International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 24(3): 204–212 (2015)
Published online 18 February 2014 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1437

One-year test–retest reliability of a Japanese web-based version of the WHO Composite International Diagnostic Interview (CIDI) for major depression in a working population†

HARUKI SHIMODA,[1,2] AKIOMI INOUE,[3] KANAMI TSUNO[4] & NORITO KAWAKAMI[1]

1 Department of Mental Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
2 Department of Mental Health Policy and Evaluation, National Institute of Mental Health, National Center of Neurology and Psychiatry, Tokyo, Japan
3 Department of Mental Health, Institute of Industrial Ecological Sciences, University of Occupational and Environmental Health, Japan, Kitakyushu, Japan
4 Department of Hygiene, School of Medicine, Wakayama Medical University, Wakayama, Japan

Key words
major depression, WHO-Composite International Diagnostic Interview (WHO-CIDI), reliability, psychiatry, demographic characteristics

Correspondence
Haruki Shimoda, Department of Mental Health, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113–0033, Japan.
Telephone (+81) 3-5841-3522
Fax (+81) 3-5841-3392
Email: harukis221@yahoo.co.jp

†This article was published online on 18 February 2014. Errors were subsequently identified. This notice is included in the online and print versions to indicate that both have been corrected 14 March 2014.

Abstract

The purpose of this study was to investigate the one-year test–retest reliability and the demographic correlates of a self-administered web-based depression section of the World Health Organization-Composite International Diagnostic Interview (WHO-CIDI) in a working population. Overall, 1060 out of all employees ($N = 1279$) from a manufacturing company in Japan responded to two web-based surveys of depression of the WHO-CIDI within a one-year interval in 2009 and 2010. The concordance between lifetime diagnoses of major depressive disorder on two occasions was calculated as percent agreement (%), Gwet's AC_1, and Yule's Q indicators were compared by gender, age, education, and marital status. For the total sample, percent agreement was 94%, AC_1 was 0.93, and Yule's Q was 0.82. The concordance rate was low (0.15) among those who were diagnosed at either time or both times. The concordance differed significantly across education and marital status. While the agreement indicators were relatively high, consistent with previous reports based on face-to-face interviews conducted within a shorter interval, the low stability of positive cases may challenge the accuracy of lifetime diagnosis of major depressive disorder using a web version of the WHO-CIDI. Education and marital status might affect the test–retest reliability.

Introduction

Common mental disorders, such as major depressive disorder, have a huge influence on people's quality of life, as expressed in disability-adjusted life years (Prince *et al.*, 2007). A number of community-based epidemiologic studies conducted in the United States and other parts of the world, including Japan, have estimated prevalence and risk factors of common mental disorders (Andrade *et al.*, 2003; Demyttenaere *et al.*, 2004). Lifetime prevalence of any mental disorder (i.e. the proportion of those who ever experienced a mental disorder in their lifetime before completing the survey) was estimated at 12–47% across countries (Kessler *et al.*, 2007) while 12-month prevalence was 4-26% (Demyttenaere *et al.*, 2004). Many large-scale studies employed structured interviews, such as the US NIMH Diagnostic Interview Schedule (DIS; Robins *et al.*, 1981) and the World Health Organization-Composite International Diagnostic Interview (WHO-CIDI; Wittchen, 1994), conducted by trained lay interviewers in order to increase inter-rater reliability of the diagnoses of mental disorders.

The WHO-CIDI is a lay-interviewer administered, highly structured interview schedule for diagnosing a wide range of common mental disorders according to the accepted definitions and criteria of the International Classification of Diseases, 10th Revision (ICD-10) and the American Psychiatric Association's Diagnostic and Statistical Manual (DSM) (Robins *et al.*, 1981; Kessler and Ustün, 2004). The WHO-CIDI, written at the request of the WHO/US Alcohol, Drug Abuse, and Mental Health Administration Task Force on Psychiatric Assessment Instruments, comprises questions from the DIS along with questions designed to elicit Present State Examination items (Robins *et al.*, 1988). It is a comprehensive, fully structured diagnostic interview for the assessment of mental disorders, which provides lifetime and current diagnoses by means of standard scoring algorithms. The latest 3.0 version of the WHO-CIDI (Kessler and Ustün, 2004) validated diagnoses of common mental disorders against clinical interviews (Haro *et al.*, 2006). The WHO-CIDI was translated into Japanese and used in epidemiologic surveys of common mental disorders in the communities in Japan (Kawakami *et al.*, 2005; Kawakami, 2006). Depression section was validated by comparing clinician diagnoses with clinical structured interview schedules (Kawakami *et al.*, 2008). The self-report WHO-CIDI sections of depression and anxiety were also administered via the web and had high concordance with interviewer-administered sections (Peters *et al.*, 1998). Many existing studies have evaluated the reliability of WHO-CIDI. Most studies reported good (> 0.60) or moderate (0.40–0.60) Cohen's kappa (κ) for mood, anxiety, and substance use disorders (Wittchen, 1994; Wittchen *et al.*, 1998). For instance, the percent agreement was, on average, greater than 0.85, and the κ was greater than 0.5 for depressive disorders among four test–retest studies with 3–7 day intervals.

Inter-rater agreement statistics can be computed in different ways. The *S* coefficient (Bennett *et al.*, 1954), the π statistics (Scott, 1955), and the κ statistics (Cohen, 1960) are traditional indicators of the extent of agreement between two observers (see Table 1). Kendall's *W* (Kendall and Smith, 1939) and weighted κ (Cohen, 1960) were also developed for ordinal variables. Intraclass correlation coefficient (ICC) and Cronbach's α are used to evaluate the inner consistency between some specific items, for example, subscales measuring the same phenomenon. However, κ statistic has been most frequently used in studies that tested the reliability of the CIDI, while it has been recognized that the κ statistic depends on the prevalence (Feinstein and Cicchetti, 1990). When the prevalence is low, κ statistics becomes extremely low. Therefore, several indicators, such as Gwet's AC_1 (Gwet, 2002) and Yule's Q (Yule and Kendall, 1957), which are relatively independent of the prevalence, have been recently proposed as alternative indicators of the concordance.

Participants' memory of a previous test might affect the result of a retest given only in a few-days interval (Wittchen *et al.*, 1989). Using an extended washout period might minimize such recall bias (Vera *et al.*, 2010). With a longer time interval, the test–retest reliability of the diagnosis might be much lower, particularly when assessing lifetime experiences. Indeed, a test–retest study with long interval reported some limitations with estimated reliability (Bromet *et al.*, 1986), since changes in environment and respondent's mental health status may have affected their responses to a questionnaire. However, not many studies investigated the test–retest reliability with a long interval. One study with a 20-month interval still reported high concordance for mood and anxiety

Table 1. Assessment for lifetime major depressive disorder (MDD) episode of N sample at baseline and follow-up

		Assessment at follow-up		
		Yes	No	Total
Assessment at baseline	Yes	a	b	f_1
	No	c	d	f_2
	Total	g_1	g_2	N

disorders (0.64 or greater in Yule's *Y*) but not specific phobias in a small random sample ($N = 85$) of the community population (Wittchen, 1994). These findings should be replicated.

In addition, factors affecting the test–retest reliability have not been fully investigated yet. Few studies have previously investigated the effects of demographic characteristics, such as gender, age, education, and marital status, on the test–retest reliability. Older age and lower education might be associated with poor recall of an episode of mental disorder and thus result in poor test–retest reliability. Marital status also might affect the test–retest reliability because divorced or widowed individuals may better recall a difficult event and a related episode of mental disorder in their life.

Recently, a web-based depression section of the WHO-CIDI 3.0 was administered to a working population in Japan at two time-points within one-year interval. Using the data from the survey, this study tested one-year concordance (test–retest reliability) of the diagnoses of major depressive disorder. We also investigated the influence of selected demographic characteristics (gender, age, education, and marital status) on the test–retest reliability of the diagnosis. Web-based methods for conducting epidemiological survey are expected to enable researchers to make large-scale research easily, although the reliability of these measures might be limited to some extent.

Participants and methods

Participants

A prospective study of employees from five branches of a manufacturing company located in the Kanto (east coast) region of Japan was conducted between August 2009 and August 2010. The data were collected using the depression section of the web-based self-administered computerized CIDI 3.0. At baseline (August 2009), all employees ($N = 1279$) were invited to participate in this study. Before collecting data using the web-based self-administered questionnaire, participants were assured that their participation was voluntary and the information they provided was confidential. Overall, 1228 questionnaires were returned. Because 115 out of 1228 employees were excluded due to transfer, retirement, leave of absence, or death during one-year follow-up, the total number of eligible employees for follow-up survey was 1113 (August 2010). Out of 1113 employees, 1060 returned the completed follow-up questionnaires. To secure the quality of data, company personnel motivated employees to complete the surveys. At the end of each survey, we measured the response quality by asking participants whether they reported correct information. After excluding seven employees who had at least one missing response for variables relevant to this study, 1053 employees were analyzed. The final sample comprised 391 males and 662 females aged 20 to 63 years [mean = 36.72, standard deviation (SD) = 7.947]. Table 2 shows the detailed characteristics of participants. The dropout group had significantly higher rate of those aged from 50 to 63 and significantly lower rate of singles, according to χ^2 analysis. It could be caused by retirement and marriage dissolution. The Ethics committee of the Graduate School of Medicine at the University of Tokyo reviewed and approved the procedures of this study (No.2580).

Diagnosis of major depressive disorder

The depression section of the WHO-CIDI first asks three screening questions on stem symptoms (two questions for dysphoric mood and one for interest loss) experienced during lifetime. If a respondent indicates a presence of any of the symptoms in their lifetime, the survey further enquires about the duration (two weeks or more) of the symptoms, other symptoms of major depressive disorder, related functional impairments, and the exclusion criteria (Kessler and Ustün, 2004). Using the obtained information, a computer program generates a diagnosis of major depressive disorder according to DSM-IV. The depression section of the WHO-CIDI was developed into a web-based questionnaire. A small modification of the questions was made. Although the web-based questionnaire posed similar stem questions as the original CIDI, the web-administered stem questions were divided into two parts. The first part assessed 12-month experience. If participants did not experience symptoms within the last 12 months, then the second part assessed lifetime experience prior to the 12 month. A preliminary study showed that web-based questionnaire had moderate sensitivity (71.4% among 14 cases with clinically diagnosed major depressive episode) and high specificity (100% among nine cases without clinically diagnosed major depressive episode) in terms of 12-month prevalence.

At baseline (in 2009), we assessed the lifetime diagnosis before the survey. At follow-up (in 2010), we assessed the lifetime diagnosis prior to the 12 months, that is, before the initiation of the survey, in order to match the timeframe with the lifetime diagnosis assessed at baseline. The diagnosis was made without the hierarchy rule, omitting exclusion criteria (i.e. no overlapping of bipolar disorder). Additionally, to see whether the modification affected test–retest reliability, we calculated a concordance between two lifetime prevalences at baseline and follow-up (not subtracting the past 12 months from the lifetime).

Table 2. Number of participants, baseline lifetime prevalence of DSM-IV major depressive disorder, relative prevalence rate in relation to total sample, multivariate odds ratio indicating the correlation between subject's characteristics and prevalence rate at baseline and follow up using the generalized estimating equations (GEE), and indicators of the one-year concordance (test-retest reliability) between the diagnoses in 2009 and 2010 for the total sample (n = 1,053) and demographic subgroups measured by a web-based version of the WHO Composite International Diagnostic Interview (CIDI)

Characteristics	n	Lifetime prevalence (%) at baseline (2009)	Relative prevalence in relation to total sample	Multivariate odds ratio for the concordance between diagnoses (se)	% Agreement	κ (se)	AC1 (se)	Yule's Q (se)
Total sample	1,053	52(5%)	1		94%	0.23(0.06)	0.93(0.01)	0.82(0.06)
Gender								
Male	391	12(3%)	0.6	1	96%	0.21(0.13)	0.96(0.01)	0.89(0.10)
Female	662	40(6%)	1.2	2.7(1.4)	92%	0.23(0.07)	0.91(0.01)	0.79(0.01)
Age in 2009								
50-65	62	3(5%)	0.9	1	95%	0.38(0.28)	0.95(0.03)	0.93(0.01)
35-49	573	27(5%)	1.0[a]	0.7(1.8)	94%	0.24(0.09)	0.94(0.01)	0.85(0.01)
20-34	418	22(5%)	1.1	0.7(1.9)	92%	0.20(0.09)	0.92(0.01)	0.76(0.01)
Education years								
16-	611	28(5%)	0.9	1	95%	0.27(0.09)	0.94(0.01)	0.88(0.06)
13-15	310	14(5%)	0.9	0.9(1.3)	91%	0.02(0.07)	0.90(0.02)	0.18(0.03)
-12[a]	132	10(8%)	1.6	1.5(1.5)	95%	0.51(0.16)	0.94(0.02)	0.98(0.02)
Marital Status								
Single	477	25(5%)	1.1	1	93%	0.26(0.09)	0.92(0.01)	0.84(0.01)
Married	516	23(5%)	0.9	1.0(1.3)[a]	94%	0.08(0.07)	0.93(0.01)	0.61(0.06)
Divorced/widowed	60	4(7%)	1.4	1.6(1.6)	95%	0.64(0.19)	0.94(0.03)	0.98(<0.01)

se = standard error
[a]Correction made here after initial online publication

Demographic characteristics

Demographic variables included gender (male or female), age (from 20 to 34 years old, 35 to 49 years old, and 50 to 63 years old), education (12 years or less, from 13 years to 15 years, and 16 years or higher), and marital status (single, married, or divorced/widowed).

Statistical analysis

Analyses were conducted using SPSS and SAS. Lifetime prevalence of major depressive disorder at baseline and diagnostic agreement using κ, AC_1 and Yule's Q were calculated for the total sample as well as for subgroups classified according to the demographic characteristics (gender, age group, education, and marital status).

In medical research, κ is a common index for evaluating the extent of agreement between raters. It is computed by subtracting chance concordance from observed concordance, but recently, it is said that κ has problems of two paradoxes (Feinstein and Cicchetti, 1990). The first paradox is that if chance agreement on calculation is large, value of κ may indicate poor reliability even with a high value of observed prevalence rate. The second one is that κ will be higher when imbalance in marginal totals is asymmetrical rather than symmetrical. AC_1 is computed differently to resolve these problems. Chance concordance is $P_e = (f_1g_1 + f_2g_2)/N^2$ for κ, and $P_e^* = 2\pi\ (1-\pi)$, $\pi = (f_1 + g_1)/2N$ for AC_1, but this indicator is quite new and has not been well validated.

Yule's Q relates to the odds ratio of $Q = (\theta - 1)/(\theta + 1)$, $\theta = ad/bc$, a monotone transformation of θ from the $[0, \infty]$ scale onto the $[-1, 1]$ scale (Agresti, 2002). In this study, we used percent agreement, κ, AC_1, and Yule's Q, as major agreement indicators. A standard error of the AC_1 was estimated based on a conditional variance by using a SAS algorithm (Blood and Spratt, 2007; Gwet, 2008).

Moreover, we compared prevalence of each characteristic to total sample by calculating simple ratio and subsequently computed multivariate odds ratio for each demographic characteristic using logistic regression analysis modeled by GEE. GEE models the association between the data collected repeatedly on each subject with a patterned correlation matrix (Williamson, 2000; Barnhart, 2001). To model the effects of demographic characteristics on diagnosis estimated by logistic regression using GEE, we assigned identical number to all respondents at baseline and follow-up and created clusters. To further investigate the effects of the demographic characteristics on the concordance and discordance in the diagnoses between the two occasions, the participants were divided into four groups based on their diagnoses at baseline and follow-up. These groups were: (i) not diagnosed on both occasions; (ii) diagnosed on both occasions; (iii) not diagnosed at baseline and diagnosed at follow-up; (iv) diagnosed at baseline and not diagnosed at follow-up. All were analyzed using GEE.

Results

One-year concordance of the diagnoses

For the total sample, the one-year concordance between the diagnosis at baseline and follow-up was 94% for the percent agreement, 0.23 for κ, 0.93 for AC_1, and 0.82 for Yule's Q (Table 2).

Concordance of the diagnoses by demographic characteristic

Among different demographic subgroups, percent agreement and AC_1 values were similar and high in all subgroups while the values of κ and Yule's Q were similar but relatively low, especially κ. Percent agreement ranged from 91% to 96% and AC_1 ranged from 0.90 to 0.96 in all subgroups. For gender and age subgroups, κ ranged from 0.20 to 0.38, Yule's Q ranged from 0.76 to 0.93, and κ ranged from 0.02 to 0.64. Yule's Q ranged from 0.18 to 0.98 among subgroups of education and marital status.

At baseline, 52 participants had lifetime history of major depression and 40 participants had lifetime history at follow-up (Table 3). Overall, the number of participants diagnosed on both occasions was 12 (23% of the 52 participants diagnosed at baseline), 68 (6% of the total sample) were diagnosed either at baseline or follow-up and 973 (92%) were not diagnosed on any occasion.

The analysis of participants based on the combination of the diagnosis at baseline and follow-up showed that females and middle educated were significantly more likely to not be diagnosed at baseline but diagnosed at follow-up ($p = 0.02$ and 0.02, respectively) (Table 4).

Table 3. Consistency of lifetime diagnoses of major depressive disorder (MDD) at baseline (2009) and follow-up (2010)

	History of MDD at follow-up	No History of MDD at follow-up	Total
History of MDD at baseline	12	40	52
No history of MDD at baselie	28	973	1001
Total	40	1013	1053

Int. J. Methods Psychiatr. Res. 24(3): 204–212 (2015). DOI: 10.1002/mpr

Table 4. Correlation between sociodemographic characteristics and the pattern of diagnoses at baseline and follow-up

	Agree (YY)					Disagree (NY)					Disagree (YN)				
Characteristics	n	Bivariate odds ratio	P	Multivariate odds ratio	P	n	Bivariate odds ratio	P	Multivariate odds ratio	P	n	Bivariate odds ratio	P	Multivariate odds ratio	P
Gender															
Male	2	1		1		4	1		1		10	1		1	
Female	10	3.1	0.14	4.0[a]	0.12	24	3.8	0.02	2.5	0.15	30	1.9	0.09	2.3	0.06
Age in 2009 (years)															
50-65	1	1		1		1	1		1		2	1		1	
35-49	6	0.7	0.69	0.5	0.51	12	1.3	0.80	0.8	0.81	21	1.1	0.86	1.0[a]	0.98
20-34	5	0.8	0.81	0.3	0.32	15	2.3	0.43	1.1	0.91	17	1.3	0.74	1.0[a]	0.97
Education (years)															
16-	7	1		1		12	2		1		21	1		1	
13-15	1	0.3	0.25	0.2	0.09	15	2.5	0.02	1.9	0.13	13	1.3	0.52	1.0[a]	0.96
-12	4	2.7	0.12	2.2	0.24	1	0.4	0.37	0.4	0.39	6	1.3	0.53	1.4	0.46
Marital status															
Single	7	1		1		15	1		1		18	1		1	
Married	2	0.3	0.09	0.3	0.15	11	0.7	0.31	1.1	0.79	21	1.1	0.86	1.4	0.34
Divorced/widowed	3	3.5	0.08	2.6	0.21	2	1.1	0.92	1.4	0.68	1	0.5	0.44	0.5	0.45

(Y-) = diagnosed at baseline, (N-) = not diagnosed at baseline, (-Y) = diagnosed at follow-up, (-N) = not diagnosed at follow-up
[a]Correction made here after initial online publication

Quality of data reported by participants and property of CIDI modification

In each survey, 97–98% of the respondents reported that they provided true information in the survey. Recent studies have reported that the prevalence of major depression in Japan is between 3% and 7% (Kawakami, 2006). In this study, 4.9% had prevalence of major depression at baseline. In addition, two lifetime prevalences at baseline and follow-up (not lifetime minus one year but lifetime prevalence) that calculated to see the affection of the modification to test–retest reliability resulted in 4.9% at baseline and 5.0% at follow-up.

Discussion

For the total sample, the study found that percent agreement, AC_1, and Yule's Q were high for one-year concordance of the lifetime diagnosis of major depressive disorder, as assessed by a web version of depression section of the WHO-CIDI in a working population in Japan. The percent agreement and Yule's Q observed in this study were comparable to those reported in previous test–retest studies of WHO-CIDI with shorter intervals and in one study with a 20 months interval (Wittchen, 1994). However, κ was relatively low compared to previous studies (Wittchen, 1994; Wittchen *et al.*, 1998), which may be attributable to the low prevalence of major depression in this sample. It was also notable that 77% of the respondents diagnosed at baseline were not diagnosed at follow-up. While most indicators showed good test–retest agreement, the instrument failed to re-diagnose most of those initially diagnosed at follow-up, which may challenge the accuracy of the diagnosis using a web version of the WHO-CIDI. This could have happened because respondents may have had a difficulty in recalling a past episode of depression or they may have hesitated to report their experience honestly after they already admitted to a past depression episode at the baseline. A further investigation into the reasons for this discrepancy and the ways to improve the accuracy of the diagnosis is needed, e.g. by improving the questions.

Among middle educated participants, the one-year concordance was lower when measured by κ and Yule's Q for which we could not make a reasonable explanation. From the viewpoint of distribution of diagnoses on both occasions, the reason why κ and Yule's Q were low among middle educated could be because very few of them were diagnosed at both times while a relatively large number of them were diagnosed on only one occasion. Kappa becomes lower regardless of observed concordance when the balance of positive-positives and negative-negatives (a and d, respectively, in Table 1) is extreme because of bigger chance of concordance calculated. Yule's Q also becomes lower with more number of discordance especially in this case. However, AC_1 becomes higher in the case that the number of positive-positives and negative-negatives is not marginal when the total amount of concordance is fixed (Gwet, 2002). Therefore, low κ and Yule's Q and high AC_1 among middle educated could be, at least partly, attributable to the distribution of the pattern of diagnosis among them.

Kappa and Yule's Q were also lower for the married participants, while percent agreement and AC_1 were almost similar across marital status categories. This is attributable to the fact that most individuals who were diagnosed at baseline tended not to report the episode at the follow-up compared to the participants in the other categories. This may be explained by the differences in mental health status between married individuals and divorced/widowed. Those who were married may be less likely to recall the past episode of major depression, possibly because of receiving more support from the spouses and having better mental health status. However, divorced/widowed individuals might recall the past episode better because they might remember difficult life events, such as divorce and loss of the loved one, clearly, which could have caused higher κ and Yule's Q among divorced/widowed participants.

Among respondents who were not diagnosed at baseline, females and middle educated participants were more likely to be diagnosed at the follow-up. A study of long-term test–retest reliability reported that clinical status during an interval between interviews significantly influenced diagnostic stability (Bromet *et al.*, 1986). Females were reported to have a higher prevalence of major depressive disorder in the community (Andrade *et al.*, 2003) as well as depressive symptoms among the working population (Kawakami *et al.*, 1995); therefore, they may be more depressed than other respondents within a one-year interval and recall a past depression episode better at the follow-up.

As expected from their characteristics, AC_1 and Yule's Q indicated that the web-version of WHO-CIDI 3.0 depression section had high test–retest reliability in this sample while κ was quite low, as kappa has been known to be quite sensitive to prevalence rate. However, high values of AC_1 and Yule's Q are counterintuitive, when considering the fact that many first positives were not diagnosed in the second assessment. Yule's Q and chance of concordance of AC_1 are calculated by multiplication and division of the number (or possibility) of positive and negative. While calculating them, the ratio of discordance (b and c) to positive-positives (a) or negative-negatives (d) are not

taken into account. So, AC_1 and Yule's Q can indicate high value even if positive predictive value or negative predictive value (the likelihood that an individual with a positive/negative test result truly has/has not the particular trait) of a rater is low when the number of positive-positives or negative-negatives is low respectively, like the present study. Considering the results of the present study, which indicated counterintuitive high reliability, we reconfirmed the importance of considering carefully the degree to which the indicators reflect response reliability, not just distribution. In most research of mental disorders, the number of positives is lower than negatives. Considering relative clinical importance of accurately diagnosing positives more than diagnosing negatives, future studies need to investigate the reliability the agreement indicators show taking a serious view of positive predictive value and sensitivity.

The study has three main limitations. First, while previous studies reported comparable reliability of an interviewer-based and web-based CIDI, only few studies examined the reliability of a web-based CIDI. While previous research showed no significant difference in test–retest reliability among face-to-face, telephone, and the internet surveys (Ritter *et al.*, 2004; Vallejo *et al.*, 2007; Donker *et al.*, 2010), the current finding may not be applicable to an interviewer-administered CIDI. Second, web-based surveys have been increasingly popular in Asian as well as in Western countries, but the quality of collected data and their response tendencies are still not clear. The present study suggests a need to clarify psychometric properties of web-based surveys in non-Western countries. However, its applicability in low-income countries is still questionable because many people may not have an access to the Internet. Third, the current web-based WHO-CIDI depression section was modified. The stem questions were divided into two parts, one assessing a 12-month episode and another assessing a lifetime episode. We have not received permission from the CIDI Editorial Committee about this modification, which may affect the findings, although the quality of data reported by participants was good and lifetime prevalences on two occasions were almost the same.

Acknowledgments

The present study was supported by a Grant-in-Aid for Scientific Research (A) 2009–1011 (No. 20240062) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

Reference

Agresti A. (2002) Categorical Data Analysis, 2nd edition, Wiley series in probability and Statistics, New Jersey, John Wiley & Sons.

Andrade L., Caraveo-Anduaga J.J., Berglund P., Bijl R.V., de Graaf R., Vollebergh W., Dragomirecka E., Kohn R., Keller M., Kessler R.C., Kawakami N., Kiliç C., Offord D., Ustün T.B., Wittchen H.U. (2003) The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys. *International Journal of Methods in Psychiatric Research*, **12**(1), 3–21, DOI: 10.1002/mpr.138

Bennett E.M., Alpert R., Goldstein A. (1954) Communications through limited response questioning. *Public Opinion Quarterly*, **18**(3), 303–308.

Barnhart H.X. (2001) Modeling Concordance Correlation via GEE to Evaluate reproducibility. *Biometrics*, **57**(3), 931–940, DOI: 10.1111/j.0006-341x.2001.00931.x

Blood E., Spratt K.F. (2007) Disagreement on agreement: two alternative agreement coefficients. *SAS Global Forum Statistics and Data Analysis*, **186**, 1–12.

Bromet E.J., Dunn L.O., Connell M.M., Dew M.A., Schulberg H.C. (1986) Long-term reliability of diagnosing lifetime major depression in a community sample. *Archives of General Psychiatry*, **43**(5), 435–40, DOI: 10.1001/archpsyc.1986.01800050033004

Cohen J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46, DOI: 10.1177/001316446002000104

Demyttenaere K., Bruffaerts R., Posada-Villa J., Gasquet I., Kovess V., Lepine J.P., Angermeyer M.C., Bernert S., de Girolamo G., Morosini P., Polidori G., Kikkawa T., Kawakami N., Ono Y., Takeshima T., Uda H., Karam E.G., Fayyad J.A., Karam A.N., Mneimneh Z.N., Medina-Mora M.E., Borges G., Lara C., de Graaf R., Ormel J., Gureje O., Shen Y., Huang Y., Zhang M., Alonso J., Haro J.M., Vilagut G., Bromet E. J., Gluzman S., Webb C., Kessler R.C., Merikangas K.R., Anthony J.C., Von Korff M. R., Wang P.S., Brugha T.S., Aguilar-Gaxiola S., Lee S., Heeringa S., Pennell B.E., Zaslavsky A.M., Ustün T.B., Chatterji S., WHO World Mental Health 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 – JAMA*, **291**(21), 2581–2590, DOI: 10.1001/jama.291.21.2581

Donker T., Straten A.V., Marks I., Cuijpers P. (2010) Brief self-rated screening for depression on the Internet. *Journal of Affective Disorders*, **122**(3), 253–259, DOI: 10.1016/j.jad.2009.07.013

Feinstein A.R., Cicchetti D.V. (1990) High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, **43**(6), 543–549, DOI: 10.1016/0895-4356(90)90158-L

Gwet K. (2002) Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. *Series: Statistical Methods for Inter-rater Reliability Assessment*, **2**, 1–9.

Gwet K.L. (2008) Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, **61**(1), 29–48, DOI: 10.1348/00071106X126600

Haro J.M., Arbabzadeh-Bouchez S., Brugha T.S., de Girolamo G., 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 Psychiatric Research*, **15**(4), 167–180, DOI: 10.1002/mpr.196

Kawakami N. (2006) Major depression in Japan and the world, state-of-art in epidemiology. *Igakunoayumi*, **219**(13), 925–929.

Kawakami N., Roberts R.E., Lee E.S., Araki S. (1995) Changes in rates of depressive symptoms in a Japanese working population: life-table analysis from a 4-year follow-up study. *Psychological Medicine*, **25**(6), 1181–1190, DOI: 10.1017/S0033291700033158

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**(4), 441–452, DOI: 10.1111/j.1440-1819.2005.01397.x

Kawakami N., Takeshima T., Ono Y., Uda H., Nakane Y., Nakamura Y., Tachimori H., Iwata N., Nakane H., Watanabe M., Naganuma Y., Furukawa T.A., Hata Y., Kobayashi M., Miyake Y., Kikkawa T. (2008) Twelve-month prevalence, severity, and treatment of common mental disorders in communities in Japan: the World Mental Health Japan 2002–2004 Survey. In Kessler R.C., Ustün T.B. (eds) The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders, New York: Cambridge University Press, pp 474–485.

Kendall M.G., Smith B.B. (1939) The problem of m rankings. *Annals of Mathematical Statistics*, **10**(3), 275–287, DOI: 10.1214/aoms/1177732186

Kessler R.C., Ustün 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 Psychiatric Research*, **13**(2), 93–121, DOI: 10.1002/mpr.168

Kessler R.C., Angermeyer M., Anthony J.C., de Graaf R., Demyttenaere K., Gasquet I., de Girolamo G., Gluzman S., Gureje O., Haro J. M., Kawakami N., Karam A., Levinson D., Medina Mora M.E., Oakley Browne M.A., Posada-Villa J., Stein D.J., Adley Tsang C.H., Aguilar-Gaxiola S., Alonso J., Lee S., Heeringa S., Pennell B.E., Berglund P., Gruber M.J., Petukhova M., Chatterji S., Ustün T.B. (2007) Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry*, **6**(3), 168–176, DOI: 10.1001/archpsyc.62.6.593

Peters L., Clark D., Carroll F. (1998) Are computerized interviews equivalent to human interviewers? CIDI-Auto versus CIDI in anxiety and depressive disorders. *Psychological Medicine*, **28**(4), 893–901, DOI: 10.1017/S0033291798006655

Prince M., Patel V., Saxena S., Maj M., Maselko J., Phillips M.R., Rahman A. (2007) No health without mental health. *Lancet*, **370**(9590), 859–877, DOI: 10.1016/50140-6736(07) 61238-0

Ritter P., Lorig K., Laurent D., Mtthews K. (2004) Internet versus mailed questionnaires: a randomized comparison. *Journal of Medical Internet Research*, **6**(3), e29, DOI: 10.2196/jmir.6.3.e29

Robins L.N., Helzer J.E., Croughan J., Ratcliff K.S. (1981) National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. *Archives of General Psychiatry*, **38**(4), 381–389, DOI: 10.1001/archpsyc.1981.01780290015001

Robins L.N., Wing J., Witchen H.U., Helzer J.E., Babor T.F., Burke J., Farmer A., Jablenski A., Pickens R., Reginer D.A., Sartorius N., Towle L.H. (1988) The Composite International Diagnostic Interview. An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry*, **45**(12), 1069–1077, DOI: 10.1001/archpsyc.1988.01800360017003

Scott W.A. (1955) Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, **19**(3), 321–325.

Vallejo M.A., Jordan C.M., Diaz M.I., Comeche M. I., Ortega J. (2007) Psychological assessment via the Internet: a reliability and validity study of online (vs paper-and-pencil) versions of the General Health Questionnaire-28 (GHQ-28) and the Symptoms Check-List 90-Revised (SCL-90-R). *Journal of Medical Internet Research*, **9**(1), e2, DOI: 10.2196/jmir.9.1.e2

Vera M.A.D., Ratzlaff C., Doerfling P., Kopec J. (2010) Reliability and validity of an internet-based questionnaire measuring lifetime physical activity. *American Journal of Epidemiology*, **172**(10), 1190–1198, DOI: 10.1093/aje/kwq273

Williamson J.M. (2000) Modeling kappa for measuring dependent categorical agreement data. *Biostatistics*, **1**(2), 191–202, DOI: 10.1093/biostatistics/1.2.191

Wittchen H.U. (1994) Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): a critical review. *Journal of Psychiatry Research*, **28**(1), 57–84, DOI: 10.1016/0022-3956(94)90036-1

Wittchen H.U., Burke J.D., Semler G., Pfister H., Cranach M.V., Zaudig M. (1989) Recall and dating of psychiatric symptoms – test–retest reliability of time related symptom questions in a standardized psychiatric interview. *Archives of General Psychiatry*, **46**(5), 437–443, DOI: 10.1001/archpsyc.1989.01810050051009

Wittchen H.U., Lachner G., Wunderlich U., Pfister H. (1998) Test–retest reliability of the computerized DSM-IV version of the Munich-Composite International Diagnostic Interview (M-CIDI). *Social Psychiatry and Psychiatric Epidemiology*, **33**(11), 568–578.

Yule G.U., Kendall M.G. (1957) An Introduction to the Theory of Statistics, 14th edition, p. 30, London, Griffin.