of exploratory and confirmatory factor analysis. The invariance tests of PHQ-9 and GAD-7 suggest the equivalence of these measurements across different sex, strata and linguistics background.

Based on the results of four different DIF tests, the PHQ-9 and GAD-7 displayed equivalent psychometrics properties across sex, patients' strata, and linguistic background. Even through some items showed statistically significant DIF, the effect size was negligible. For instance, the somatic items about appetite, movements, and death thoughts showed significant DIF in two out of four tests between patients in different strata. Further analyses showed that, perhaps due to their physical condition, patients who had been in the ICU found the questions about movement and thoughts about death easier to respond to with regard to the item difficulty coefficients (Appendix 5). However, the difference was not very large and the effect size in LRT and Wald tests showed that the difference is negligible, supporting the invariance of the respective test items (Zumbo, 2007). Given that the total score of the scales were incorporated in the logistic regression DIF tests, they do not seem to affect the mean of the total score.

Although our study has important strengths such as a large and cross-national sample, there are some shortcomings. Firstly, despite a relatively high number of patients from the ICU stratum, the sample mostly consists of patients after mild TBI. This is in line with other studies that examine the TBI severity levels amongs patients admitted to trauma centers (for a review, see Peeters et al., 2015). One of the reasons for high percentage of mild TBI in ICU

may have been the patients' other accompanying injuries in addition to their TBI, hence their admission to the ICU. Another reason for this finding might be differences in policies and infrastructure between hospitals and countries, which result in variations in admission of patients to the ICU. Moreover, this study does not evaluate the association of depression and anxiety with other physical, emotional and social functioning and clinical variables, a task that will be addressed in the future studies. Finally, it is important to note that the PHQ-9 and GAD-7 are screening instruments. Although the tests are more reliable for people with moderate and severe depression and anxiety (Figure 1), for an accurate clinical diagnosis and treatment purposes, a detailed medical and psychiatric/psychological evaluation are required.

Taken together, the current study provides evidence for good psychometric properties of the PHQ-9 and GAD-7 scales in a large observational sample of patients after TBI with a special focus on the detection of potential DIF due to sex, patient strata, and linguistic background. The study confirmed the unidimensionality of each scale and that there was no serious violation of their item functioning across different groups. These results suggest that researchers can interpret these two instruments as unidimensional and use summary scores for screening patients with TBI for depression and anxiety symptoms, and for comparison of the scores across different sex, strata and linguistic backgrounds.

References

Arthurs, E., Steele, R. J., Hudson, M., Baron, M., Thombs, B. D., & Group, (CSRG) Canadian Scleroderma Research. (2012). Are Scores on English and French Versions of the PHQ-9 Comparable? An Assessment of Differential Item Functioning. *PLOS ONE*, *7*(12), e52028. https://doi.org/10.1371/journal.pone.0052028

Bay, E., Sikorskii, A., & Saint-Arnault, D. (2009). Sex differences in depressive symptoms and their correlates after mild-to-moderate traumatic brain injury. *Journal of Neuroscience Nursing*, *41*(6), 298–309.

Bryant, R. A., O'donnell, M. L., Creamer, M., McFarlane, A. C., Clark, C. R., & Silove, D. (2010). The psychiatric sequelae of traumatic injury. *American Journal of Psychiatry*, *167*(3), 312–320.

Byrne, B. M. (2016). *Structural Equation Modeling With AMOS : Basic Concepts, Applications, and Programming, Third Edition*. https://doi.org/10.4324/9781315757421

Clauser, B. E., & Mazor, K. M. (1998). Using Statistical Procedures to Identify Differentially Functioning Test Items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Fann, J. R., Bombardier, C. H., Dikmen, S., Esselman, P., Warms, C. A., Pelzer, E., … Temkin, N. (2005). Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury. *The Journal of Head Trauma Rehabilitation*, *20*(6), 501–511.

Fogelberg, D. J., Hoffman, J. M., Dikmen, S., Temkin, N. R., & Bell, K. R. (2012). Association of Sleep and Co-Occurring Psychological Conditions at 1 Year After Traumatic Brain Injury. *Archives of Physical Medicine and Rehabilitation*, *93*(8), 1313–1318. https://doi.org/10.1016/j.apmr.2012.04.031

Galenkamp, H., Stronks, K., Snijder, M. B., & Derks, E. M. (2017). Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: the HELIUS study. *BMC Psychiatry, 17*(1), 349. https://doi.org/10.1186/s12888-017-1506-9

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *Journal of General Internal Medicine, 22*(11), 1596–1602.

Hambleton, R. K. (2006). Good Practices for Identifying Differential Item Functioning. *Medical Care, 44*(11), S182–S188. Retrieved from JSTOR.

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

Jorge, R. E., Robinson, R. G., Moser, D., Tateno, A., Crespo-Facorro, B., & Arndt, S. (2004). Major depression following traumatic brain injury. *Archives of General Psychiatry, 61*(1), 42–50.

Keum, B. T., Miller, M. J., & Inkelas, K. K. (2018). Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse US college students. *Psychological Assessment*.

Kline, P. (2014). *An Easy Guide to Factor Analysis*. https://doi.org/10.4324/9781315788135

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals, 32*(9), 509–515.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*(9), 606–613.

Kroenke, K., Spitzer, R. L., Williams, J. B., & Löwe, B. (2010). The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *General Hospital Psychiatry, 32*(4), 345–359.

Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O., & Löwe, B. (2007). Anxiety Disorders in Primary Care: Prevalence, Impairment, Comorbidity, and Detection. *Annals of Internal Medicine*, *146*(5), 317. https://doi.org/10.7326/0003-4819-146-5-200703060-00004

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical Care*, *46*(3), 266–274.

Maas, A. I. R., Menon, D. K., Adelson, P. D., Andelic, N., Bell, M. J., Belli, A., … Chesnut, R. M. (2017). Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. *The Lancet Neurology*, *16*(12), 987–1048.

Maas, A. I. R., Menon, D. K., Steyerberg, E. W., Citerio, G., Lecky, F., Manley, G. T., … Sorgner, A. (2015). Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injury (CENTER-TBI)A Prospective Longitudinal Observational Study. *Neurosurgery*, *76*(1), 67–80. https://doi.org/10.1227/NEU.0000000000000575

Magis, D., Beland, S., & Raiche, G. (2018). difR: Collection of Methods to Detect Dichotomous Differential Item Functioning (DIF) (Version 5.0). Retrieved from https://CRAN.R-project.org/package=difR

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Marmarou, A., Lu, J., Butcher, I., McHugh, G. S., Murray, G. D., Steyerberg, E. W., … Maas, A. I. (2007). Prognostic value of the Glasgow Coma Scale and pupil reactivity in traumatic brain injury assessed pre-hospital and on enrollment: an IMPACT analysis. *Journal of Neurotrauma*, *24*(2), 270–280.

Mooney, G., & Speed, J. (2001). The association between mild traumatic brain injury and psychiatric conditions. *Brain Injury*, *15*(10), 865–877.

Moore, E. L., Terryberry-Spohr, L., & Hope, D. A. (2006). Mild traumatic brain injury and anxiety sequelae: a review of the literature. *Brain Injury*, *20*(2), 117–132.

Nagelkerke, N. J. D. (1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika*, *78*(3), 691–692. https://doi.org/10.2307/2337038

Peeters, W., van den Brande, R., Polinder, S., Brazinova, A., Steyerberg, E. W., Lingsma, H. F., & Maas, A. I. R. (2015). Epidemiology of traumatic brain injury in Europe. *Acta Neurochirurgica*, *157*(10), 1683–1696. https://doi.org/10.1007/s00701-015-2512-7

Perry, D. C., Sturm, V. E., Peterson, M. J., Pieper, C. F., Bullock, T., Boeve, B. F., … Kramer, J. H. (2016). Association of traumatic brain injury with subsequent neurological and psychiatric disease: a meta-analysis. *Journal of Neurosurgery*, *124*(2), 511–526.

Plummer, F., Manea, L., Trepel, D., & McMillan, D. (2016). Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. *General Hospital Psychiatry*, *39*, 24–31. https://doi.org/10.1016/j.genhosppsych.2015.11.005

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Rapoport, M. J., Kiss, A., & Feinstein, A. (2006). The impact of major depression on outcome following mild-to-moderate traumatic brain injury in older adults. *Journal of Affective Disorders*, *92*(2–3), 273–276.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566. https://doi.org/10.1037/0033-2909.114.3.552

Revelle, W. (2017). *psych: Procedures for Personality and Psychological Research*. Retrieved from https://cran.r-project.org/web/packages/psych/index.html

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*(2), 1–36.

Rutter, L. A., & Brown, T. A. (2017). Psychometric Properties of the Generalized Anxiety Disorder Scale-7 (GAD-7) in Outpatients with Anxiety and Mood Disorders. *Journal of Psychopathology and Behavioral Assessment*, *39*(1), 140–146. https://doi.org/10.1007/s10862-016-9571-9

Seel, R. T., Kreutzer, J. S., Rosenthal, M., Hammond, F. M., Corrigan, J. D., & Black, K. (2003). Depression after traumatic brain injury: a National Institute on Disability and Rehabilitation Research Model Systems multicenter investigation. *Archives of Physical Medicine and Rehabilitation*, *84*(2), 177–184.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3–46. https://doi.org/10.1177/1094428114553062

Teasdale, G., Maas, A., Lecky, F., Manley, G., Stocchetti, N., & Murray, G. (2014). The Glasgow Coma Scale at 40 years: standing the test of time. *The Lancet Neurology*, *13*(8), 844–854. https://doi.org/10.1016/S1474-4422(14)70120-6

van der Horn, H. J., Spikman, J. M., Jacobs, B., & van der Naalt, J. (2013). Postconcussive complaints, anxiety, and depression related to vocational outcome in minor to severe traumatic brain injury. *Archives of Physical Medicine and Rehabilitation*, *94*(5), 867–874.

Van Reekum, R., Bolago, I., Finlayson, M. A. J., Garner, S., & Links, P. S. (1996). Psychiatric disorders after traumatic brain injury. *Brain Injury*, *10*(5), 319–328. https://doi.org/10.1080/026990596124340

Wilson, J. t. L., Pettigrew, L. E. l., & Teasdale, G. M. (1998). Structured Interviews for the Glasgow Outcome Scale and the Extended Glasgow Outcome Scale: Guidelines for Their Use. *Journal of Neurotrauma*, *15*(8), 573–585. https://doi.org/10.1089/neu.1998.15.573

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832