International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 19(4): 211–222 (2010)
Published online 19 July 2010 in Wiley Online Library
(wileyonlinelibrary.com) **DOI**: 10.1002/mpr.320

Evaluation of the DSM-IV and ICD-10 criteria for depressive disorders in a community population in Japan using item response theory

MARI SAITO,[1] NOBORU IWATA,[2] NORITO KAWAKAMI,[3] YUTAKA MATSUYAMA[4] & WORLD MENTAL HEALTH JAPAN 2002–2003 COLLABORATORS (YUTAKA ONO,[5] YOSHIBUMI NAKANE,[6] YOSHIKAZU NAKAMURA,[7] HISATERU TACHIMORI,[8] HIDENORI UDA,[9] HIDEYUKI NAKANE,[10] MAKOTO WATANABE,[7] YOICHI NAGANUMA,[8] TOSHIAKI A. FURUKAWA,[11] YUKIHIRO HATA,[12] MASAYO KOBAYASHI,[7] YUKO MIYAKE,[8] TADASHI TAKESHIMA[8] & TAKEHIKO KIKKAWA[13])

1 Clinical Research Center, National Center for Child Health and Development, Tokyo, Japan
2 Department of Clinical Psychology, Hiroshima International University, Higashi-Hiroshima, Japan
3 Department of Mental Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
4 Department of Epidemiology and Biostatistics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
5 Health Center, Keio University, Tokyo, Japan
6 Division of Human Sociology, Nagasaki International University Graduate School, Nagasaki, Japan
7 Department of Public Health, Jichi Medical School, Saitama, Japan
8 National Institute of Mental Health, National Center of Neurology and Psychiatry, Tokyo, Japan
9 Director General of the Health, Social Welfare, and Environmental Department, Osumi Regional Promotion Bureau, Kagoshima, Japan
10 Division of Neuropsychiatry, Department of Translational Medical Sciences, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan
11 Department of Psychiatry and Cognitive-Behavioral Medicine, Nagoya City University Graduate School of Medical Sciences, Nagoya City, Japan
12 Department of Psychiatry, Field of Social and Behavioral Medicine, Kagoshima University Graduate School of Medical and Dental Sciences, Kagoshima City, Japan
13 Department of Human Well-Being, Chubu Gakuin University, Seki City, Japan

Key words
depression, World Mental Health Japan Survey, DSM-IV, ICD-10, item response theory

Abstract
The DSM-IV and ICD-10 are both operational diagnostic systems that classify known psychological disorders according to the number of criteria symptoms. Certain discrepancies between the criteria exist and may lead to some inconsistencies in psychiatric research. The purpose of this study was to investigate these

Correspondence
Mari Saito, Clinical Research Center, National Center for Child Health and Development, Tokyo, Japan; 2-10-1 Ookura Setagaya-ku, Tokyo 157-8535, Japan.
Telephone (+81)-3-3416-0181
Email: saito-mr@ncchd.go.jp

Received 2 July 2008;
revised 4 December 2008;
accepted 19 January 2009

differences in the assessment of depression with item response theory (IRT) analyses. The World Mental Health-Japan (WMHJ) Survey is an epidemiological survey of the general population in Japan. We analyzed data from the WMHJ completed by 353 respondents who had either depressive mood or diminished interest. A two-parameter logistic model was used to evaluate the characteristics of the symptoms of the DSM-IV and ICD-10. IRT analyses revealed that the symptoms about psychomotor activity, worthlessness and self-reproach were more informative and suggestive of greater severity, while the symptoms about dietary habits were less informative. IRT analyses also revealed that the ICD-10 seems more sensitive to the mild range of the depression spectrum compared to the DSM-IV. Although there were some variations in severity among respondents, most of the respondents diagnosed with a severe or moderate depressive episode according to the ICD-10 were also diagnosed with a major depressive episode according to the DSM-IV.

Introduction

Mood disorders are the second most common group of mental disorders, with the 12-month Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) prevalence reported to vary from 0.8% in Africa to 9.6% in the United States (World Health Organization [WHO] World Mental Health [WMH] Survey Consortium, 2004). Among these disorders, depression has been shown to be associated with an increased mortality risk of nearly double in a meta-analysis of 25 community surveys (Cuijpers and Smit, 2002); further, it has been described as the leading cause of disability (Üstun *et al.*, 2004). The point and lifetime prevalence of major depression in Japan are reported to be 0–2.7% and 2.9–6.6%, respectively, using DSM-III-R (American Psychiatric Association, 1987; Kawakami *et al.*, 2004; Otsubo, 2005). The variability of these estimates is attributed to many factors, including not only the research population but differences in the diagnostic systems applied.

The World Health Organization (WHO) has established the World Mental Health (WMH) Survey, which aims to address the prevalence and treatment of mental disorders all over the world (WHO WMH Survey Consortium, 2004). The WMH Survey is comprised of many questions used to make diagnoses according to the major criteria of the DSM-IV (American Psychiatric Association, 1994) and the International Classification of Disease-10 Classification of Mental and Behavioral Disorders (ICD-10) (WHO, 1992).

The DSM-IV and ICD-10 are both operational diagnostic systems that attempt to classify all known psychological disorders according to the number of criteria symptoms that are simultaneously present and their adverse impacts on social functioning. However, there exist certain discrepancies between the DSM-IV and ICD-10 criteria in making diagnoses. The psychopathological phenomena specified by a particular diagnosis may therefore differ according to which system is used, and thus the severities of the symptoms of patients, even among those with corresponding diagnoses and prevalences, can vary greatly. This situation leads to inconsistencies and even contradictions in psychiatric research, and thus provides a clear rationale to investigate their differences in diagnosed entities in detail.

Only a few studies have compared the DSM-IV and ICD-10 for depression. Faravelli *et al.* (1996) compared the diagnoses of these systems against other indexes and showed that the number of symptoms reflects the severity of depression. Vilalta-Franch *et al.* (2006) compared the prevalence of depression in patients with Alzheimer's disease according to the ICD-10, DSM-IV, and other classification systems. Besides depression, several studies have addressed the prevalence or severity of eating disorders (Fontenelle *et al.*, 2005), hyperkinetic disorders (Lahey *et al.*, 2006), dysthymic disorders (Han *et al.*, 1995; Lopez Ibor *et al.*, 1994) and panic disorders (Ietsugu *et al.*, 2007) among patients in comparisons of the DSM-IV and ICD-10 diagnostic systems. No study, however, has ever addressed the quantitative comparability of the characteristics of these systems, including a detailed inspection of individual symptoms.

Item response theory (IRT) is a psychometric theory representing mathematical functions that relate person and item parameters to the probability of the responses on a discrete outcome (Boeck and Wilson, 2004; Lord and Novick, 1968). IRT can quantify item level, item sensitivity

and respondent ability level. An IRT model treats items as ordered indicators of risk rather than as a count, and it assumes that the severity of a respondent follows the normal distribution on the same latent trait continuum scale. The model has been applied to questionnaire data in medical settings for the validation or shortening of major questionnaires, such as the Hamilton Rating Scale (Gibbons *et al.*, 1982) and SF-36 (Haley *et al.*, 1994; McHorney *et al.*, 1997).

Some previous studies have used IRT to investigate depression scales. Aggen *et al.* (2005) analyzed the DSM-IV criteria for depression with IRT and reported that these symptoms had different relative levels of severity and case discrimination parameters on the same latent trait continuum. Simon *et al.* (2002) also analyzed the DSM-IV for depression with IRT, demonstrating that there were no differences in the patterns of symptom severity among high, middle and low prevalence countries. Simon also found that some items in the DSM-IV were less informative as indicators of depression in patients with chronic medical illness (Simon and Von Korff, 2006). Although such investigations of the DSM-IV are available, no attempt has been made to compare the DSM-IV and ICD-10.

The purposes of this study are to analyze the item characteristics of symptoms for depression according to the DSM-IV and ICD-10 using IRT, and to clarify the relationships between individual symptoms based on cross-sectional WMHJ (WMHJ 2002–2004) survey data. We also compare the two scales through respondent diagnoses and severities estimated with IRT.

Method

Sample and data collection

The WMHJ is an epidemiological survey of Japanese-speaking household residents aged 20 and older. Seven community populations in Japan were selected as study sites in 2002–2004. These sites included two urban cities (Okayama City [population 660 000] and Nagasaki City [population 450 000]) and five rural municipalities (Kushikino City [population 25 000], Fukiage Town [population 8500], Ichiki Town [population 7000], and Higashiichiki Town [population 14 000] in Kagoshima Prefecture and Tamano City [population 70 000] in Okayama Prefecture). These sites were selected considering their geographic variation, availability of site investigators, and local government cooperation. A total of 2436 respondents completed a face-to-face interview. The total averaged response rate was 58.4%, calculated as the proportion of these respondents among eligible subjects ($N = 4173$) excluding those who had died, moved, or become institutionalized ($N = 408$) from the time of the initially selected sample ($N = 4581$); the respective study site response rates were as follows: Okayama City, 65.7%; Nagasaki City, 26.4%; Kushikino City, 65.7%; Fukiage Town, 81.6%; Ichiki Town, 71.2%; Higashiichiki Town, 69.8%; and Tamano City, 56.4%. Non-respondents tended to be male and younger (Kawakami *et al.*, 2005), and thus the present sample more reflected the characteristics of older and female populations. Written consent was obtained from each respondent at each site. The Human Subjects Committees of Okayama University (for the Okayama City and Tamano City sites), the Japan National Institute of Mental Health (for the four sites in Kagoshima Prefecture), and Nagasaki University (for the Nagasaki City site) approved the recruitment, consent, and field procedures. More detailed information on these field procedures are described elsewhere (Kawakami *et al.*, 2005).

The WMHJ surveys used the WHO Composite International Diagnostic Interview Version 3.0 (WHO-CIDI 3.0) (Haro *et al.*, 2006; Kessler and Üstun, 2004), which is a comprehensive, fully-structured interview designed to be used by trained lay interviewers for the assessment of mental disorders according to the definitions and criteria of the DSM-IV and ICD-10. An internal sampling strategy was used in all surveys to reduce the respondent burden by dividing the interview into two parts. The Part I module included a core diagnostic assessment of all respondents ($N = 2436$). Respondents who endorsed the core items of any disorder then entered the Part II module of that disorder and were asked additional questions. We analyzed the response data of respondents who entered the module for depression of the WHO-CIDI 3.0. For this module, there were a total of 353 respondents who had symptoms of either depressive mood lasting most of the day or of markedly diminished interest.

Items and diagnostic criteria

The depression module for the WHO-CIDI 3.0 assessed 14 groups of symptoms: question 1 (Q1) 'depressed mood,' Q2 'loss of interest (anhendonia),' Q3 'weight or appetite changes,' Q4 'sleep problems,' Q5 'psychomotor activity (objective or subjective),' Q6 'fatigability,' Q7 'worthlessness,' Q8 'concentration difficulty,' Q9 'suicidal ideation,' Q10 'loss of confidence,' Q11 'self-reproach,' Q12 'objective psychomotor activity,' Q13 'weight and appetite changes,' and Q14 'non-reactive depressed mood.' Because each symptom is identified by several questions, some symptoms are similar. Q1 to Q9 comprise the components of the DSM-IV, while Q2, Q4, Q6, and Q8 to Q14 are the

components of the ICD-10 (details of these items are given in Appendix 1). Diagnoses were based on criteria of both the DSM-IV and ICD-10. A major depressive episode (MDE) according to the DSM-IV requires five or more out of nine symptoms, while a minor depressive episode, which is not an official diagnosis, requires two to four symptoms. In both types of episodes, the symptom of either depressive mood lasting most of the day or diminished interest is the least required symptom. The ICD-10 criteria have three levels in depressive episode disorder: severe, moderate and mild depressive episodes. A severe depressive episode requires eight or more out of 10 criteria symptoms including loss of interest, fatigability, and depressive mood; a moderate depressive episode requires five or more criteria symptoms including at least loss of interest or fatigability; and a mild depressive episode requires three or more criteria symptoms including at least loss of interest or fatigability.

Analysis methods

IRT model

Item response models calibrate items and respondents on a common latent scale simultaneously. The two-parameter logistic (2PL) model used here is commonly used in medical settings. The 2PL model expresses the logit (log odds) of probability p_{ij} that respondent i endorses item j as a combination of the parameters called item threshold b_j, item discrimination a_j and respondent's latent trait θ_i. Respondent latent trait θ_i is assumed to follow a standard normal distribution. The threshold parameter b_j and latent trait parameter θ_i represent the location point. The threshold of the item is estimated for an endorsement rate of 50%. Both are on the same latent continuum of depression.

$$\text{logit } p_{ij} = 1.7\, a_j(\theta_i - b_j)$$

Discrimination parameter a_j represents the slope of this function that depicts the probability of endorsing item j. It determines the extent to which each symptom is able to discriminate the cases from non-cases with respect to the underlying severity of depression. Items with low discrimination may indicate that the item measures a trait unrelated to the overall construct it is supposed to measure. Items with high discrimination, however, are considered to be sensitive to the severity. We applied this 2PL model to all 14 items in the DSM-IV and ICD-10 simultaneously.

IRT assumes a unidimensionality nature of measurement, which means that a single dominant trait is sufficient to account for respondent performance. Among some indexes proposed for unidimensionality (Falissard, 1999), here we report the results of exploratory factor analyses and the goodness-of-fit statistics, that is, infit and outfit, which refer to how well the model fits the data (Haley *et al.*, 1994). Infit is an information-weighted fit statistic that is more sensitive to unexpected responses to items near a respondent's severity level. Outfit is an outlier-sensitive fit statistic that is more sensitive to unexpected responses by respondents on items far from their severity level. Values substantially less than one indicate dependencies in the data or redundant items, and values substantially greater than one indicate respondents' unexpected behavior. The formulas used to calculate these fit statistics are shown in Appendix 2.

Information function

Item information is defined as the inverse of the variance of conditional probability given θ; it is summed to produce the total information function for criteria. The total information function represents how much information the criteria have at each θ; it is used to check the information provided by the criteria across all latent severity ranges (i.e. on the latent continuum of depressive severity). We summed this information for the DSM-IV and ICD-10 separately.

Computation

We took a Bayesian approach to estimate parameters. Specified prior distributions for parameter θ was the standard normal distribution, while a and b were non-informative priors. We used Gibbs sampling to obtain the marginal posterior distribution for the parameters without analytical approximations (Gillks *et al.*, 1993). We simulated three independent sequences of length 10,000 to sample each parameter, and assessed the convergence of sequence for all parameters with the potential scale reduction factor (Gelman and Rubin, 1996). To reduce dependency on initial values, we removed the first 1000 samples of each sequence. Analyses were performed with the statistical package SAS/IML Version 9.1.

Results

Item parameter estimates

The fact that the respondents had to respond affirmatively to either Q1 'depressed mood' or Q2 'loss of interest (anhendonia)' before they would be asked subsequent questions is

a critical part of the CIDI 3.0 interview system protocol. As for the analyses, first an exploratory factor analysis for the 14 symptoms was carried out by principal component analysis to assess the dimensionality of the symptoms. The eigenvalue of the first factor accounted for 36.1% of the common variance, and subsequent factors accounted for less than 15%, suggesting a one-factor structure for these 14 symptoms. However, the factor loadings of Q1 (depressed mood) and Q14 (non-reactive depressed mood) were relatively low, presumably because the interview system required the presence of at least one of depressed mood or loss of interest to continue on to additional interview symptoms. We also performed an exploratory factor analysis for the 12 symptoms (i.e. all except Q1 and Q2) and obtained an analogous result (the eigenvalue of the first factor accounted for 40.3% of the common variance, and subsequent factors accounted for less than 17%).

Table 1 shows the proportion of affirmative responses, the parameter estimates, and the fit statistics for the individual symptoms. The convergence of Gibbs sampling were quite well. The infit and outfit statistics of the symptoms were within the acceptable range (0.48 to 1.08). The outfit values of Q5 'objective psychomotor activity' and Q7 'worthlessness' were relatively low, indicating a high correlation with other symptoms. The results of the item response model excluding the core symptoms are also shown in the right side of Table 1. Although the estimated parameters are not the same, the relative values are consistent. Thus, we henceforth report the results of the model that includes the core symptoms. In addition, Figure 1 plots the item characteristic curves (ICCs), which plot the respondents' probability of endorsement of each item related to latent severity [(a) nine symptoms in the DSM-IV, and (b) 10 symptoms in the ICD-10]. In this population, due to the interview structure of the WHO-CIDI 3.0, almost all respondents endorsed Q1 and Q14, the core symptom for the module, and therefore their thresholds and discriminations were estimated to be extremely low, with wide 95% confidence intervals (CIs). Of the remaining 12 symptoms, Q2 'loss of interest (anhendonia)' and Q4 'sleep problems' had slightly lower thresholds, while Q13 'weight and appetite changes' had a high threshold. Q5 'objective psychomotor activity' and Q7 'worthlessness' also had relatively high thresholds. The discrimination parameters of Q5 and Q7 were high; the probabilities of endorsing these symptoms changed markedly along with a slight difference in depression severity around the threshold level. However, Q3 'weight or appetite changes' and Q13 'weight and appetite changes' showed low discrimination; the change in endorsement probability was little even when depression severity differed.

Information functions

Figure 2 shows the curves of the information functions of the DSM-IV criteria (solid line) and ICD-10 criteria (dashed line) across the entire severity spectrum. The function of the DSM-IV was rather flat, with small peaks at −1.3 and 1.3, and the ICD-10 was unimodal with its highest peak at −1. The higher peak seen in the DSM-IV line is due to Q5 and Q7, which have high discriminations at high thresholds. These information functions suggest that the symptoms of the ICD-10 criteria may cover a lower range of severity than the symptoms of the DSM-IV criteria.

Respondent diagnoses and the distribution of severity

We summarize the respondent diagnoses in Table 2. Almost all the cases diagnosed according to ICD-10 criteria as having had severe or moderate depressive episodes were diagnosed as having MDE according to the DSM-IV.

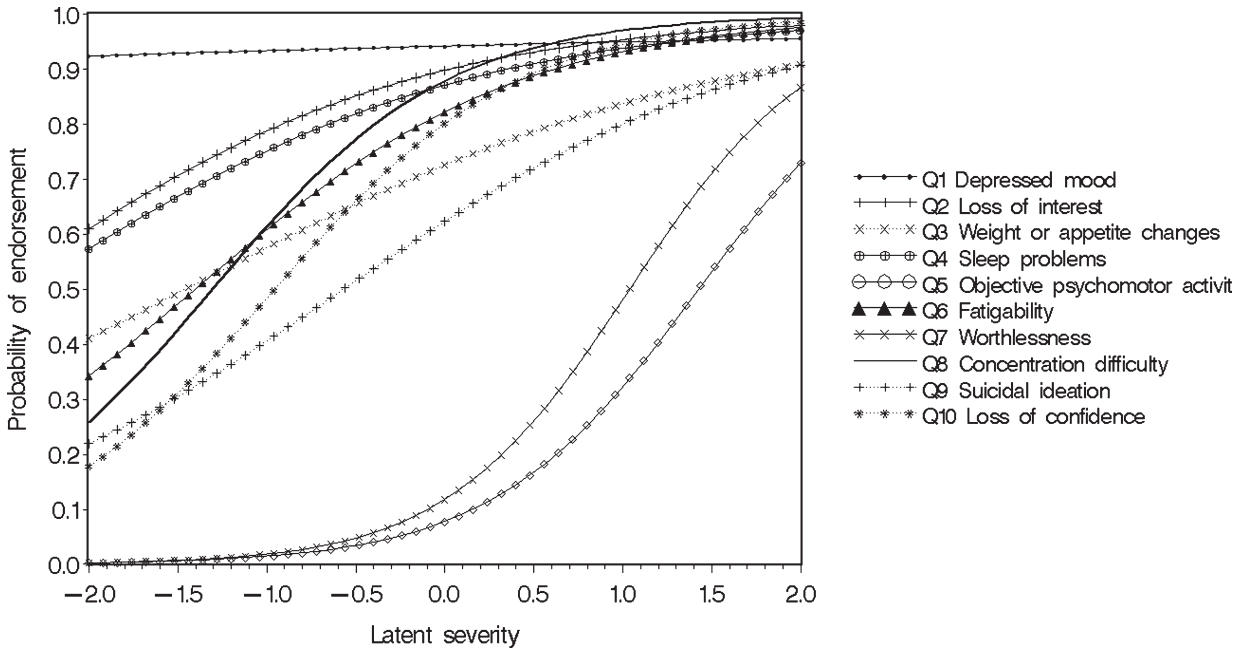
The IRT model also estimated the individual severity of depression (θ) of each respondent on the same latent trait scale as item threshold. The lower histogram in Figure 2 shows the distribution of the estimated severity of the 353 respondents with mean zero and standard deviation 0.86. This latent trait was assumed to be the standard normal prior, but the posterior distribution was narrower because it is the shrinkage parameter in a Bayesian framework.

Figure 3 shows the box-whisker plots of the respondents' severity according to the diagnoses. The more severe the diagnosis of an episode, the higher the severity. Although the DSM-IV requires at least five of nine symptoms to be met to be diagnosed as an MDE and at least two of nine to be diagnosed as a minor depressive episode, the results revealed that the estimated severity level might vary even within the same diagnosis. More respondents were excluded by the ICD-10 ($N = 90$) than by the DSM-IV ($N = 37$). These excluded respondents showed relatively high severity. The reason for this exclusion was either the fewer number of symptoms endorsed or their meeting of the exclusion criteria. Specifically, the exclusion criteria for the DSM-IV were 'due to the effects of a substance' ($N = 19$ or 51%), 'no social impairment' ($N = 9$ or 24%), 'core symptoms lasting less than two weeks' ($N = 6$ or 16%). The exclusion criteria for the ICD-10 were 'no loss of interest or fatigability' ($N = 47$ or 52%), 'depressive mood lasting less than two weeks' ($N = 29$ or 32%), 'due to the effects of a substance' ($N = 18$ or 20%), 'hypomanic or manic symptoms' ($N = 2$ or 2%). Among them, the respondents who were not diagnosed due to substance use were estimated as having a higher severity of depression.

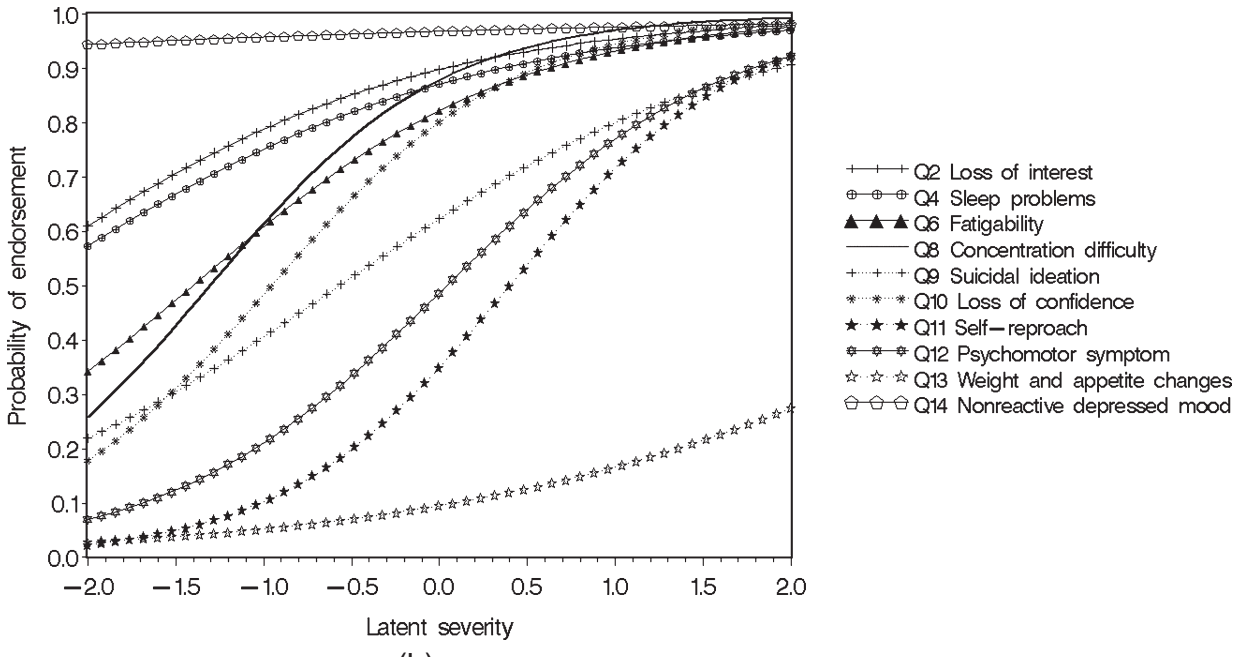
Table 1 Proportions of patients endorsed the items, the parameter estimates (95% confidence intervals [CIs]) in the model with all 14 items (left side), and the parameter estimates in the model excluding two core symptoms (right side), using IRT among respondents in a community of Japan who endorsed either depressive mood or markedly diminished interest for two weeks ($N = 353$): WMHJ 2002–2004 Survey

			Item parameters in the model with all items					Item parameters in the model excluding core items				
			Discrimination		Threshold			Discrimination		Threshold		
Symptom	Criteria	Proportion	Median	(95% CI)	Median	(95% CI)	Infit/outfit	Median	(95% CI)	Median	(95% CI)	Infit/outfit
Depressed mood	DSM	0.94	0.09	(–0.23:0.42)	–7.50	(–139.34:137.29)	1.03/1.02	–	–	–	–	–
Loss of interest (anhendonia)	DSM & ICD	0.87	0.51	(0.27:0.79)	–2.52	(–4.48:–1.76)	1.00/0.90	–	–	–	–	–
Weight or appetite changes	DSM	0.71	0.40	(0.21:0.60)	–1.45	(–2.61:–0.93)	0.95/0.94	0.39	(0.20:0.59)	–1.50	(–2.79:–0.97)	0.96/0.96
Sleep problems	DSM & ICD	0.85	0.48	(0.25:0.74)	–2.36	(–4.20:–1.62)	0.98/0.92	0.46	(0.24:0.73)	–2.43	(–4.46:–1.66)	0.98/0.96
Objective psychomotor activity	DSM	0.17	1.03	(0.67:1.56)	1.42	(1.12:1.88)	0.88/0.54	1.06	(0.68:1.64)	1.34	(1.05:1.79)	0.89/0.55
Fatigability	DSM & ICD	0.78	0.64	(0.41:0.92)	–1.40	(–2.06:–1.03)	0.91/0.84	0.70	(0.45:0.98)	–1.35	(–1.92:–1.00)	0.91/0.87
Worthlessness	DSM	0.23	1.15	(0.76:1.72)	1.03	(0.79:1.36)	0.76/0.49	1.51	(0.94:2.44)	0.89	(0.68:1.17)	0.69/0.38
Concentration difficulty	DSM & ICD	0.81	0.89	(0.60:1.26)	–1.30	(–1.76:–100)	0.86/0.66	0.94	(0.63:1.32)	–1.29	(–1.72:–1.00)	0.87/0.68
Suicidal ideation	DSM & ICD	0.61	0.52	(0.33:0.74)	–0.58	(–0.99:–0.28)	0.91/0.88	0.55	(0.36:0.77)	–0.58	(–0.97:–0.30)	0.92/0.89
Loss of confidence	ICD	0.74	0.86	(0.58:1.20)	–0.96	(–1.33:–0.70)	0.83/0.68	0.95	(0.66:1.32)	–0.93	(–1.25:–0.69)	0.83/0.68
Self–reproach	ICD	0.41	0.91	(0.61:1.33)	0.40	(0.20:0.63)	0.78/0.67	1.13	(0.74:1.72)	0.32	(0.15:0.53)	0.75/0.61
Psychomotor symptom (objective or subjective)	ICD	0.50	0.75	(0.51:1.05)	0.04	(–0.18:0.26)	0.83/0.76	0.79	(0.53:1.11)	0.00	(–0.21:0.22)	0.85/0.78
Weight and appetite change	ICD	0.11	0.38	(0.14:0.65)	3.47	(2.20:8.50)	1.02/0.92	0.38	(0.14:0.65)	3.41	(2.13:8.76)	1.02/0.93
Non-reactive depressed mood	ICD	0.97	0.18	(–0.26:0.61)	–6.86	(–107.69:95.15)	1.07/1.01	0.30	(–0.12:0.75)	–5.94	(–59.29:46.18)	1.08/0.95

Int. J. Methods Psychiatr. Res. 19(4): 211–222 (2010). DOI: 10.1002/mpr



(a)



(b)

Figure 1 Item characteristic curves of the criteria (Q1–Q14, see Table 1) for depressive disorders derived from the DSM-IV (a) and ICD-10 (b) as a function of the latent severity (horizontal) and the probability of endorsement (vertical) among respondents in a community of Japan who endorsed either depressive mood or markedly diminished interest ($N = 353$): WMHJ 2002–2004 Survey.

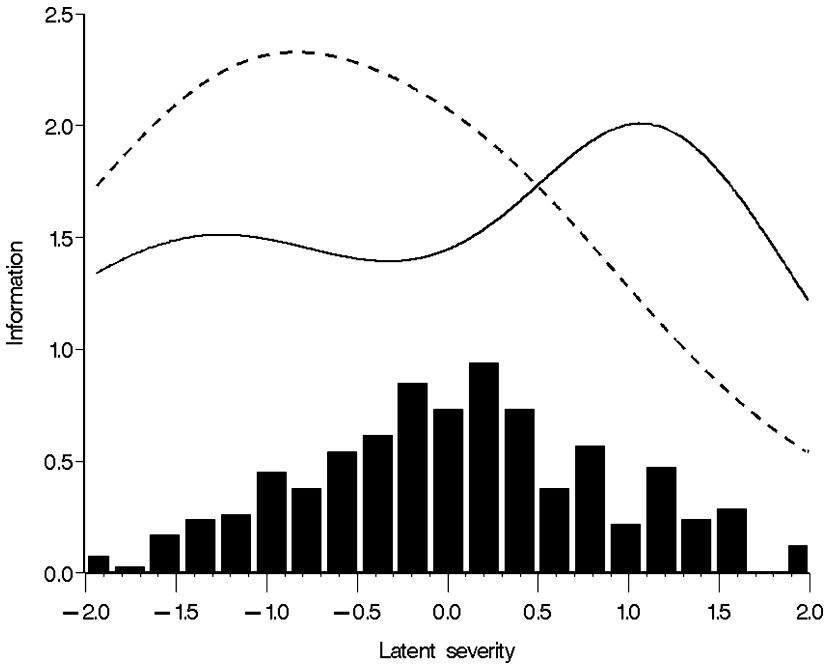


Figure 2 Information functions for symptom criteria for the DSM-IV (solid line) and ICD-10 (dotted line) diagnoses of depressive disorders, and the distribution of respondents by severity (lower histogram): WMHJ 2002–2004 Survey.

Table 2 Comparison of the DSM-IV and ICD-10 diagnoses of depressive disorders among respondents in a community of Japan who endorsed either depressive mood or markedly diminished interest ($N = 353$): WMHJ 2002–2004 Survey

		ICD-10				
		Severe	Moderate	Mild	None	Total
DSM-IV	Major	107	97	24	29	257
	Minor	0	6	25	28	59
	None	0	0	4	33	37
	Total	107	103	53	90	353

Discussion

In the current study we investigated the depressive episode-related symptoms in the DSM-IV and ICD-10 with IRT analysis of the WHO-CIDI 3.0 data obtained from community samples in Japan. Particular attention was paid to the measurement properties of individual symptoms, and to the difference in the depressive severity between the DSM-IV and ICD-10 diagnostic criteria. Before discussing the results, we should point out the limitations of this study. Our sample contained only those who had experienced either one of the two core symptoms, i.e. having a depressive mood or diminished interest. Therefore, the results obtained here may not necessarily be generalized to a general community population. In particular, the predominant endorsements of the core symptoms, Q1 or Q2, resulted in wide 95% CIs and lower discriminations of Q1 and Q14, and thus these symptoms could have possibly detracted the parameter estimation of the other symptoms. However, in the analyses of the remaining 12 symptoms, we found the relation between these symptoms and the respondents' severities were almost the same, suggesting that both Q1 and Q2 had little effect on the overall IRT analyses.

The present IRT analyses provide important information on how items of the DSM-IV and ICD-10 criteria for depressive episodes work on the same latent continuum of depressive severity. As shown in Figure 3, the θ level ranges of the diagnosed episodes overlapped each other in the distribution, while the estimated severity of respondents without any diagnoses seemed relatively high. This is likely attributable to the difference in the thresholds of the symptoms that the diagnosed patients endorsed. In the IRT analysis, one respondent who endorsed symptoms with a low threshold was estimated as low severity, while another who endorsed a few symptoms with a high threshold was estimated as high severity.

Most of the respondents diagnosed as having had a severe or moderate depressive episode according to the ICD-10 were also diagnosed as having had an MDE

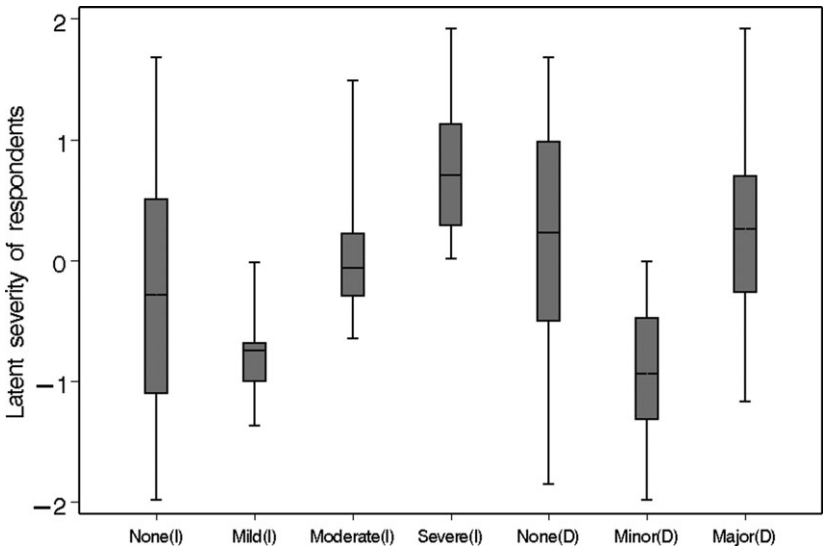


Figure 3 Box-whisker plots of the latent severity of depressive disorders estimated by IRT model according to the DSM-IV (D) and ICD-10 (I) diagnosis categories among respondents in a community of Japan who endorsed either depressive mood or markedly diminished interest for two weeks ($N = 353$): WMHJ 2002–2004 Survey.

according to the DSM-IV. The distribution of the estimated severities of these respondents also shows a similar result. That is, the diagnoses were comparable between the moderate and severe depressive episodes of the ICD-10 and the MDE of the DSM-IV. However, according to the information function curve, which is the unique information derived by IRT analysis, items of the ICD-10 criteria have much more information at the lower range of severity compared to the DSM-IV items. Thus, the ICD-10 seems more sensitive to the mild range of the depressive continuum, while the DSM-IV may be more sensitive to the moderate or severe ranges.

The IRT analyses also revealed that the items about dietary habits (Q3 and Q13) were less informative, while the items about psychomotor activity (Q5), feeling worthless (Q7) and self-reproach (Q11) were more informative for discriminating the severity of depression. The WHO-CIDI 3.0 contains some items asking about the same or similar symptoms, but these are treated as different questions in the DSM-IV and ICD-10 criteria. For example, Q1 'depressed mood' and Q14 'non-reactive depressed mood,' both of which inquire about depressive mood, are used for the DSM-IV and ICD-10 criteria, respectively. Similarly, Q5 'objective psychomotor activity,' used in the DSM-IV, and Q12 'psychomotor activity, objective or subjective,' used in the ICD-10, both inquire about experiencing a symptom related to retardation and agitation. Q3 'weight or appetite changes' and Q13 'weight and appetite changes' are quite similar, but the difference between 'and' and 'or' might lead to a difference in their thresholds. In particular, the cut-off value for weight change commonly used worldwide seems to be high for Japanese women, and thus the threshold of Q13 is estimated to be high. These wording differences do not seem essential; therefore these items could be unified. In this way, some modifications can be justified for making both sets of criteria more equivalent.

Aggen *et al.* (2005) examined the item characteristics of the DSM-IV symptoms for depression and reported that 'psychomotor activity' showed a lower threshold, while 'concentration difficulty' and 'suicidal ideation' showed higher thresholds compared to our results. Further, the discrimination parameters of all symptoms in their study were high. While we cannot yet sufficiently explain the former discrepancy, the latter discrepancy is likely due to the fact that the participants in their study comprised a community sample, with many respondents endorsing no item at all. Simon *et al.* (2002) also conducted IRT analyses according to the DSM-IV depression criteria with data obtained from primary care patients in a cross-national epidemiologic study. In their study, 'sleep problems,' 'fatigability,' 'depressed mood' and 'loss of interest' were the low-threshold symptoms, while 'suicidal ideation,' 'weight or appetite changes,' and 'psychomotor activity' were the high-threshold symptoms. These results appear consistent with the results of our study except for the item 'weight or appetite changes.' This similarity may be attributed to the characteristics of those respondents who were aware of their illness.

In this study we found a difference in the range of sensitivity of the DSM-IV and ICD-10 depressive episode criteria on the same latent trait continuum. To our knowledge this is the first investigation of the comparability of the DSM-IV and ICD-10 criteria using a methodologically sound analytic technique. To obtain more

general results, additional research including the response data of community samples who had not experienced the core symptoms is needed, and, further, similar IRT analyses should be conducted using data collected in other countries.

Conclusion

We explored the measurement characteristics of the DSM-IV and ICD-10 criteria for depressive episode-related symptoms using IRT in a population somewhat aware of their depressive state. The symptoms regarding psychomotor activity and feeling worthless had higher thresholds and were more informative with respect to discriminating depression, while those regarding dietary habits were less informative. The ICD-10 seems more sensitive to the mild range of the depressive continuum, while the DSM-IV seems to be more sensitive to the moderate and severe ranges. Our investigation also revealed that the MDE of the DSM-IV criteria corresponds to the severe and moderate depressive episodes of the ICD-10.

Acknowledgments

The World Mental Health Japan (WMHJ) set of surveys is supported by a Grant for Research on Psychiatric and Neurological Diseases and Mental Health (H13-SHOGAI-023, H14-TOKUBETSU-026, H16-KOKORO-013, H19-KOKORO-IPPAN-011) from the Japan Ministry of Health, Labor, and Welfare. We thank the staff members, filed coordinators, and interviewers of the WMHJ 2002–2004 Survey. The WMHJ 2002–2004 Survey was carried out in conjunction with the World Health Organization (WHO) World Mental Health (WMH) Survey Initiative. We also thank the WMH staff for assistance with instrumentation and fieldwork. These activities were supported by the US National Institute of Mental Health (R01MH070884), the John D. and Catherine T. MacArthur Foundation, the Pfizer Foundation, the US Public Health Service (R13-MH066849, R01-MH069864, and R01 DA016558), the Fogarty International Center (FIRCA R01-TW006481), the Pan American Health Organization, Eli Lilly and Company, Ortho-McNeil Pharmaceutical, Inc., GlaxoSmithKline, and Bristol-Myers Squibb. A complete list of WMH publications can be found at http://www.hcp.med.harvard.edu/wmh/

Declaration of interest statement

The authors have no competing interests.

References

American Psychiatric Association (1987) *Diagnostic and Statistical Manual of Mental Disorders, 3rd edition-revised*, American Psychiatric Association.

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition*, American Psychiatric Association.

Aggen S.H., Neale M.C., Kendler K.S. (2005) DSM criteria for major depression: evaluating symptom patterns using latent-trait item response models. *Psychological Medicine*, **35**, 475–487, DOI: 10.1017/S0033291704003563

Boeck P.D., Wilson M. (2004) *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach* (3rd ed.), Springer.

Cuijpers P., Smit F. (2002) Excess mortality in depression: a meta-analysis of community studies. *Journal of Affective Disorders*, **72**, 227–236, DOI: 10.1016/S0165-0327(01)00413-X

Falissard B. (1999) The unidimensionality of a psychiatric scale: a statistical point of view. *International Journal of Methods in Psychiatric Research*, **8**, 162–167, DOI: 10.1002/mpr.66

Faravelli C., Servi P., Arends J.A., Strik W.K. (1996) Number of symptoms, quantification, and qualification of depression. *Comprehensive Psychiatry*, **37**, 307–315, DOI: 10.1016/S0010-440X(96)90011-5

Fontenelle L.F., Mendlowicz M.V., Moreira R.O., Appolinario J.C. (2005) An empirical comparison of atypical bulimia nervosa and binge eating disorder. *Brazilian Journal of Medical and Biological Research*, **38**, 1663–1667, DOI: 10.1590/S0100-879X2005001100014

Gelman A., Rubin D.B. (1996) Markov chain Monte Carlo. *Statistical Methods in Medical Research*, **5**, 339–355, DOI: 10.1177/096228029600500402

Gibbons R.D., Clark D.C., Davis J.M. (1982) A statistical model for the classification of imipramine response in depressed inpatients. *Psychopharmacology*, **78**, 185–189, DOI: 10.1007/BF00432260

Gillks W.R., Clayton D.G., Spiegelhalter D.J., Best N.G., McNeil A.J., Sharples L.D., Kirby A.J. (1993) Modeling complexity: applications of Gibbs sampling in medicine. *Journal of the Royal Statistical Society: Series B*, **55**, 39–52.

Haley S.M., McHorney C.A., Ware J.E. Jr. (1994) Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimentionality and reproducibility of the Rasch item scale. *Journal of Clinical Epidemiology*, **47**, 671–684, DOI: 10.1016/0895-4356(94)90215-1

Han L., Schmaling K.B., Dunner D.L. (1995) Descriptive validity and stability of diagnostic criteria for dysthymic disorder. *Comprehensive Psychiatry*, **36**, 338–343, DOI: 10.1016/S0010-440X(95)90114-0

Haro J.M., Arbabzadeh-Bouchez S., Brugha T.S., Girolamo G.D., Guyer M.E., Jin R., Lepine J.P., Mazzi F., Reneses B., Vilagut G., Sampson N.A., Kessler R.C. (2006) Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health Surveys. *International Journal of Methods in Psychiatry Research*, **15**, 167–180, DOI: 10.1002/mpr.196

Ietsugu T., Sukigara M., Furukawa T.A. (2007) Evaluation of diagnostic criteria for panic attack using item response theory: findings from the National Comorbidity Survey in USA. *Journal of Affective Disorders*, **104**, 197–201, DOI: 10.1016/j.jad.2007.03.005

Kawakami N., Shimizu H., Haratani T., Iwata N., Kitamura T. (2004) Lifetime and 6-month

prevalence of DSM-III-R psychiatric disorders in an urban community in Japan. *Psychiatry Research*, **121**, 293–301, DOI: 10.1016/S0165-1781(03)00239-7

Kawakami N., Takeshima T., Ono Y., Uda H., Hata Y., Nakane Y., Nakane H., Iwata N., Furukawa T.A., Kikkawa T. (2005) Twelve-month prevalence, severity, and treatment of common mental disorders in communities in Japan: preliminary finding from the World Mental Health Japan Survey 2002–2003. *Psychiatry and Clinical Neurosciences*, **59**, 441–452, DOI: 10.1111/j.1440-1819.2005.01397.x

Kessler R.C., Üstun T.B. (2004) The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatry Research*, **13**, 93–121, DOI: 10.1002/mpr.168

Lahey B.B., Pelham W.E., Chronis A., Massetti G., Kipp H., Ehrhardt A., Lee S.S. (2006) Predictive validity of ICD-10 hyperkinetic disorder relative to DSM-IV attention-deficit/hyperactivity disorder among younger children. *Journal of Child Psychology and Psychiatry*, **47**, 472–479, DOI: 10.1111/j.1469-7610.2005.01590.x

Lopez Ibor J.J., Frances A., Jones C. (1994) Dysthymic disorder: a comparison of DSM-IV and ICD-10 and issues in differential diagnosis. *Acta Psychiatrica Scandinavica*, Supplement, **s383**: 12–18, DOI: 10.1111/j.1600-0447.1994.tb05878.x

Lord F.M., Novick M.R. (1968) *Statistical Theories of Mental Test Scores*, Addison Wesley Longman Publishing.

McHorney C.A., Haley S.M., Ware J.E. Jr. (1997) Evaluation of the MOS SF-36 physical functioning scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *Journal of Clinical Epidemiology*, **50**, 451–461, DOI: 10.1016/S0895-4356(96)00424-6

Otsubo T. (2005) Epidemiology of the depressive disorder. *Rinsho Seishin Igaku*, **34**, 871–880 (in Japanese).

Simon G.E., Goldberg P., Von Korff M., Ustun T.B. (2002) Understanding cross-national differences in depression prevalence. *Psychological Medicine*, **32**, 585–594, DOI: 10.1017/S0033291702005457

Simon G.E., Von Korff M. (2006) Medical co-morbidity and validity of DSM-IV depression criteria. *Psychological Medicine*, **36**, 27–36, DOI: 10.1017/S0033291705006136

Üstun T.B., Ayuso-mateo J.L., Chatterji S., Mathers C., Murray C.J.L. (2004) Global burden of depressive disorders in the year 2000. *British Journal of Psychiatry*, **184**, 386–392, DOI: 10.1192/bjp.184.5.386

Vilalta-Franch J., Garre-Olmo J., López-Pousa S., Turon-Estrada A., Lozano-Gallego M., Hernàndez-Ferràndiz M., Pericot-Nierga I., Feijóo-Lorza R. (2006) Comparison of different clinical diagnostic criteria for depression in Alzheimer disease. *American Journal of Geriatric Psychiatry*, **14**, 589–597, DOI: 10.1097/01.JGP.0000209396.15788.9d

World Health Organization (WHO) (1992) *The ICD-10 Classification of Mental and Behavioural Disorders; Clinical Descriptions and Diagnostic Guidelines*, World Health Organization.

The World Health Organization (WHO) World Mental Health (WMH) Survey Consortium (2004) Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *Journal of the American Medical Association*, **291**, 2581–2590, DOI: 10.1001/jama.291.21.2581

Appendix 1

Details of the 14 symptoms used to evaluate depression in the WHO-CIDI 3.0.

Q1 Depressed mood
Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g. feels sad or empty) or observations made by others.

Q2 Loss of interest (anhendonia)
Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation made by others).

Q3 Weight or appetite changes
Significant weight loss when not dieting or weight gain (e.g. change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day.

Q4 Sleep problems
Insomnia or hypersomnia nearly every day (sleep disturbance of any type).

Q5 Objective psychomotor activities
Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).

Q6 Fatigability
Fatigability or loss of energy nearly every day.

Q7 Worthlessness
Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).

Q8 Concentration difficulty
Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others). Complaints or evidence of diminished ability to think or concentrate, such as indecisiveness or vacillation.

Q9 Suicidal ideation
Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

Q10 Loss of confidence
Loss of confidence or self-esteem.

Q11 Self-reproach
Unreasonable feelings of self-reproach or excessive and inappropriate guilt.

Q12 Psychomotor symptom (objective or subjective)
Change in psychomotor activity, with agitation or retardation either objectively observed by others or subjectively-assessed

Q13 Weight and appetite changes
Change in appetite (decrease or increase) with corresponding weight change.

Q14 Non-reactive depressed mood
Depressed mood to a degree that is definitely abnormal for the individual, present for most of the day and almost every day, largely uninfluenced by circumstances, and sustained for at least two weeks.

Appendix 2

Formula used to calculate fit statistics

The value O_{ij} is an observed response of respondent i to item j, E_{ij} is the expected value, and σ_j^2 is the variance of expectation. The squared standardized residual of respondent i to item j is then expressed as follows:

$$Z_{ij} = (O_{ij} - E_{ij})^2 \big/ \sigma_{ij}^2 \quad \text{where} \quad \sigma_{ij}^2 = E_{ij}(1 - E_{ij})$$

The outfit statistic is the mean of the squared standardized residuals in the respondents. The infit statistic is the information-weighted (i.e. inverse of variance-weighted) mean of the standardized residuals.

$$\text{Outfit}_j = \sum\nolimits_i Z_{ij} / N_j$$

$$\text{Infit}_j = \sum\nolimits_i \sigma_{ij}^2 Z_{ij} \Big/ \sum\nolimits_i \sigma_{ij}^2$$

Int. J. Methods Psychiatr. Res. 19(4): 211–222 (2010). DOI: 10.1002/mpr