How to compare scores from different depression scales: equating the Patient Health Questionnaire (PHQ) and the ICD-10-Symptom Rating (ISR) using Item Response Theory

H. FELIX FISCHER,^{1,2} KARIN TRITT,^{3,4} BURGHARD F. KLAPP² & HERBERT FLIEGE²

1 Institut für Sozialmedizin, Epidemiologie und Gesundheitsökonomie, Charité - Universitätsmedizin Berlin, Germany

2 Medizinische Klinik mit Schwerpunkt Psychosomatik, Charité Universitätsmedizin Berlin, Germany

3 Institut für Qualitätsentwicklung in der Psychotherapie und Psychosomatik (IQP), München, Germany

4 Institut für Epidemiologie und Präventivmedizin der Universität Regensburg, Regensburg, Germany

Key words

Item Response Theory, equating, PHQ, ISR, depression self-rating scales

Correspondence

Felix Fischer, Institut für Sozialmedizin, Epidemiologie und Gesundheitsökonomie Charité - Universitätsmedizin Berlin Luisenstrasse 57, 10117 Berlin, Germany.
Telephone (+49) 30 450 529 104
Fax (+49) 30 450 529 902
Email: felix.fischer@charite.de

Received 26 August 2009;
revised 7 April 2010;
accepted 26 April 2010

Abstract

A wide range of questionnaires for measuring depression are available. Item Response Theory models can help to evaluate the questionnaires exceeding the boundaries of Classical Test Theory and provide an opportunity to equate the questionnaires.

In this study after checking for unidimensionality, a General Partial Credit Model was applied to data from two different depression scales [Patient Health Questionnaire (PHQ-9) and ICD-10-Symptom Rating (ISR)] obtained in clinical settings from a consecutive sample, including 4517 observations from a total of 2999 inpatients and outpatients of a psychosomatic clinic. The precision of each questionnaire was compared and the model was used to transform scores based on the assumed underlying latent trait.

Both instruments were constructed to measure the same construct and their estimates of depression severity are highly correlated. Our analysis showed that the predicted scores provided by the conversion tables are similar to the observed scores in a validation sample.

The PHQ-9 and ISR depression scales measure depression severity across a broad range with similar precision. While the PHQ-9 shows advantages in measuring low or high depression severity, the ISR is more parsimonious and also suitable for clinical purposes. Furthermore, the equation tables derived in this study enhance the comparability of studies using either one of the instruments, but due to substantial statistical spread the comparison of individual scores is imprecise. *Copyright* © 2011 John Wiley & Sons, Ltd.

Introduction

The frequency and clinical relevance of depressive symptoms in clinical practice have led to the development of a number of questionnaires assessing depression [e.g. Beck Depression Inventory (BDI), Hospital Anxiety and Depression Scale (HADS) and Symptom Checklist (SCL)]. Recent instruments rely on the descriptive, atheoretic and criterion-oriented approach to operationalize depression, which is the foundation of modern classifications of mental and behavioral disorders such as the International Classification of Diseases, 10th Revision (ICD-10) chapter F (Dilling *et al.*, 2005) and the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) (American Psychiatric Association, 2000). Most recent examples of such questionnaires include the Patient Health Questionnaire (PHQ) (Spitzer *et al.*, 1999) and the ICD-10-Symptom Rating (ISR) (Tritt *et al.*, 2008).

The development and evaluation of most depression questionnaires are based on Classical Test Theory (CTT), which focuses on the concepts of true scores and measurement error. Although all of these instruments ascertain to measure the same or a similar construct, namely “depression”, the comparability of classically derived sum scores is not trivial because sum scores are “test-dependent” (Hambleton and Jones, 1993). This implies, for example, that a low sum score could either derive from a person’s low ability or from a difficult test (Zickar and Broadfoot, 2008).

Item Response Theory (IRT) provides a general theoretical framework which extends the possibilities of test evaluation (Edelen and Reeve, 2007) in comparison to CTT. The idea of IRT is that response behavior to a test (i.e. solving an item or choosing a specific category) can be explained by underlying person (ability or latent trait) and item (difficulty) parameters. Relation between latent trait and probability of a certain item response can be expressed by different estimation models for dichotomous (e.g. Rasch-model) or polytomous [(General Rating Model (GRM) or General Partial Credit Model (GPCM)) items (Embretson and Reise, 2000)]. In this approach, item and person parameters are estimated from the obtained data and considered as sample and test independent (Hambleton and Jones, 1993). In contrast to CTT, easy and hard tests should lead to the same person parameter.

The IRT-framework inherits capabilities to construct “computer-adaptive-tests” (CATs), which reach a high degree of precision and lower respondent burden by making use of large item banks – presenting only those items to a test-taker which are expected to yield information relevant to the range of the person’s ability. Examination of differential item functioning (DIF) for subgroups (e.g. gender

and age) or estimation of probabilities of specific response patterns to detect improbable response behavior is also possible (Reise and Waller, 2009).

Another advantage of IRT models is the inherent possibility of calibrating several questionnaires on a common scale (Embretson and Reise, 2000). This ensures comparability of ability estimates over different forms of one questionnaire, as well as over different questionnaires measuring the same construct and enhances the possibilities for comparing studies. In clinical contexts, the procedures for linking questionnaires were used, for example, with instruments measuring headache (Bjorner *et al.*, 2003b) or health-related quality of life in cancer patients (Holzner *et al.*, 2006).

In this paper, we focus on the depression scales from two questionnaires with a similar theoretical background, the widely used PHQ, with its short-forms (PHQ-9 and PHQ-2) and the recently developed ISR. Both instruments are based on conventions about symptoms indicating depression, which are formulated in DSM-IV (PHQ) and ICD-10 (ISR).

On the one hand, this study aims at comparing the ISR depression scale to the PHQ depression scales PHQ-9 and PHQ-2. We assessed this by examining the similarities and differences in both instruments for measuring depression in psychosomatic patients. According to IRT, this comparison qualifies as a validation study of the ISR depression scale.

On the other hand, we used the IRT approach to link both questionnaires on a common scale, providing data to enable the conversion of test scores. This should enhance the comparability of the studies using either one of the questionnaires. Predicted and observed scores were examined to assess the quality of our predictions.

Method

Instruments

The PHQ was developed as an instrument for screening and assessing presence, severity and temporal course of a total of eight mental and behavioral disorders (Spitzer *et al.*, 1999). The depression scale constitutes the short-form PHQ-9 and consists of nine items (1a to 1i) derived from the DSM-IV diagnostic criteria for a depressive disorder (Kroenke *et al.*, 2001). Use of only the first two (1a and 1b) items of the depression scale (PHQ-2) is also proposed for screening depression economically and parsimoniously in primary care (Löwe *et al.*, 2005). The PHQ-9 is considered as a valid instrument for the evaluation of depressive symptoms in medical care (Löwe *et al.*, 2004a; Löwe *et al.*, 2004b; Löwe *et al.*, 2004c) and is

available free of charge in a range of languages (Patient Health Questionnaire Screeners, <http://www.phqscreeners.com>). Items are rated across four categories and respondents are asked how often they were bothered by symptoms in the past two weeks. The items of the questionnaire are shown in Table 1.

The ISR is a recently developed, free of charge questionnaire for the evaluation of a broad range of mental or behavioral symptoms ratable by patients (Tritt *et al.*, 2008). The instrument focuses on assessing the presence and severity of mental stress and consists of five syndrome scales with three to four items each (depressive, anxiety, obsessive-compulsive, somatoform and eating disorders), as well as an additional scale covering 12 symptoms, which either occur in different syndromes (e.g. concentration difficulties), are clinically relevant (e.g. suicidality) or function as a first indicator of a specific syndrome [e.g. for post-traumatic stress disorder (PTSD)]. All items are derived from the diagnostic criteria of the ICD-10. The anticipated factor structure and suitability for measuring treatment outcomes has been substantiated in preceding studies (Fischer *et al.*, 2011; Fischer *et al.*, 2009). The ISR is available in German (Institut für Qualitätsentwicklung in der Psychotherapie und Psychosomatik, <http://www.iqp-online.de>) and Turkish (Kizilhan *et al.*, submitted for publication); translations of the instrument in several other languages are in preparation. Items are rated across five categories and respondents are asked about the extent

to suffer from the specific symptoms in the past two weeks. The ISR depression scale items are also presented in Table 1.

As external validation criteria, data from the depression-CAT (D-CAT) were used measuring depression based on the principles of IRT (Fliege *et al.*, 2005; Fliege *et al.*, 2009). It has been shown to measure depression with high precision (standard error of measurement amounts to less than 0.32, which is equal to a reliability greater than 0.90 in terms of CTT) and leads to low respondent burden by administering only those depression-related items out of a large item bank with 64 items that are expected to contain information highly relevant to assessing the severity of the respondent's depressive condition. In view of the large number of possible items used in D-CAT, they will not be listed here. To give an impression of the items of the D-CAT, note that 12 out of the 13 most frequently administered items (occurrence in more than 10% of all observations) cover depressive mood. Other aspects of depression, like fatigue, are barely covered by the D-CAT (Fliege *et al.*, 2009).

Sample

All three instruments were used for routine psychometric diagnostics at the Clinic for Internal Medicine, Department of Psychosomatic Medicine and Psychotherapy, Charité University of Medicine, Berlin. All consecutive inpatients and outpatients were asked to complete the questionnaires on a handheld computer at (a) first interview, as well as

Table 1 Item content and estimated item parameter

					Thresholds	
	Item	Content	Number of categories	Slope	Location (mean)	Location (range)
ISR	1	Depressed mood	5	2.36	0.31	−0.90 to 1.58
	2	Lack of joy	5	2.17	0.48	−0.52 to 1.53
	3	Lack of energy	5	1.29	0.08	−0.95 to 1.08
PHQ-9	4	Low self esteem	5	1.33	0.87	0.16 to 1.56
	1 ^a	Little interest	4	1.49	0.06	−1.14 to 0.68
	2 ^a	Feeling down, depressed or hopeless	4	2.73	0.26	−0.74 to 0.95
	3	Sleep disturbance	4	0.79	−0.45	−1.52 to 0.41
	4	Feeling tired or having little energy	4	1.65	−0.34	−1.62 to 0.40
	5	Poor appetite or overeating	4	0.71	0.37	−0.47 to 1.22
	6	Feeling bad about oneself	4	1.18	0.79	0.14 to 1.43
	7	Trouble concentrating	4	1.22	0.36	−0.70 to 1.01
	8	Moving or speaking slowly	4	0.80	1.12	0.37 to 1.73
	9	Thoughts of death	4	1.40	1.69	0.89 to 2.26

^aPHQ-2 items.

(b) at hospital admission and (c) discharge of inpatient treatment. In general, about 5% of the patients do not complete the psychometric evaluation at each time of assessment, mainly due to reading difficulties or language problems.

Since it is recommended to use a large and heterogeneous sample for the estimation of polytomous IRT models (Embretson and Reise, 2000), we decided to pool all available observations in one sample. Due to the orientation of our clinic, this sample covers a broad variety of patients with different somatic and mental symptoms. Depressive disorders are prevalent as main disorder as well as a comorbidity in 30% of our patients, including mild, moderate, severe and other episodes of depression as well as recurrent and persistent depressive disorders. This ensures coverage of the whole spectrum of depressive severity in clinical contexts and a sufficient number of responses in each item category and is necessary for precise estimation of item and person parameters.

The large sample size and pooling of data across different assessments could lead to deflated *p*-values and confidence intervals; we accepted this issue because the main results do not rely on *p*-values and confidence intervals.

The data set contained 4517 observations from a total of 2999 inpatients and outpatients. Repeated measurement mostly occurred with outpatients, who continued with inpatient treatment. Single measurement mainly took place with outpatients with no subsequent inpatient treatment. A total of 2148 patients completed the instruments once, while 851 patients completed the instruments several times. The ratio of 73% (*N* = 2181) women and 27% (*N* = 818) men reflects a relatively common gender distribution in German Psychosomatic Medicine patient samples (Tritt *et al.*, 2003). The mean age was 44.2 [standard deviation (SD) = 14.8] years, with ages ranging from 14 to 86 years. Mean ISR depression score in the sample is 1.59 (SD = 1.06, range = 0–4), mean PHQ-9 score is 10.56 (SD = 6.22, range = 0–27) and mean PHQ-2 score is 2.70 (SD = 1.84, range = 0–6).

The ICD-10 F diagnoses found in the sample according to clinical interviews were, in order of prevalence: affective disorders (F32–34 present in 30% of the cases), somatoform disorders (F45; 26%), reaction to severe stress and adjustment disorders (F43; 25%), eating disorders (F50; 22%), anxiety disorders (F40–41; 14%), mental and behavioral disorders due to psychoactive substance abuse (F1x; 12%), psychological and behavioral factors associated with disorders or diseases classified elsewhere (F54; 4%), dissociative disorders (F44; 4%) and others (2%).

No clinical diagnosis of a mental disorder (ICD-10 F) was given in 494 of 4517 cases (10.9%). Most frequently, only one (in 2113 cases, 46.8%) or two (in 1359 cases,

30.1%) F-diagnoses were given; in 551 cases (12.2%), three or more diagnoses were present.

A construction sample was used to check for unidimensionality and local independence, as well as to apply an item response model and derive the equation tables. A validation sample was then used to validate the results. The sample was randomly divided into a construction sample (*n* = 2258) and a validation sample (*n* = 2259). There were no differences in gender ($\chi^2 = 0.0006$, *df* = 1, *p* = 0.9806), mean age ($t = 0.25$, *df* = 4513, *p* = 0.80), ISR depression score ($t = 0.81$, *df* = 4515, *p* = 0.41), PHQ depression score ($t = 1.52$, *df* = 4515, *p* = 0.13) and D-CAT theta ($t = -0.27$, *df* = 2577, *p* = 0.79) between these two samples, which could effect validation.

Statistical analysis

The free software R 2.8.1 (R Development Core Team, 2008) was used for all statistical procedures, especially the packages nFactors (Raiche and Magis, 2009) for parallel analysis, sem: Structural Equation Models (Fox and Kramer, 2009) for confirmatory factor analysis (CFA) and ltm (Rizopoulos, 2006) for IRT model fitting.

Unidimensionality and local independence

We first ensured unidimensionality and local independence of the 13 items from both questionnaires as a requirement for IRT analysis (Orlando, 2004). We addressed this issue by using a factor analytic approach. First, we checked the ratio of eigenvalues and conducted a parallel analysis to determine the number of extractable factors. CFA for several unidimensional models was then conducted, using the method of maximum likelihood, assuming multinormal errors (Fox and Kramer, 2009). Goodness of fit was examined following general guidelines (Brown, 2006). Additionally, the modification indices as well as the matrix of standardized residuals were examined.

IRT model

Next, we applied an IRT model. Response categories in both instruments were polytomous and ordered. Since the items were derived from different questionnaires and had different response categories, we chose the General Partial Credit Model (Muraki, 1992) allowing items to have different thresholds and slopes for appropriate representation of the data. Using this model, we were then able to estimate a person's value of the measured underlying construct (in our case, depression severity).

Evaluating the goodness of fit of polytomous IRT-models is crucial. While some authors argue that a large

sample size is sufficient to ensure adequate goodness of fit (Embretson and Reise, 2000; Fliege *et al.*, 2005), some fit indices have also been suggested in this context (e.g. $S-G^2$ or $S-X^2$) (Bjorner *et al.*, 2007). According to McDonald (1989), a large sample size could compromise fit indices. Given the rather large sample size, we decided to neglect use of the goodness of fit indices after checking for unidimensionality and local independence. Instead use of the construction sample and the validation sample is an approach used to ensure generalizability. While the model derived from the construction sample was used for further calculations, we compared item slopes and thresholds from the construction sample with item slopes and thresholds estimated from the validation sample (Lord, 1980).

Respondents' scores of depression severity and measurement error were calculated in the validation sample by using information provided solely from one questionnaire (ISR theta, PHQ-9 theta, PHQ-2 theta) or by use of all information (Overall theta) applying an empirical Bayes method (Rizopoulos, 2006). This enabled us to compare the quality of measurement of the questionnaires. Correlations of these theta values to classical sum scores and to D-CAT as an external criterion are presented.

Equating

The overall IRT model provides the probabilities for each answer category of each item dependent on the latent trait. Therefore, it is possible to calculate the expected value of every item answer (sum of answer probabilities $\times$ answer value) and hence of the whole original scale (here: PHQ-9, PHQ-2 and ISR) based on the calculated latent trait. Using these expected values, it is possible to create conversion tables with the corresponding sum scores. For a certain score in instrument 1 (e.g. five points in PHQ-9), there is an interval of corresponding theta values (e.g. -1.0 to -0.3 with stepwidth 0.1). These theta values correspond to certain scores in the second instrument (e.g. 0.75 to 1.25 in ISR). The expected score was then calculated as the mean of the expected scores of the second instrument (i.e. ISR scores). To validate the conversion, we predicted ISR scores from PHQ-9 (respectively PHQ-2 scores) and vice versa in the validation sample and compared them to the observed PHQ-9, PHQ-2 and ISR scores.

Results

Unidimensionality and local independence

The first eigenvalue of the correlation matrix of the 13 items is 6.99 and is substantially greater than the second eigenvalue, which is 1.00. The first factor accounts for

54% of the total variance. The second eigenvalue (1.00) does not exceed 95% of second eigenvalues derived from random correlation matrices in parallel analysis; it can therefore be viewed as being random (Brown, 2006). In exploratory factor analysis, this result would strongly suggest a one-factor solution.

CFA was initially conducted for ISR and PHQ-9 separately with one latent factor. With two items not being sufficient to set up an identified latent factor model (Brown, 2006), a model consisting only of the PHQ-2 items could not be evaluated. Unidimensionality of PHQ-2 was therefore assumed through showing unidimensionality of PHQ-9. Subsequently, a one-factor model for all 13 items was tested. A second model with two correlated factors was also tested to examine possible questionnaire-related differences. To examine goodness of fit, measures for absolute fit (standardized root mean square residual, SRMR), fit adjusted for model parsimony (root mean square error of approximation, RMSEA) and comparative fit (comparative fit index, CFI) are proposed (Brown, 2006). Goodness of fit measures dependent on sample size (such as χ^2/df and goodness of fit/adjusted goodness of fit, GFI/AGFI) were not used. Rules of thumb suggest that SRMR values < 0.08 , RMSEA values < 0.08 and CFI > 0.95 indicate adequate model fit (Brown, 2006; Schermelleh-Engel *et al.*, 2003).

The results (see Table 2) indicate a good fit for a unidimensional model of the ISR depression scale and an acceptable fit for the PHQ-9 depression scale. Both combined models had strong SRMR values for absolute fit, whereas RMSEA and CFI had poor fit. Standardized estimates of factor loadings in the one-factor model ranged between 0.54 for PHQ-9, item 8 (experience of slow-downs or restlessness) and 0.85 for item ISR, item 1 (depressed mood).

In the two-factor model, both factors correlated very highly (0.95) and the goodness of fit measures were comparable to the one-factor model. This suggests that no information is gained by adding one factor for each instrument.

Modification indices revealed the strongest connections between ISR, item 4 and PHQ-9, item 6, as well as between ISR, item 3 and PHQ-9, item 4. Both item pairs focus on nearly identical aspects of depression (lack of self esteem, resp. lack of energy). In order to prevent local dependency, removing one of two such similar items is recommended (Orlando, 2004). After removal of PHQ-9, item 6 and ISR, item 3, achieved goodness of fit was clearly better. For our later purposes, excluding items was not possible, therefore it was necessary to ensure that neither one of the item pairs became the definition of the latent trait (Orlando, 2004). Given the remaining number of items, this was easily managed.

Table 2 Results of CFA

CFA model	Number of factors	Number of items	SRMR	RMSEA	CFI
ISR	1	4	0.007	0.037	0.99
PHQ-9	1	9	0.041	0.093	0.94
ISR and PHQ-9	1	13	0.050	0.116	0.88
ISR and PHQ-9	2	13	0.049	0.114	0.89
ISR and PHQ-9 ^a	1	11	0.038	0.087	0.94

Note: SRMR, standardized root mean square residual; RMSEA, root mean square error of approximation; CFI, comparative fit index.

^aRemoved items.

The residual correlations for the model show values between -0.09 and 0.23 , with the highest correlations for the earlier-mentioned item pairs with almost identical content. The cutoff-criteria indicating local dependence proposed in other studies are 0.20 (Bjorner *et al.*, 2003a) or 0.25 (Fliege *et al.*, 2005).

Considering these results and the common theoretical background of both depression scales, unidimensionality and local independence can be assumed. Therefore, the data can be viewed as suitable for IRT analysis.

IRT model

The GPCM was applied to the data from the construction sample to estimate item parameters (Table 1). Slope parameters indicated an item discrimination range between 0.71 and 2.73 . The lowest slopes were observed for items focused on the somatic symptoms of depression (poor appetite, sleep and moving or speaking slowly). Items with a possible local dependency (ISR item 3 and PHQ-9 item 4,

resp. ISR item 4 and PHQ-9 item 6) did not show exceeded slopes. Exceeded slopes would indicate that one of the item pairs became the definition of the latent trait, which is at risk with local dependency (Orlando, 2004). The original PHQ-9-scale contains items with more extreme difficulties, as indicated by the locations (threshold means) and ranges. Slopes and thresholds of the IRT model correlated with the slopes and thresholds from an equal model derived from the validation sample ($r = 0.99$) and can therefore be assumed to be valid.

The response curves for the ISR items indicate good characteristics with steep and separable trace lines, response options in the correct order and one response curve exceeding every other option in only one of the defined ranges for depressive symptom severity. The response curves for the PHQ-9 items showed less desirable characteristics. The probability functions for the third response category ("More than half the days") did not exceed the probability functions for categories 2 and 4 in most items (PHQ-9 items 1, 3, 5, 6, 7, 8 and 9). Figure 1 shows item

Figure 1 Exemplary item response functions.

response curves for ISR item 1 (revealing good characteristics, exemplary for all ISR items), PHQ-9 item 2 (showing good characteristics, exemplary for PHQ-9 items 2 and 4) and PHQ-9, item 3 (with less desirable characteristics, exemplary for PHQ items 1, 3, 5, 6, 7, 8 and 9).

Figure 2 shows the standard error of measurement indicating measurement precision for the PHQ-2, PHQ-9 and ISR depression scales dependent on the latent trait. The measurement precision when using information from both questionnaires (13 items) is also shown, in addition to the cutoff-criterion for D-CAT (Fliege *et al.*, 2005). Due to IRT prerequisites, about 95% of person's parameters are located between -2 and 2 . Not surprisingly, the short PHQ-2 is the most imprecise of the three static questionnaires. Within a theta range from -1 to 2 , PHQ-9 and ISR show similar measurement error and are close to precision

Figure 2 Standard errors of measurement dependent on latent trait.

achieved by the D-CAT, which is equal to a reliability of 0.9 in terms of CTT (Embretson and Reise, 2000). PHQ-9 is also able to measure theta values < -1 and over 2 . Accordingly, use of the information from all 13 items provides more precise measurement over the whole scale.

Table 3 shows correlations between classical sum scores, estimated theta values from different item sets according to the questionnaires and the D-CAT score in the validation sample as an external criterion. Classical sum scores and estimated theta values show correlations near one (Table 3, shown in italic typeface). A correlation of 0.85 was found for estimated thetas from the four ISR items and the nine PHQ-9 items. Differences between theta estimates by ISR and PHQ-9 are distributed around zero (mean = 0.03, SD = 0.48). In 77% of the 2259 cases, the absolute value of the difference is below or equal to 0.5.

Equating

In order to enable easy transformation between the two questionnaires, Tables 4 and 5 present the range of possible PHQ-9, PHQ-2 and ISR scores, together with the corresponding expected scores from the other questionnaire. As an example, persons who achieved a PHQ-9 score of seven are expected to achieve an ISR-score of 1.00, based on their calculated latent trait.

In order to substantiate the validity of the conversion tables, we compared the converted scores with the actual scores in the validation sample. For all 2259 cases, the PHQ-9 (respectively PHQ-2) scores were converted to ISR scores and ISR scores were converted to PHQ-9 (respectively PHQ-2) scores. Figure 3 shows the converted PHQ-9 (respectively PHQ-2) scores, the converted ISR scores and the means of the actual scores of the instruments, as well

Table 3 Correlations of depression scores and estimates

		CTT sum scores (<i>n</i> = 2259)			IRT theta values (<i>n</i> = 2259)				External criterion (<i>n</i> = 1284)
		ISR	PHQ-9	PHQ-2	ISR	PHQ-9	PHQ-2	Overall	D-CAT theta
CTT	ISR	1	0.83	0.77	<i>0.99</i>	0.85	0.79	0.94	0.77
	PHQ-9		1	0.85	0.83	<i>0.98</i>	0.85	0.94	0.72
	PHQ-2			1	0.78	0.89	<i>0.98</i>	0.86	0.70
IRT	ISR				1	0.85	0.80	0.95	0.79
	PHQ-9					1	0.91	0.97	0.75
	PHQ-2						1	0.89	0.73
	Overall							1	0.80

All correlations significant with *p* < 0.001.

Table 4 Conversion tables

ISR	Expected PHQ-9	Expected PHQ-2
0.00	0.91	0.00
0.25	3.00	0.57
0.50	4.75	1.00
0.75	6.00	1.00
1.00	7.33	2.00
1.25	8.50	2.00
1.50	9.50	2.50
1.75	11.50	3.00
2.00	12.50	3.00
2.25	14.50	4.00
2.50	16.00	4.00
2.75	17.50	5.00
3.00	19.00	5.00
3.25	20.67	5.33
3.50	22.33	6.00
3.75	24.29	6.00
4.00	26.14	6.00

as intervals which contain about 66% (mean $\pm$ 1 SD) and 95% (mean $\pm$ 2 SD) of the observed scores. Means and SDs were calculated for each value of the questionnaire. For example, a PHQ-9 score of seven was actually found in 128 cases and correspond to a converted ISR score of 1.00. The mean of the observed ISR scores in these 128 cases was 1.10, with a standard deviation of 0.58. Note that for most categories, sample size was greater than 50, but highest sum scores (PHQ-9 > 21; ISR $\geq$ 3.5) were observed less often (PHQ-9: 22 to 27 points in 42, 30, 22, 18, 3 resp. 4 cases; ISR: 3.50 to 4.00 points in 48, 40 resp. 42 cases).

Discussion

In this study, we used IRT methods to equate the depression scales of the PHQ-9, PHQ-2 and the ISR to provide easy-to-use conversion tables. Viewed in the light of IRT, the statistical analyses of this study also provide validation for the more-recently developed ISR depression scale on the basis of comparison to the well-established PHQ-9 scale.

The examination of unidimensionality and local independence as a requirement for the use of unidimensional IRT-models yielded one underlying factor responsible for item responses. The PHQ-9 and ISR depression scales can therefore be viewed as measuring the same construct. Based on the item contents (diagnostic criteria for depressive disorders in ICD-10 or DSM-IV), the construct can be identified as "depression". It was also demonstrated that

Table 5 Conversion tables

	Expected ISR
<i>PHQ-2</i>	
0	0.03
1	0.45
2	1.17
3	1.80
4	2.38
5	3.00
6	3.84
<i>PHQ-9</i>	
0	0.00
1	0.00
2	0.08
3	0.25
4	0.33
5	0.50
6	0.75
7	1.00
8	1.12
9	1.38
10	1.50
11	1.75
12	1.88
13	2.00
14	2.25
15	2.25
16	2.50
17	2.75
18	2.75
19	3.00
20	3.25
21	3.25
22	3.50
23	3.62
24	3.75
25	3.81
26	4.00
27	4.00

different answer formats between both questionnaires did not affect unidimensionality. Two pairs of items were identified to possibly cause local dependency, which was refuted in later IRT analyses (Orlando, 2004).

Although some multidimensional IRT models are available, most researchers use unidimensional models as we did. In fact, this is a limiting factor in measuring depression, e.g. because some aspects of depression like somatic symptoms might not be covered in a unidimensional model. While there was empirically no doubt in unidimensionality of the questionnaires in this study, which can be ascribed

Figure 3 Evaluation of conversion tables. Legends: ○, expected score; —, mean observed score; - - -, mean ± 1 SD or mean ± 2 SD.

to the similar theoretical and conceptual background of both instruments, equating between questionnaires covering broader aspects of depression might require a more advanced methodological procedure.

The applied GPCM (Muraki, 1992) produced nearly identical slope and threshold parameters in the construction and the validation samples, indicating that the item's measurement characteristics are very stable (Lord, 1980). While the PHQ-9 contained a broader variety of item difficulties and discrimination parameters than the ISR, its response format showed drawbacks for most items. At no point of the scale did the third response category ("More than half the days") have the highest probability of all response categories to be chosen. It is likely that a distinction between the second and fourth response categories does not reflect any further refinements by the

respondents. The ISR response format showed overall good characteristics (Embretson and Reise, 2000).

The standard error of measurement in the IRT, which is similar to reliability assessment in the CTT, was almost equal for ISR and PHQ-9 across a wide range of depression severities and as precise as a computer-adaptive-test (D-CAT) for the middle range of depression severity. The PHQ-9 is able to differentiate better between respondents with a low ($\theta < -1$) or very high ($\theta > 2$) latent trait in comparison to the ISR, as can be expected for an extremely short scale. The PHQ-2 measured depression with higher standard error and provides less information. For screening purposes however, the gain of lower respondents' burden may compensate for the loss of information.

Correlations between classical sum scores and theta estimates reached almost one, which is in accordance with

similar studies (Embretson and Reise, 2000; Reise and Waller, 2009). The theta estimates correlated a little higher to the external criterion in comparison to classical sum scores and the ISR correlated a little higher than the PHQ-9. This little difference pertaining to concordance of the instruments provides no information about differences of the questionnaires' diagnostic accuracy and might result from differences in item concordance. These highly similar correlations point to the question "What is the advantage of using an IRT model?" It should be kept in mind that with IRT methods, persons and items parameters are located on the same scale for both instruments and can be compared more easily and more precisely (e.g. person one theta value estimated with ISR items and person two theta value estimated with PHQ-9 items are directly comparable). However, the measurement model of IRT reflects our underlying assumptions of a person's trait or state, which is responsible for response behavior (Embretson and Reise, 2000).

All these findings (similar standard error of measurement, high correlation to PHQ-9, PHQ-2 and D-CAT) suggest that the ISR depression scale with four items measures depression almost as well as the PHQ-9 depression scale with nine items. Given the similar orientation of both questionnaires based on DSM-IV and ICD-10, this finding may seem trivial, but analysis revealed that the higher degree of parsimony of the ISR scale is not bought with a loss of precision within a broad range of the latent trait continuum. Therefore, it can be concluded that the ISR depression subscale is suitable for assessing depressive symptoms. Thus in clinical practice it seems to suffice using only one instead of both questionnaires. The advantage of the PHQ-9 lies in having a slightly more precise measurement of depressive severity, while the overall ISR addresses a broader variety of symptoms appearing in anxiety, compulsive, somatoform and eating disorders (Tritt *et al.*, 2008). In addition, the ISR with 29 items is substantially shorter than the complete PHQ with 78 items.

In this study no standardized diagnostic interviews were applied. In light of the required large sample sizes, which would be accompanied by an immense workload if all the patients in routine care would receive standardized interviews, this decision seems justifiable. Momentarily we are working on a comparison of the ISR and standardized diagnostic interviews in selected sample, to examine the diagnostic accuracy of the ISR in future research.

The provided conversion tables are derived from the expected classical sum scores for PHQ-9, PHQ-2 and ISR. They should enable users of one instrument to translate scores into the other. This may also enhance comparability between studies.

The expected and mean observed scores in the validation sample proved to be very similar; therefore, the assumed IRT model demonstrated its suitability. Conversions using the short PHQ-2 scale were more imprecise than conversions using PHQ-9 scale. Substantial statistical spread was observed, making comparison of individual scores imprecise (e.g. a PHQ-9 score of 15 corresponds with an ISR score of 2.25; note that given a PHQ-9 score of 15, approximately 95% of observed ISR scores lie between 1.00 and 3.75). This is a finding in accordance to prior equating studies (Holzner *et al.*, 2006). The SD decreases with sample size by the square root of sample size, so that the conversion of sum scores on the group level would be more exact. The biggest differences between the observed and expected scores were found for extremely low and high values. We interpret this finding to be the result of a general tendency towards the middle of the range. Extreme values were observed less often – and even if a person reached highest values in one questionnaire, it is unlikely that this person also scored that high in the other questionnaire. Therefore, theoretical model fit at the ends of the continuum could be less precise.

Based on a large sample of psychosomatic patients with a wide variety of somatic and mental symptoms and diseases, the conversion tables showed suitability in this domain. Further validation in different samples (e.g. patients in primary care) should be conducted.

Some questions remain for further research. While earlier studies demonstrated the positive dimension structure as well as stable measuring characteristics over time (Fischer *et al.*, 2011; Fischer *et al.*, 2009), further evaluation of the anxiety, obsessive-compulsive, somatoform and eating disorder scales of the ISR, e.g. through comparison with other established questionnaires, has yet to be accomplished.

Initially, the ISR was developed to overcome certain difficulties of other scales; for example, the insufficient replication of factor structure (Cyr *et al.*, 1985; Schmitz *et al.*, 2000) of the SCL-90 (Derogatis and Savitz, 1999), which is a widely used instrument to assess mental illness. Therefore, a comparison of the SCL and ISR scales in terms of IRT models would be useful for clinical applications.

Finally, a more methodological issue concerns the extent to which equating questionnaires by calibrating the scores on a common scale is more helpful in applied research than the use of a linear regression estimation of scores. This might be addressed in future research by a direct comparison of these two methodologies.

Acknowledgements

No funding was granted for this study. We reproduced no material which has appeared in another publication.

Declaration of interest statement

The authors have no competing interests.

References

- American Psychiatric Association. (2000) *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., Text Revision), Washington, DC, American Psychiatric Association.
- Bjorner J.B., Kosinski M., Ware J.E. (2003a) Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT). *Quality of Life Research*, **12**(8), 913–933. DOI: 10.1023/A:1026163113446
- Bjorner J.B., Kosinski M., Ware J.E. (2003b) Using item response theory to calibrate the Headache Impact Test (HIT-6) to the metric of traditional headache scales. *Quality of Life Research*, **12**(8), 981–1002. DOI: 10.1023/A:1026123400242
- Bjorner J.B., Smith K.J., Stone C., Sun X. (2007) *IRTFIT: A Macro for Item Fit and Local Dependence Tests under IRT Models*, Lincoln, RI, QualityMetric. http://outcomes.cancer.gov/areas/measurement/irt_model_fit.html.
- Brown T. (2006) *Confirmatory Factor Analysis for Applied Research* (1st ed.), New York, The Guilford Press.
- Cyr J.J., McKenna-Foley J.M., Peacock E. (1985) Factor Structure of the SCL-90-R: Is there one? *Journal of Personality Assessment*, **49**(6), 571–578.
- Derogatis L.R., Spitzer K.L. (1999) The SCL-90-R, brief symptom inventory and matching clinical rating scales. In *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment* (2nd ed.), Maruish M.E. (ed) Mahwah, NJ, Lawrence Erlbaum Associates.
- Dilling H., Mombour W., Schmidt M.H. (2005) *Internationale Klassifikation psychischer Störungen. ICD-10, Kapitel V (F). Klinisch-diagnostische Leitlinien*, Bern, Huber.
- Edelen M.O., Reeve B.B. (2007) Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, **16**(Supplement 1), 5–18. DOI: 10.1007/s11136-007-9198-0
- Embretson S.E., Reise S.P. (2000) *Item Response Theory for Psychologists*, Mahwah, NJ, Lawrence Erlbaum.
- Fischer H.F., Schirmer N., Tritt K., Klapp B.F., Fliege H. (2011) Retest–Reliabilität und Änderungssensitivität des ICD-10-Symptom-Rating (ISR) in verschiedenen Stichproben. *Psychotherapie, Psychosomatik, Medizinische Psychologie*, **61**(3/04), 162–169. DOI: 10.1055/s-0030-1249683
- Fischer H.F., Tritt K., Klapp B.F., Fliege H. (2009) Faktorstruktur und psychometrische Eigenschaften des ICD-10-Symptom-Rating (ISR) an Stichproben psychosomatischer Patienten. *Psychotherapie, Psychosomatik, Medizinische Psychologie*, **8**. DOI: 10.1055/s-0029-1214419
- Fliege H., Becker J., Walter O.B., Bjorner J.B., Klapp B.F., Rose M. (2005) Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, **14**(10), 2277–2291. DOI: 10.1007/s11136-005-6651-9
- Fliege H., Becker J., Walter O.B., Rose M., Bjorner J.B., Klapp B.F. (2009) Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. *International Journal of Methods in Psychiatric Research*, **18**(1), 23–36. DOI: 10.1002/mpr.274
- Fox J., Kramer A. (2009) *sem: Structural Equation Models* (R package version 0.9-17). <http://cran.r-project.org/web/packages/sem/index.html>
- Hambleton R.K., Jones R.W. (1993) Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practices*, **12**(3), 38–47.
- Holzner B., Bode R.K., Hahn E.A., Cella D., Kippa M., Sperner-Unterwieser B., Kemmler G. (2006) Equating EORTC QLQ-C30 and FACT-G scores and its use in oncological research. *European Journal of Cancer*, **42**(18), 3169–3177. DOI: 10.1016/j.ejca.2006.08.016
- Kizilhan J.I., Roniger A., Tritt K., von Heymann F. (Submitted for publication) Psychometrische Überprüfung der türkische Version des ICD-10-Symptom-Ratings.
- Kroenke K., Spitzer R.L., Williams J.B.W. (2001) The PHQ-9 – validity of a brief depression severity measure. *Journal of General Internal Medicine*, **16**(9), 606–613. DOI: 10.1046/j.1525-1497.2001.016009606.x
- Lord F.M. (1980) *Applications of item response theory to practical testing problems*, Hillsdale, NJ, Erlbaum.
- Löwe B., Kroenke K., Herzog W., Gräfe K. (2004b) Measuring depression outcome with a brief self-report instrument: Sensitivity to change of the Patient Health Questionnaire (PHQ-9). *Journal of Affective Disorders*, **81**(1), 61–66. DOI: 10.1016/S0165-0327(03)00198-8
- Löwe B., Üntzler J., Callahan C.M., Perkins A.J., Kroenke K. (2004c) Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical Care*, **42**(12), 1194–1201.
- Löwe B., Gräfe K., Zipfel S., Witte S., Loecherer W. (2004a) Diagnosing ICD-10 depressive episodes: Superior criterion validity of the Patient Health Questionnaire. *Psychotherapy and Psychosomatics*, **73**(6), 386–390. DOI: 10.1159/000080393
- Löwe B., Kroenke K., Gräfe K. (2005) Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *Journal of Psychosomatic Research*, **58**(2), 163–171. DOI: 10.1016/j.jpsychores.2004.09.006
- McDonald R.P. (1989) Future directions for item response theory. *International Journal of Educational Research*, **13**(2), 205–220.
- Muraki E. (1992) A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, **16**(2), 159–176. DOI: 10.1177/014662169201600206
- Orlando M. (2004) *Critical Issues to Address when Applying Item Response Theory (IRT) Models. DIA and Advances in Health Outcomes Measurement*. <http://www.outcomescancergov/conference/irt>
- R Development Core Team. (2008) *R: A Language and Environment for Statistical Computing*, Vienna, R Foundation for Statistical Computing.
- Raiche G., Magis D. (2009) *nFactors* (R package version 2.3.1). <http://cran.r-project.org/web/packages/nFactors/index.html>
- Reise S.P., Waller N.G. (2009) Item Response Theory and Clinical Measurement. *Annual Review of Clinical Psychology*, **5**, 27–48. DOI: 10.1146/annurev.clinpsy.032408.153553
- Rizopoulos D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, **17**(5), 1–25.
- Schermelleh-Engel K., Moosbrugger H., Müller H. (2003) Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, **8**(2), 23–74.
- Schmitz N., Hartkamp N., Kiuse J., Franke G.H., Reister G., Tress W. (2000) The Symptom Check-List-90-R (SCL-90-R): A German

- validation study. *Quality of Life Research*, **9**(2), 185–193. DOI: 10.1023/A:1008931926181
- Spitzer R.L., Kroenke K., Williams J.B.W. (1999) Validation and utility of a self-report version of PRIME-MD. *JAMA: The Journal of the American Medical Association*, **282**(18), 1737–1744. DOI: 10.1001/jama.282.18.1737
- Tritt K., Heymann F., Loew T.H., Benker B., Bleichner F., Buchmüller R., Findeisen P., Galuska J., Kalleder W., Lettner F., Michelitsch B., Pfitzer F., Stadtmüller G., Zaudig M. (2003) Patienten in stationärer psychosomatischer Krankenhausbehandlung: Patientencharakterisierung und Behandlungsergebnisse anhand der Psy-BaDo-PTM. *Psychotherapie in Psychiatrie, Psychotherapeutische Medizin und Klinischer Psychologie*, **8**(2), 244–251.
- Tritt K., Heymann F., Zaudig M., Zacharias L., Söllner W., Loew T.H. (2008) Entwicklung des Fragebogens 'ICD-10-Symptomrating' (ISR). *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, **54**(4), 425–434.
- Zickar M.J., Broadfoot A.A. (2008) The partial revival of a dead horse? Comparing classical test theory and item response theory. In *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in Organizational and Social Sciences*, Lance C.E., Vandenberg R.J. (eds) pp. 37–60, New York, Routledge Academic.