International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 17(3): 152–158 (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/mpr.257

The performance of the Japanese version of the K6 and K10 in the World Mental Health Survey Japan

TOSHI A. FURUKAWA,[1] NORITO KAWAKAMI,[2] MARI SAITOH,[3] YUTAKA ONO,[4] YOSHIBUMI NAKANE,[5] YOSIKAZU NAKAMURA,[6] HISATERU TACHIMORI,[7] NOBORU IWATA,[8] HIDENORI UDA,[9] HIDEYUKI NAKANE,[10] MAKOTO WATANABE,[6] YOICHI NAGANUMA,[7] YUKIHIRO HATA,[11] MASAYO KOBAYASHI,[6] YUKO MIYAKE,[7] TADASHI TAKESHIMA,[7] TAKEHIKO KIKKAWA[12]

1 Department of Psychiatry and Cognitive-Behavioral Medicine, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan
2 Department of Mental Health, University of Tokyo Graduate School of Medicine, Tokyo, Japan
3 Department of Biostatistics/Epidemiology and Preventive Health Sciences, University of Tokyo Graduate School of Medicine, Tokyo, Japan
4 Health Center, Keio University, Keio, Japan
5 Division of Human Sociology, Nagasaki International University Graduate School, Nagasaki, Japan
6 Department of Public Health, Jichi Medical School, Jichi, Japan
7 National Institute of Mental Health, National Center of Neurology and Psychiatry, Tokyo, Japan
8 Department of Clinical Psychology, Hiroshima International University, Hiroshima, Japan
9 Sensatsu Public Health Center, Kagoshima Prefecture, Kagoshima, Japan
10 Division of Neuropsychiatry, Department of Translational Medical Sciences, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan
11 Oshima Hospital, Kagoshima Prefecture, Kagoshima, Japan
12 Department of Human Well-being, Chubu Gakuin University, Seki City, Japan

Abstract

Two new screening scales for psychological distress, the K6 and K10, have been developed using the item response theory and shown to outperform existing screeners in English. We developed their Japanese versions using the standard back-translaton method and included them in the World Mental Health Survey Japan (WMH-J), which is a psychiatric epidemiologic study conducted in seven communities across Japan with 2436 participants. The WMH-J used the WMH Survey Initiative version of the Composite International Diagnostic Interview (CIDI) to assess the 30-day Diagnostic and Statistical Manual of Mental Disorders – Fourth Edition (DSM-IV). Performance of the two screening scales in detecting DSM-IV mood and anxiety disorders, as assessed by the areas under receiver operating characteristic curves (AUCs), was excellent, with values as high as 0.94 (95% confidence interval = 0.88 to 0.99) for K6 and 0.94 (0.88 to 0.995) for K10. Stratum-specific likelihood ratios (SSLRs), which express screening test characteristics and can be used to produce individual-level predicted probabilities of being a case from screening scale scores and pretest probabilities in other samples, were strikingly similar between the Japanese and the original versions. The Japanese versions of the K6 and K10 thus demonstrated screening performances essentially equivalent to those of the original English versions. Copyright © 2008 John Wiley & Sons, Ltd.

Key words: general psychological distress, screening instrument, ROC curve, stratum-specific likelihood ratio

Introduction

We now have a plethora of dimensional scales of general psychological distress to be used in community epidemiological surveys, starting with the 20-item Health Opinion Survey in the Stirling County Study (MacMillan, 1957) and the 22-item Langner Scale in the Midtown Manhattan Study (Langner, 1962). Some are better validated and in wider use [e.g. the General Health Questionnaire (Goldberg and Williams, 1988)] than others but until recently none had been developed using modern psychometric methods to maximize the screening precision.

Dohrenwend and his colleagues' review of screening instruments showed that the scales typically included questions about a heterogeneous set of cognitive, behavioral, emotional and psychophysiological symptoms that are elevated among people with a wide range of different mental disorders (Dohrenwend et al., 1980). In other words, despite this heterogeneous content, the vast majority of these symptoms have high factor loadings on a first principal factor, which therefore can be regarded as representing non-specific psychological distress. Assuming this uni-dimensionality of general psychological distress, the modern item response theory can allow us to select items that are maximally discriminative at a certain point of the general population distribution of this general factor (van den Linden and Hambleton, 1997). Kessler and his colleagues (Kessler et al., 2002) developed 10-item and 6-item very short screening instruments using modern item response theory methods to select questions that are maximally discriminative of respondents in the 90th to 99th percentile range of the population distribution, because it was known that between 5–10% of the general population suffer from serious mental illness at any point in time. Only items displaying constant psychometric characteristics across gender, age, race/ethnicity, and educational and socio-demographic subsamples were included in the final questionnaires (Kessler et al., 2002).

The resulting scales, referred to as K10 and K6 (K6 is a subset of K10), strongly discriminated between community cases and non-cases of the Diagnostic and Statistical Manual of Mental Disorders – Fourth Edition (DSM-IV) in an American epidemiological survey done to test its performance. The area under the receiver operating characteristic curve (AUC) was 0.88 for both scales (Kessler et al., 2002). In a subsequent validation study using a convenience sample in the US, both K6 and K10 worked at least as well as the more extensive Composite International Diagnostic Interview – Short Form (CIDI-SF) in identifying cases of clinically significant mental illness. The AUC was 0.85 for K10 and 0.86 for K6 (Kessler et al., 2003). Using data from a large nationally representative household survey undertaken in Australia, Furukawa and his colleagues found that K10 was marginally more discriminative than K6 and that both significantly outperformed the widely used 12-question General Health Questionnaire (GHQ). The AUC was 0.90, 0.89 and 0.80 for K10, K6 and GHQ, respectively. Moreover, K6 was more robust than K10 to subsample variation (Furukawa et al., 2003).

The items in K10 and K6 ask respondents how frequently they experienced symptoms of psychological distress (e.g. feeling so sad that nothing can cheer you up) during the past 30 days. Responses are recorded using a five-category scale (0 = all of the time, 1 = most of the time, 2 = some of the time, 3 = a little of the time, and 4 = none of the time), producing therefore a score range 0–40 for K10 and 0–24 for K6. K10 and K6 are now core measures in the annual US National Heath Interview Survey, the US National Household Survey of Drug Abuse, and the Canadian National Health Interview Survey. K10 is now being used as a standard patient-completed outcome measure for all mental health services in the state of New South Wales, Australia (Brooks et al., 2006).

Given these strong psychometric properties and the brevity, development of Japanese versions of K10 and K6 was urgently needed. Because these instruments were incorporated into the World Health Organization (WHO) World Mental Health Survey being undertaken in multiple countries including Japan (Demyttenaere et al., 2004), we had the unique opportunity to develop the Japanese version and to test it in a methodologically sound community survey in Japan. The present study reports the development and psychometric characteristics of the newly developed Japanese K10 and K6. When high validity as a screening instrument is ascertained, we will calculate the scales' stratum-specific likelihood ratios (SSLRs) and illustrate their use, as SSLRs are increasingly used in diagnostic processes in other branches of medicine (Furukawa et al., 2008).

DOI: 10.1002/mpr

Methods

Subjects

The World Mental Health Survey Japan (WMH-J) is an epidemiological survey of Japanese-speaking household residents aged 20 and older. It was conducted in seven communities across Japan in 2002–2004, including two urban cities and five rural municipalities. These sites were selected on the basis of geographic variation, availability of site investigators, and cooperation of the local government. A random sample was selected from residents aged ≥20 years old in each survey site, based on a voter registration list or a resident registry. After a letter of invitation was sent, trained interviewers contacted the subjects and interviewed those who agreed to participate in the survey using the standardized instrument. The total response rate was 58.4%.

An internal sampling strategy was used in all surveys to reduce respondent burden by dividing the interview into two parts. Part I included a core diagnostic assessment of all respondents that took an average of about one hour to administer. Part II included questions about risk factors, consequences and others, including the K6 and K10. Part II was administered to all Part I respondents with one or more lifetime disorders plus a probability subsample of approximately 25% of the remaining respondents. Part II respondents were weighted by the inverse of their probability of selection to adjust for the differential sampling of cases and non-cases.

The Ethics Committees of Okayama University, National Institute of Mental Health Japan, and Nagasaki University approved the recruitment, consent and field procedures. Written informed consent was obtained from each respondent. More details of the study procedures are given in a separate article (Kawakami et al., 2005).

Measures

The diagnostic interview included in the survey was the World Mental Health Survey Initiative Version of the World Health Organization Composite International Diagnostic Interview (WMH-CIDI) (Haro et al., 2006; Kessler and Ustun, 2004). The CIDI is a fully structured interview designed to generate DSM-IV and International Classification of Diseases – 10th revision (ICD-10) diagnoses based on responses obtained in face-to-face interviews by trained lay interviewers.

The Japanese version of the K6 and K10 questionnaires were developed in accordance with the WHO translation guidelines by experts in psychiatric interviewing and/or psychiatric epidemiology. TAF first translated the original English versions into Japanese. An expert panel of two psychiatric epidemiologists (NK and Yoshiharu Kim) then examined this preliminary Japanese version in view of the original English, and modified the translation where necessary. A Japanese woman with a BSc in psychology and fluent in English backtranslated the Japanese version into English, which was then checked by Todd Strauss under Ron C. Kessler. This process was repeated until TS found the backtranslation to be equivalent to the original. The Japanese version was then field tested.

Analysis methods

SPSS (SPSS Inc., 2006), SAS (SAS Institute Inc., 2004) and Microsoft Excel were used for statistical analyses of the data.

Evaluating performance of the screening scales

The purpose of the screening scales considered here is to screen for broadly defined mental disorders rather than for one particular diagnosis. As a result, we defined the 'gold standard' of caseness as any current DSM-IV mood disorder (depression, dysthymia) or anxiety disorder (panic disorder, agoraphobia, social phobia, generalized anxiety disorder, post-traumatic stress disorder). Taking the ICD-10 mood and anxiety disorders as the 'gold standard' yielded similar results and herein we report the results for DSM-IV only. The performance of the continuous versions of the K6 and K10 as screening scales was analyzed using receiver operating characteristic (ROC) curves. Areas under ROC curves (AUC) and their 95% confidence intervals were calculated by the non-parametric method, and compared between the screening scales by the standard method (Hanley and McNeil, 1983).

Stratum-specific likelihood ratios (SSLRs)

Instead of dichotomizing the originally continuous ROC curve by use of a single cutoff, we can calculate the multi-level or SSLRs. An SSLR is a ratio of two likelihoods, one of showing the test result in question among those with the target disorder, over one of showing the same test result among those without the disorder. According to the Bayes theorem, it can be shown that:

DOI: 10.1002/mpr

Population odds × SSLR = individual respondent odds

where

Odds = Probability/(1 − Probability)

where Odds and Probability refer to those having the target disorder in the total population.

This means that the SSLR indicates by how much a given screening score value will raise or lower the odds of having the target disorder for an individual respondent in comparison to the total population. Based on this convenient relationship between population odds or prevalence and SSLR, SSLR values generated in a benchmark sample can be used to compute individual-level predicted probabilities of being a case from screening scale scores in other samples with different prevalences. SSLRs are presented for the screening scales evaluated here in order to help investigators who use these scales make such computations.

Because the two screening scales have wide ranges (0–24 for the K6 and 0–40 for the K10), they were collapsed into a smaller number of strata in order to improve precision of estimation. Peirce and Cornell (1993) developed a spreadsheet program to arrive at the optimal number of strata of test scores by calculating likelihood ratios specific to different strata along with their 95% confidence intervals. Because with too many strata the likelihood ratio becomes unstable and degenerate, the following rules of thumb were recommended: (1) to provide sufficient abnormal and normal cases in each stratum to allow the relative-odds of caseness across strata to be monotonically related, and (2) to collapse those strata where the relative-odds are close to one another and their 95% confidence intervals easily overlap (Peirce and Cornell, 1993). It is further advisable (3) to keep strata with SSLRs smaller than 0.1 or greater than 10 distinct at the two tails of the screening scale distribution because values at these extremes can be decisively informative in the diagnostic process and often rule in or rule out the disorders (Furukawa et al., 2008).

The 95% confidence interval of each SSLR was estimated using bootstrapping following Furukawa et al. (2001). The bias-corrected 95% confidence intervals were estimated by the BCa method (Efron and Tibshirani, 1993). Each 95% confidence interval was calculated by 1000 bootstrap samples. When there was no case in a stratum, the 95% confidence interval was estimated as follows. The upper limit p_u for the probability of having the disorder was obtained as a solution to $(1 - p_u)^n = 0.05$. The interval $(0, p_u)$ provides 95% coverage for probability (Jovanovic, 2005). Then we calculated the 95% confidence interval of odds and SSLR for the stratum.

Subpopulation variation

In order to look for subgroups where SSLRs may be substantively different, we used logistic regression analyses involving interaction terms between each screening instrument and demographics (sex, age). When statistically significant interactions were noted, we intended to examine SSLRs separately for that variable.

Results

Prevalences

The total sample size was 2463, of whom 887 proceeded to Part II survey and 864 completed K6 and K10 questionnaires. These 864 respondents constitute the subjects of the following analyses: 60.1% were women and the mean age was 52.2 [standard deviation (SD) = 16.4, range = 20 to 93); 28 persons (weighted prevalence 1.4%) had at least one of the target disorders listed in the Methods section in the 30 days preceding the interview.

Comparative accuracies of the screening scales

Figure 1 displays the ROC curves of the K6 and K10 against the target diagnoses. The two screening instruments had very similar AUC, namely 0.94 (0.88 to 0.995) for K10 and 0.94 (95% confidence interval = 0.88 to 0.99) for K6. The differences were not statistically significant ($p = 1.0$).

SSLRs

SSLRs of the K10 and K6 were calculated (Table 1). The number of strata was set to six, so that they can be compared to the published SSLRs of the English versions of the K6 and K10 (Furukawa et al., 2003). Both K10 and K6 were associated with informative SSLRs, i.e. smaller than 0.1 or greater than 10, towards both ends of the population spectrum. Neither sex nor age was a significant modifier for the SSLRs of the K10 or K6.

DOI: 10.1002/mpr

Discussion

The accuracies of the K10 and K6 as screening scales

The Japanese versions of both K10 and K6 demonstrated excellent AUCs, comparable to or even somewhat superior to those reported from the US and Australia. The standard backtranslation procedure succeeded in producing a functionally equivalent scale from English into Japanese.

It is also important to note that, in all the samples tested so far, the shorter K6 demonstrated AUCs very similar to those of the K10. The additional four items of K10 do not seem to add much to the screening performance of the scale, and the apparent brevity of the K6 argues strongly in favor of the shorter version.

The usefulness of the SSLR approach

Naïve users of screening scales usually specify a single caseness threshold that is applied to all populations. For example, Kessler et al. (2003) once argued that, in order to equalize false-positive and false-negative results, the optimum cutoff was 12/13 for K6. More thoughtful users, however, recognize that accuracy of classification increases incrementally as screening scale scores become more and more extreme and that the probabilities of caseness at a given value on the screening scale vary with population prevalence (Goldberg et al., 1998).

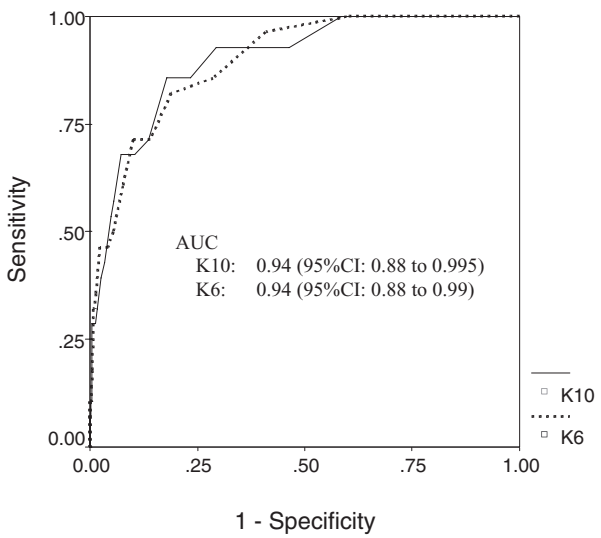

Figure 1. ROC curves of K6 and K10.

The SSLR approach is ideally suited to deal with both of these observations. The calculation of a separate SSLR for each meaningful scale range allows the user to assign more fine-grained classifications than those available with the more conventional dichotomous screening approach. A useful nomogram proposed by Fagan (1975) (an interactive Web-based version is available at: http://meta.cche.net/clint/templates/calculators/lr_nomogram.asp and at the Centre for Evidence-Based Medicine http://www.cebm.net/nomogram.asp) can facilitate the translation of SSLRs into predicted probabilities. Alternatively, if the clinician is at ease with computers, a spreadsheet program will generate the individual's probability of caseness based on the population prevalence and the SSLR. (An example Microsoft Excel spreadsheet can be downloaded from the website of Nagoya City University Evidence-based Psychiatry Center, http://www.ebpcenter.com.)

Table 2 shows the SSLRs of K6 and K10 in the Australian nationally representative epidemiological survey (n = 10,641). Because of the relatively small sample size of the Japanese survey (n = 864), the confidence intervals of the SSLRs of the Japanese versions are much wider than those of the Australian ones, but one is naturally struck with the similarities of each SSLR value for each stratum between the English and the Japanese versions.

Table 1. SSLRs of the K6 and K10 in the World Mental Health Survey Japan (WMH-J)

K6	0	1,2	3–5	6–8	9–13	14–24
SSLR	0.00 (0 to 0.0001)	0.29 (0.07 to 0.85)	1.5 (0.41 to 3.5)	4.9 (1.7 to 11.2)	16 (6.1 to 34)	110 (11 to 400)
K10	0,1	2–4	5–9	10–14	15–19	20–40
SSLR	0.00 (0 to 0.0001)	0.16 (0.00 to 0.71)	1.8 (0.70 to 3.4)	6.1 (2.1 to 14)	11 (2.3 to 32)	110 (32 to 280)

95% confidence intervals are in parentheses.

DOI: 10.1002/mpr

Table 2. SSLRs of the K6 and K10 in Australia (Furukawa et al., 2003)

K6	0	1,2	3–5	6–8	9–13	14–24
SSLR	0.09 (0.06 to 0.13)	0.23 (0.18 to 0.28)	1.0 (0.84 to 1.1)	3.8 (3.3 to 4.4)	11 (8.9 to 13)	46 (33 to 65)
K10	0,1	2–4	5–9	10–14	15–19	20–40
SSLR	0.08 (0.05 to 0.12)	0.29 (0.23 to 0.37)	1.3 (1.2 to 1.5)	5.4 (4.7 to 6.2)	15 (12 to 18)	83 (40 to 170)

95% confidence intervals are in parentheses.

However, the same SSLRs do not lead to the same post-test probabilities, if the pre-test probabilities or population prevalences are different. For example, in the Australian survey, the prevalence of the target disorders was 8.0% (Furukawa et al., 2003), while in the Japanese survey it was 1.4%. If a patient scores 13 on K6, i.e. just above the cutoff proposed by Kessler and his colleagues, the post-test probability for an Australian patient would be calculated as 49% and that for a Japanese patient as 19%, according to the nomograms or spreadsheet mentioned earlier. Or, if another patient scores five on K6, the post-test probabilities would be 8% and 2%, respectively. These finer-grained interpretations of screening test results should lead to more accurate interpretations and less false negative or false positive readings.

In other words, advantages of the SSLR approach over the dichotomous fixed threshold approach may be summarized as follows. First, for a test that produces continuous scores, SSLRs retain as much information as possible that is originally contained in the test by deriving multiple level indices instead of reducing the test into a dichotomous value below or above the cutoff. Secondly, SSLRs are themselves independent of prevalence of the target disorder but choice of the optimum threshold is often not, for example when it is determined by equalizing the numbers of false-positives and false-negatives. The polychotomous SSLR approach is hence more informative and generalizable than dichotomous fixed threshold approach.

Conclusions

The original K10 and K6 were developed with the modern item response theory so that they are sensitive to the 90th to 99th percentile range of the population distribution of non-specific psychological distress. We developed their Japanese versions according to the standard backtranslation procedure. The Japanese versions demonstrated screening performances that are essentially equivalent to those reported with the original English versions, suggesting that the backtranslation procedure, coupled with the modern test development method, succeeded in producing cross-culturally applicable screening scales.

Acknowledgments

The World Mental Health Japan (WMH-J) is supported by the Grant for Research on Psychiatric and Neurological Diseases and Mental Health (H13-SHOGAI-023, H14-TOKUBETSU-026, H16-KOKORO-013) from the Japan Ministry of Health, Labor, and Welfare. We would like to thank the staff members and other field coordinators in the WMH-J 2002–2004 Survey. The WMH-J 2002–2004 Survey is carried out in conjunction with the World Health Organization World Mental Health (WMH) Survey Initiative. We also thank the WMH staff for assistance with instrumentation, fieldwork, and data analysis. 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 Interests

TAF has received research funds and speaking fees from Asahi Kasei, Astellas, Dai-Nippon Sumitomo, Eisai, Eli Lilly, GlaxoSmithKline, Janssen, Kyowa Hakko, Meiji, Nikken Kagaku, Organon, Otsuka, Pfizer, and Yoshitomi. He was on research advisory board for Pfizer, Janssen, Mochida and Meiji, and is currently on research advisory board for Sekisui Chemicals. He has received royalties from Igaku-Shoin and Seiwa-Shoten

DOI: 10.1002/mpr

Publishers. None of the other authors have competing interests.

References

Brooks RT, Beard J, Steel Z (2006). Factor structure and interpretation of the K10. Psychol Assess 18: 62–70. DOI: 16594813.

Demyttenaere K, Bruffaerts R, Posada-Villa J, Gasquet I, Kovess V, Lepine JP, Angermeyer MC, Bernert S, de Girolamo G, Morosini P, Polidori G, Kikkawa T, Kawakami N, Ono Y, Takeshima T, Uda H, Karam EG, Fayyad JA, Karam AN, Mneimneh ZN, Medina-Mora ME, Borges G, Lara C, de Graaf R, Ormel J, Gureje O, Shen Y, Huang Y, Zhang M, Alonso J, Haro JM, Vilagut G, Bromet EJ, Gluzman S, Webb C, Kessler RC, Merikangas KR, Anthony JC, Von Korff MR, Wang PS, Brugha TS, Aguilar-Gaxiola S, Lee S, Heeringa S, Pennell BE, Zaslavsky AM, Ustun TB, Chatterji S (2004). Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. JAMA 291: 2581–90. DOI: 15173149.

Dohrenwend BP, Shrout PE, Egri G, Mendelsohn FS (1980). Nonspecific psychological distress and other dimensions of psychopathology. Measures for use in the general population. Arch Gen Psychiatry 37: 1229–36. DOI: 7436685.

Efron B, Tibshirani RJ (1993). An Introduction to Bootstrap. London: Chapman & Hall.

Fagan TJ (1975). Nomogram for Bayes theorem. N Eng J Med 293: 257. DOI: 1143310.

Furukawa TA, Goldberg DP, Rabe-Hesketh S, Ustun TB (2001). Stratum-specific likelihood ratios of two versions of the general health questionnaire. Psychol Med 31: 519–29. DOI: 11305860.

Furukawa TA, Kessler RC, Slade T, Andrews G (2003). The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being. Psychol Med 33: 357–62. DOI: 12622315.

Furukawa TA, Strauss S, Bucher HC, Guyatt G (2008). Diagnostic Tests. In Guyatt G, Drummond R, Meade MO et al. (eds) Users' Guides to the Medical Literature: A Manual for Evidence-Based Practice (2nd edn). New York: McGraw-Hill. pp. 419–38.

Goldberg D, Williams P (1998). A User's Guide to the General Health Questionnaire. Berkshire: NFER-NELSON.

Goldberg DP, Oldehinkel T, Ormel J (1998). Why GHQ threshold varies from one place to another. Psychol Med 28: 915–21. DOI: 9723146.

Hanley JA, McNeil BJ (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148: 839–43. DOI: 6878708.

Haro JM, Arbabzadeh-Bouchez S, Brugha TS, de Girolamo G, Guyer ME, Jin R, Lepine JP, Mazzi F, Reneses B, Vilagut G, Sampson NA, Kessler RC (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. Int J Methods Psychiatr Res 15: 167–80. DOI: 17266013.

Jovanovic BD (2005). Confidence intervals, binomial, when no events are observed. In Armitage P, Colton T (eds) Encyclopedia of Biostatistics. Chichester: Wiley, pp. 1103–4.

Kawakami N, Takeshima T, Ono Y, Uda H, Hata Y, Nakane Y, Nakane H, Iwata N, Furukawa TA, 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 Clin Neurosci 59: 441–52. DOI: 16048450.

Kessler RC, Ustun TB (2004). The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). Int J Methods Psychiatr Res 13: 93–121. DOI: 15297906.

Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SL, Walters EE, Zaslavsky AM (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 32: 959–76. DOI: 12214795.

Kessler RC, Barker PR, Colpe LJ, Epstein JF, Gfroerer JC, Hiripi E, Howes MJ, Normand SL, Manderscheid RW, Walters EE, Zaslavsky AM (2003). Screening for serious mental illness in the general population. Arch Gen Psychiatry 60: 184–9. DOI: 12578436.

Langner TS (1962). A twenty-two item screening score of psychiatric symptoms indicating impairment. J Health Human Behav 3: 269–76. DOI: 13928667.

MacMillan AM (1957). The Health Opinion Survey: techniques for estimating prevalence of psychoneurotic and related typyes of disorders in communities. Psychol Rep 3: 325–39.

Peirce JC, Cornell RG (1993). Integrating stratum-specific likelihood ratios with the analysis of ROC curves. Med Decis Making 13: 141–51. DOI: 8483399.

SAS Institute Inc (2004). SAS/STAT 9.1. Cary, NC: SAS Institute Inc.

SPSS Inc (2006). SPSS for Windows Version 15.0. Chicago, IL: SPSS Inc.

Van den Linden WJ, Hambleton RK (1997). Handbook of Modern Item Response Theory. New York: Springer.

Correspondence: Toshi A. Furukawa, Department of Psychiatry and Cognitive-Behavioral Medicine, Nagoya City University Graduate School of Medical Sciences, Mizuho-cho, Mizuho-ku, Nagoya 467-8601, Japan.
Tel: +81-52-853-8271
Fax: +81-52-852-0837
Email: furukawa@med.nagoya-cu.ac.jp

DOI: 10.1002/mpr