International Journal of Methods in Psychiatric Research
*Int. J. Methods Psychiatr. Res.* 19(Supplement 1): 4–22 (2010)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/mpr.310

# Screening for serious mental illness in the general population with the K6 screening scale: results from the WHO World Mental Health (WMH) survey initiative

RONALD C. KESSLER,[1] JENNIFER GREIF GREEN,[1] MICHAEL J. GRUBER,[1] NANCY A. SAMPSON,[1] EVELYN BROMET,[2] MARIUS CUITAN,[3] TOSHI A. FURUKAWA,[4] OYE GUREJE,[5] HRISTO HINKOV,[6] CHI-YI HU,[7] CARMEN LARA,[8] SING LEE,[9] ZEINA MNEIMNEH,[10] LANDON MYER,[11] MARK OAKLEY-BROWNE,[12] JOSE POSADA-VILLA,[13] RAJESH SAGAR,[14] MARIA CARMEN VIANA[15] & ALAN M. ZASLAVSKY[1]

1 Department of Health Care Policy, Harvard Medical School, Boston, MA, USA
2 Department of Psychiatry, State University of New York, Stony Brook, NY, USA
3 National School of Public Health & Health Services Management of Bucharest, Romania
4 Department of Psychiatry and Cognitive-Behavioral Medicine, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan
5 Department of Psychiatry, University College Hospital, Ibadan, Nigeria
6 Department of Global Mental Health, National Center for Public Health Protection, Sofia, Bulgaria
7 Shenzhen Institute of Mental Health & Shenzhen Kangning Hospital, Shenzhen, China
8 Department of Psychiatry, Autonomous University of Puebla, Puebla, Mexico
9 Department of Psychiatry, The Chinese University of Hong Kong, Shatin, Hong Kong, China
10 Institute for Social Research, University of Michigan, Ann Arbor, MI, USA and Institute for Development, Research, Advocacy, and Applied Care (IDRAAC), Beirut, Lebanon
11 School of Public Health & Family Medicine, University of Cape Town, Cape Town, South Africa
12 Discipline of Psychiatry, the University of Tasmania, Australia
13 Colegio Mayor de Cundinamarca University, Bogota, Colombia
14 Department of Psychiatry, All India Institute of Medical Sciences, New Delhi, India
15 Section of Psychiatric Epidemiology, Institute of Psychiatry, School of Medicine, University of São Paulo, São Paulo, Brazil

**Key words**
K6 screening scale, psychiatric epidemiology, serious mental illness (SMI)

**Correspondence**
Ronald C. Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston MA 02115, USA.
Telephone (+1) 617-432-3587
Fax (+1) 617-432-3588
Email: Kessler@hcp.med.harvard.edu

Received 8 March 2010; accepted 8 March 2010

**Abstract**

Data are reported on the background and performance of the K6 screening scale for serious mental illness (SMI) in the World Health Organization (WHO) World Mental Health (WMH) surveys. The K6 is a six-item scale developed to provide a brief valid screen for Diagnostic and Statistical Manual of Mental Disorders 4th edition (DSM-IV) SMI based on the criteria in the US ADAMHA Reorganization Act. Although methodological studies have documented good K6 validity in a number of countries, optimal scoring rules have never been proposed. Such rules are presented here based on analysis of K6 data in nationally or regionally representative WMH surveys in 14 countries (combined $N$ = 41,770 respondents). Twelve-month prevalence of DSM-IV SMI was assessed with the fully-structured WHO Composite International Diagnostic Interview. Nested logistic regression analysis was used to generate estimates of the predicted probability of SMI for each respondent from K6 scores, taking into consideration the possibility of variable concordance as a function of respondent age, gender, education, and country. Concordance, assessed by calculating the area under the receiver operating characteristic curve, was generally substantial (median 0.83; range 0.76–0.89; inter-quartile range 0.81–0.85). Based on this result, optimal scaling rules are presented for use by investigators working with the K6 scale in the countries studied. 

## Introduction

The purpose of the current report is to present rules for optimal scoring of the K6 screening scale of non-specific psychological distress (Kessler *et al.*, 2002, 2003), a widely-used short scale that screens for the presence of serious mental illness. As described below, the K6 was developed for use in community epidemiological needs assessment surveys in the USA but has subsequently been validated and used in surveys in a number of other countries. Optimal scoring rules have never before been proposed for the K6. The rules proposed in the current report are based on analyses of representative general population surveys carried out in 14 countries throughout the world in conjunction with the World Health Organization (WHO) World Mental Health (WMH) Survey Initiative (Kessler and Üstün, 2008). The scoring rules are provided separately for each country to convert K6 scores into predicted probabilities of serious mental illness.

It is important to acknowledge at the outset that it is preferable to base such rules on clinical calibration studies embedded in larger epidemiological surveys whenever possible. However, it is not possible to carry out a new clinical calibration study every time a scale is used. The scoring rules presented here are made available with that reality in mind for researchers who want to use optimal scoring rules based on community samples in their countries when independent calibration is not possible. The samples on which the current scoring rules are based range from a low of 1031 in Lebanon to a high of 5692 in the USA. The largest sample is from New Zealand ($N$ = 7435), but random half-samples in New Zealand were administered either of two versions of the scale described below, so only 3705–3730 respondents received each version. The combined sample size across all 14 countries is 41 770 respondents.

### Background

Dimensional scales of non-specific psychological distress have been used in community epidemiological surveys since the end of World War II, beginning with the 20-item Health Opinion Survey in the Stirling County Study (Leighton, 1975; MacMillan, 1957) and the 22-item Langner Scale in the Midtown Manhattan Study (Langner, 1962; Srole *et al.*, 1962). Although originally used as first-stage screens to target respondents with broadly defined emotional problems for more in-depth clinical assessment, these dimensional scales came to be used without clinical follow-up in later surveys (e.g. Myers *et al.*, 1975). Controversy regarding the appropriate cut-point for case thresholds on these scales in community surveys (e.g. Seiler, 1973) led in later surveys to scale scores being reported primarily in dimensional terms (e.g. means)

rather than in terms of proportions of respondents screening positive (e.g. Pearlin *et al.*, 1981).

Dimensional scales continue to be widely used to screen for mental illness in primary care (Coyne *et al.*, 2001) and to assess symptom severity and treatment effectiveness in clinical studies (Rush *et al.*, 2000). However, influenced by the widely published results of the Epidemiological Catchment Area Study, dimensional screening scales went out of vogue in community psychiatric epidemiology beginning in the early 1980s (Robins and Regier, 1991). Fully structured research diagnostic interviews administered by lay interviewers have become the standard measures of psychopathology in community epidemiological surveys since that time. A number of such structured diagnostic interviews now exist, including the Diagnostic Interview Schedule (Robins *et al.*, 1981), the Composite International Diagnostic Interview (CIDI) (Robins *et al.*, 1988), the PRIME-MD (Spitzer *et al.*, 1994); and the Mini-International Neuropsychiatric Interview (Sheehan *et al.*, 1998).

We now know, based on the use of fully structured research diagnostic interviews in a number of large community epidemiological surveys, that up to half the general population meet criteria for one or more lifetime International Classification of Diseases (ICD) or Diagnostic and Statistical Manual of Mental Disorders (DSM) disorders and up to one-fifth carry a DSM or ICD diagnosis at any one point in time (Kessler *et al.*, 2005b, 2007). Although the published reports of these high prevalence estimates were initially met with a good deal of skepticism, subsequent clinical calibration studies showed that they are accurate (Haro *et al.*, 2006; Kessler *et al.*, 1998), but that many community cases have considerably less severe disorders than those of cases in treatment (Demyttenaere *et al.*, 2004; Kessler *et al.*, 2005d). The finding that clinical severity is related to treatment is, of course, not surprising. However, given the high proportion of people in the population who meet criteria for a mental disorder in relation to the societal resources available for treatment, policy-oriented interpreters of the epidemiological evidence have called for (Regier *et al.*, 2000), and in some cases created (National Advisory Mental Health Council, 1993), distinctions to be made between people with severe and less severe mental disorders in an effort to define medical necessity for policy-planning purposes. For example, the US Substance Abuse and Mental Health Service Administration, which administers Block Grants to States to fund public mental health services for low-income people who are not otherwise insured, limits coverage to cases defined as having a Serious Mental Illness (SMI). The criteria for SMI require not only a DSM diagnosis but also specified indicators of severity that characterize fewer than one-third of the people in the US population who meet criteria for a current DSM-IV-R disorder (Kessler *et al.*, 2005d).

## Development of the K6

Dimensional measures of non-specific psychological distress have come to take on new importance in the context of this movement to distinguish community cases based on severity rather than purely on diagnosis. In particular, a number of recent large-scale community epidemiologic surveys have included brief screening scales to provide a rapid assessment of the prevalence of SMI. Included here are a number of large ongoing health tracking surveys carried out in the USA and Australia as well as large needs assessment surveys carried out in Europe and Asia. The most widely used screening scale of SMI in these studies is the K6 scale (Furukawa *et al.*, 2003; Kessler *et al.*, 2002, 2003), a six-question scale that was developed explicitly to estimate the prevalence of SMI as defined by US Public Law (PL) 102-321, the Alcohol, Drug Abuse, and Mental Health Administration (ADAMHA) Reorganization Act. This law established a US federal Block Grant for states to fund Community Mental Health Services for adults with SMI and the law required states to include incidence and prevalence estimates in their annual applications for Block Grant funds. The law also required the US Substance Abuse and Mental Health Services Administration (SAMHSA) to develop an operational definition of SMI and to create an estimation methodology based on this definition for use by the states. The definition of SMI stipulated in PL 102-321 requires the person to have at least one 12-month DSM disorder, other than a substance-use disorder, and to have 'serious impairment'.

Given the importance for policy-planning purposes of knowing the prevalence and socio-demographic distribution of SMI in the US population for purposes of allocating Block Grant funds (which are in excess of $1 billion each year), the architects of all major US federal health-tracking surveys decided to include a measure of SMI in their interviews shortly after the ADAMHA Reorganization Act was published. The K6 was developed for this purpose to be included in the US National Health Interview Survey (NHIS), a national survey of close to 50 000 households that has been carried out on an ongoing basis in the USA for more than half a century (www.cdc.gov/nchs/nhis.htm). The goal was to create a very brief (6–10 items) scale that would provide accurate aggregate estimates of SMI prevalence and correlates. Although a

number of distress scales existed that had been used for many year as of the time the K6 was developed (Derogatis, 1983; Dohrenwend *et al.*, 1980; Gurin *et al.*, 1960), only a few of them were brief enough to meet this time requirement (Pearlin *et al.*, 1981; Ware and Sherbourne, 1992) and none was developed using modern psychometric methods to maximize precision in the clinical range of the population distribution (van der Linden and Hambleton, 1997). Based on these considerations, the decision was made to develop a new screening scale for use in the redesigned NHIS.

The conceptualization of this task relied importantly on the work of Bruce Dohrenwend and his colleagues (Dohrenwend *et al.*, 1980; Link and Dohrenwend, 1980). Their review of screening scales of non-specific psychological distress showed that these scales typically include 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. However, despite this heterogeneous content, the vast majority of the symptoms in these scales have high factor loadings on a first principal factor. People with a wide range of mental disorders typically have high scores on this core dimension of non-specific distress. Based on this result, this core dimension of non-specific psychological distress was taken as the focus of the new scale. Because the requirements called for a very short scale, the K6 was developed using modern psychometric methods to select questions with the maximum precision at the clinical threshold of the scale. Based on the fact that no more than 10%, and probably closer to 6%, of the US population were estimated to meet criteria for SMI in a given year (Kessler *et al.*, 1996), the decision was made at the onset to seek maximum precision around the 90th percentile of the general population distribution.

## K6 validation studies

As detailed elsewhere (Kessler *et al.*, 2002, 2003), independent clinical validation studies showed that the K6 has very good concordance with blinded clinical diagnoses of SMI in general population samples from the USA. Based on this evidence, the other two major ongoing national health tracking surveys in the USA, the CDC Behavioral Risk Factors Surveillance Survey (www.cdc.gov/BRFSS) and the SAMHSA National Household Survey on Drug Use and Health (www.oas.samhsa.gov/nhsda.htm), both adopted the K6 as part of their assessment of health shortly thereafter. Taken together, these three ongoing US surveys interview representative samples of nearly 500 000 people each year, creating the potential for making them the largest tracking series on the prevalence and correlates of SMI in the world.

Based on the adoption of the K6 by these three large ongoing US federal health tracking surveys, K6 validation studies were carried out in a number of other countries throughout the world (Fassaert *et al.*, 2009; Furukawa *et al.*, 2003, 2008; Gill *et al.*, 2007; Patel *et al.*, 2008). These studies uniformly found the K6 and a larger related scale known as the K10 (which includes the K6 in addition to four other items) to have very good concordance with independent clinical ratings of SMI. These studies also found, consistent with results in the USA, that the K6 performed as well as the K10, leading to the recommendation that the six-item version be used rather than the 10-item version. Additional studies found similarly good concordance in special patient populations that included primary-care attenders (Haller *et al.*, 2009), postnatal females (Baggaley *et al.*, 2007; Tesfaye *et al.*, 2009), and patients with substance-use disorders (Hides *et al.*, 2007; Swartz and Lurigio, 2006). Methodological research also showed that the K6 has little bias with regard to sex and education (Baillie, 2005), a feature that was built into the scale from the outset, as items were selected for the K6 based on formal comparisons of age, sex, and education differences in differential item functioning to minimize biases with regard to these variables (Kessler *et al.*, 2002).

## Alternative approaches to K6 scoring

The widespread adoption of the K6 in epidemiological surveys throughout the world is based on the very good performance of this short screening in the validity studies reviewed in the last paragraph. However, despite this wide use, no clear standards have yet emerged for optimal K6 scoring. As each scale item has five categories and there are six items, the unweighted scale has values in the range 0–24. The scoring rule used in most applications based on standard validation studies is to classify respondents with scores of 13–24 as having probable SMI and those with scores of 0–12 as probably not having SMI (Kessler *et al.*, 2003). However, Furukawa and colleagues (Furukawa *et al.*, 2003, 2008) have shown that this simple dichotomous scoring approach can be refined by using polychotomous rather than dichotomous scoring rules that collapse K6 scores into strata based on analysis of data in a clinical calibration study such that the observed prevalence of SMI differs significantly across strata. For example, one such scoring rule might collapse K6 scores into strata with K6 score values of 0, 1–7, 8–12, 13–18, and 19–24, with respondents in each stratum assigned a

predicted probability of SMI based on the results of a clinical calibration study.

Rather than interpret the precision of the K6 in terms of sensitivity (the proportion of true cases who are detected in the screening scale) and specificity (the percentage of true non-cases who are correctly classified as non-cases by the screening scale) based on a single diagnostic threshold, as in the dichotomous approach, the stratum-specific predicted probabilities generated in this polychotomous approach can be assigned as outcome variable scores and used directly for purposes of estimating prevalence and studying correlates. In other words, each respondent's K6 score is transformed into a score in the range 0.0 to 1.0 that represents the predicted probability of having SMI.

Furukawa and associates proposed to use the K6 in clinical screening to assign individual patients predicted scores of SMI based on the method of Stratum-Specific Likelihood Ratio (SSLR) analysis, a method that begins with estimates of sensitivity and specificity for each K6 stratum and estimates each patient's predicted probability of SMI based on external assumptions about prevalence in the population of interest (Guyatt and Rennie, 2001). The use of this approach is based on the assumption that sensitivity and specificity are more stable across populations than is positive predictive value (PPV; the prevalence of SMI among respondents with a given K6 score), an assumption that is widely accepted in the methodological literature on medical decision-making (Rao, 2006). When this assumption holds, PPV for given values of sensitivity and specificity depends on the prevalence of the disorder in the population being screened, making it necessary either to obtain independent data on this prevalence or to make an informed assumption about this prevalence before estimating PPV from data on sensitivity and specificity. The SSLR approach provides a convenient way to do this when sensitivity and specificity are assumed to be known (presumably based on a previous clinical calibration study) and prevalence is estimated externally. Examples of using SSLR analysis in this way are reported in the literature (Furukawa *et al.*, 2001, 2002).

We propose in the next paragraph a different approach than SSLR analysis for use in epidemiological surveys. However, it should be noted that it is possible to use maximum-likelihood methods to estimate prevalence in clinical situations to avoid the requirement suggested by Furukawa and colleagues of estimating prevalence based on external information. In brief, in a clinical situation where SSLR analysis is being used and sample data on K6 scores are available from the clinical population from which an individual patient comes, a predicted distribution of K6 scores can be generated for every possible value of SMI prevalence based on the known (presumably from some independent clinical calibration sample that is assumed to apply to the population) sensitivity and specificity of each K6 score or category. The SMI prevalence estimate that generates a predicted K6 distribution most closely approximating the observed distribution in the clinical population from which the patient comes is the maximum likelihood estimate of prevalence in that population based on the assumption that the sensitivity and specificity estimates actually apply to that population.

The weakness of this approach is that it requires assumptions to be made about the values of sensitivity and specificity. A preferable approach would be to embed a clinical calibration study in the data collection so as to estimate PPV directly rather than have to estimate PPV and prevalence based on external data using a pre-established set of estimates of sensitivity and specificity. This kind of internal clinical calibration study is a common feature of psychiatric epidemiological surveys (Haro *et al.*, 2006; Kessler and Üstün, 2004), where a probability sub-sample of survey respondents that over-samples screened positives is re-interviewed by clinical interviewers who make diagnoses blinded to the K6 scores in the main survey. Once data of this sort are available, the SSLR approach can be expanded to use K6 scores along with measures of other predictors of SMI, such as socio-demographic variables, in a multiple regression analysis within the clinical calibration sub-sample (appropriately weighted to adjust for the over-sampling of respondents with high K6 scores) that explores both the functional form of the association between K6 scores and SMI and the possibility that this association varies as a function of the respondent's age, sex, education, or other characteristics. When a best-fitting model is found, a predicted probability of SMI based on this model can be assigned to each sample respondent. These predicted probabilities can then be used to estimate prevalence and correlates of SMI. As noted in the introduction, the purpose of the current report is to present scoring rules based on such analyses of general population survey data obtained in the WHO WMH Survey Initiative (Kessler and Üstün, 2008).

### Adjusting for the imprecision of estimates

In providing these transformation rules, it is important to recognize that the uncertainty of inference from the prediction equations needs to be taken into consideration in analysing estimates of the prevalence and correlates

of SMI based on transformed K6 data. Conventional significance testing would treat the individual-level predicted probabilities of SMI as known rather than estimated from a model. The method of multiple imputation (MI) (Rubin, 1987) can be used to overcome this limitation by generating a number of different estimates of the predicted probability of SMI for each respondent and using information about variation across these predictions to adjust estimates of standard errors for imprecision. In our use of the approach, this was done by estimating the final prediction equation 10 times, once in each of 10 pseudo-samples. Each pseudo-sample consisted of a random sample of respondents equal to the actual sample size, but selected *with replacement* from the actual sample. The with-replacement option means that some respondents in the actual sample were included zero times, others once, and others more than once in each pseudo-sample. The precise values of the regression coefficients varied across pseudo-samples because of this variation in sample composition.

The MI method requires us to make all estimates 10 times, once in each pseudo-sample, and then to combine these estimates in such a way as to account both for between-person variation and for within-person variation. The MI parameter estimates are defined as the means across the 10 pseudo-samples of the within-sample estimates and the MI standard error of any given parameter estimate is then defined as the square root of the sum of two components. The first component is the mean of the square of the 10 within-sample standard errors (i.e. the between-person variance component). The second component is a transformation of the variance of the parameter estimates across the 10 samples (i.e. the within-person variance component). In the extreme case where the K6 is totally unrelated to SMI in a particular population, the only systematic information in the multiply imputed dataset will be the consistent 0.0 and 1.0 values in the sub-sample of respondents who were in the clinical calibration sub-sample.

The expected value of predicted disorder prevalence for each respondent who was not in the clinical calibration sub-sample will be the SMI prevalence in the clinical calibration sub-sample. In a case of this sort, the MI predicted SMI prevalence estimate will be unbiased and the standard error of the estimate will be equivalent to the design-based standard error in the clinical calibration sub-sample. At the other extreme, where the K6 perfectly predicts SMI, the MI standard error of the SMI prevalence estimate will be equivalent to the design-based standard error in the total sample. In more realistic cases, in which concordance between the K6 and the clinical diagnoses is significant but imperfect, the MI standard error will take into consideration both the size of the clinical calibration sub-sample and the strength of the association between the K6 and clinical diagnoses. The situation is similar for higher-order statistics, with the exception that measures of association will be biased towards zero by lack of concordance between predicted and true SMI diagnoses. The practical use of this approach is illustrated in a more detailed methodological exposition published previously (Kessler and Üstün, 2004) as well as in a number of subsequent substantive reports that used this approach to estimate the prevalence and correlates of several different DSM-IV disorders (Fayyad *et al.*, 2007; Huang *et al.*, 2009; Kessler *et al.*, 2005c).

To allow researchers to implement this MI approach to estimation when they use our transformation rules to score K6 responses, we generated 10 pseudo-samples for each of the 14 countries in the WMH series and then estimated the coefficients for the best-fitting prediction equation for the country (which was developed in analysis of the original sample rather than pseudo-samples) separately in each of those pseudo-samples. These 10 separate sets of coefficients are provided in appendix tables for each of the 14 countries (available at http://www.hcp.med.harvard.edu/wmh/publications.php). The remainder of the paper describes the methods used to carry out the analyses that selected the best-fitting equations, presents descriptive statistics describing the accuracy of these equations, and discusses a number of special substantive analysis issues in working with data of the sort generated by these MI methods.

## Methods

### Samples

The WMH surveys were carried out in 14 countries in Africa (Nigeria, South Africa), the Americas (Brazil, Colombia, Mexico, USA), Asia and the Pacific [India, Japan, New Zealand, and separate surveys in Beijing, Shanghai, and Shenzhen in the People's Republic of China (PRC), described below as Metropolitan PRC], Europe (Bulgaria, Romania, Ukraine), and the Middle East (Lebanon) (Table 1). Eleven of these countries are classified by the World Bank as less developed (Brazil, Bulgaria, China, Colombia, India, Lebanon, Mexico, Nigeria, Romania, South Africa, Ukraine). The others are developed. Country-level sample sizes range from a low of 1031 (Lebanon) to a high of 5692 (US). The total sample size is 41,770. The weighted average cross-national response rate was 75.5% with a range between 55.1% (Japan) and 98.8% (India).

**Table 1** WMH Sample Characteristics

| Country | Survey[1] | Sample characteristics[2] | Field dates | Age range | Sample size[3] | Response rate[4] |
|---|---|---|---|---|---|---|
| Brazil | São Paulo Megacity | Stratified multistage clustered area probability sample of household residents in the São Paulo metropolitan area | 2005–7 | 18+ | 2942 | 81.3 |
| Bulgaria | NSHS | Stratified multistage clustered area probability sample of household residents. NR | 2003–7 | 18+ | 2233 | 72.0 |
| Colombia | NSMH | Stratified multistage clustered area probability sample of household residents in all urban areas of the country (approximately 73% of the total national population) | 2003 | 18–65 | 2381 | 87.7 |
| India | WMHI | Stratified multistage clustered area probability sample of household residents in Pondicherry region. NR | 2003–5 | 18+ | 1373 | 98.8 |
| Japan | WMHJ 2002–2006 | Unclustered two-stage probability sample of individuals residing in households in 11 metropolitan areas | 2002–6 | 20+ | 1682 | 55.1 |
| Lebanon | LEBANON | Stratified multistage clustered area probability sample of household residents. NR | 2002–3 | 18+ | 1031 | 70.0 |
| Mexico | M-NCS | Stratified multistage clustered area probability sample of household residents in all urban areas of the country (approximately 75% of the total national population) | 2001–2 | 18–65 | 2362 | 76.6 |
| New Zealand[5] | NZMHS | Stratified multistage clustered area probability sample of household residents. NR | 2004–5 | 18+ | 7435 | 73.3 |
| Nigeria | NSMHW | Stratified multistage clustered area probability sample of households in 21 of the 36 states in the country, representing 57% of the national population. The surveys were conducted in Yoruba, Igbo, Hausa and Efik languages | 2002–3 | 18+ | 2143 | 79.3 |
| People's Republic of China | B-WMH S-WMH | Stratified multistage clustered area probability sample of household residents in the Beijing and Shanghai metropolitan areas | 2002–3 | 18+ | 1628 | 74.7 |
| People's Republic of China | Shenzhen | Stratified multistage clustered area probability sample of household residents and temporary residents in the Shenzhen area | 2006–7 | 18+ | 2476 | 80.0 |
| Romania | RMHS | Stratified multistage clustered area probability sample of household residents. NR | 2005–6 | 18+ | 2357 | 70.9 |
| South Africa | SASH | Stratified multistage clustered area probability sample of household residents. NR | 2003–4 | 18+ | 4315 | 87.1 |
| Ukraine | CMDPSD | Stratified multistage clustered area probability sample of household residents. NR | 2002 | 18+ | 1720 | 78.3 |
| United States | NCS-R | Stratified multistage clustered area probability sample of household residents. NR | 2002–3 | 18+ | 5692 | 70.9 |
| Total | | | | | 41770 | |

[1]NSHS (Bulgaria National Survey of Health and Stress); NSMH (The Colombian National Study of Mental Health); WMHI (World Mental Health India); WMHJ 2002–2006 (World Mental Health Japan Survey); LEBANON (Lebanese Evaluation of the Burden of Ailments and Needs of the Nation); M-NCS (The Mexico National Comorbidity Survey); NZMHS (New Zealand Mental Health Survey); NSMHW (The Nigerian Survey of Mental Health and Wellbeing); B-WMH (The Beijing World Mental Health Survey); S-WMH (The Shanghai World Mental Health Survey); RMHS (Romania Mental Health Survey); SASH (South Africa Health Survey); CMDPSD (Comorbid Mental Disorders during Periods of Social Disruption); NCS-R (The US National Comorbidity Survey Replication).

[2]Most WMH surveys are based on stratified multistage clustered area probability household samples in which samples of areas equivalent to counties or municipalities in the USA were selected in the first stage followed by one or more subsequent stages of geographic sampling (e.g. towns within counties, blocks within towns, households within blocks) to arrive at a sample of households, in each of which a listing of household members was created and one or two people were selected from this listing to be interviewed. No substitution was allowed when the originally sampled household resident could not be interviewed. These household samples were selected from census area data in all countries. The Japanese sample is the only totally unclustered sample, with households randomly selected in each of the 11 sample areas and one random respondent selected in each sample household. Of the 15 surveys, 13 are based on nationally representative (NR) household samples, while two others are based on nationally representative household samples in urbanized areas (Colombia, Mexico).

[3]Aa noted in the text, the WMH surveys were administered in two parts in all countries other than Romania and South Africa. The K6 was administered in Part 2. The full sample was administered Part 1, while Part 2 was administered to 100% of the Part 1 respondents who had a disorder assessed in Part 1 plus a probability sub-sample of other Part 1 respondents. The Part 2 sample was weighted to adjust for the undersampling of the Part 1 respondents who did not have a disorder. The sample sizes reported here are the unweighted numbers of respondents in the Part 2 sample.

[4]The response rate is calculated as the ratio of the number of households in which an interview was completed to the number of households originally sampled, excluding from the denominator households known not to be eligible either because of being vacant at the time of initial contact or because the residents were unable to speak the designated languages of the survey. The weighted average response rate is 75.5%.

[5]New Zealand interviewed respondents 16+ but for the purposes of cross-national comparisons we limit the sample to those 18+.


Int. J. Methods Psychiatr. Res. 19(Supplement 1): 4–22 (2010). DOI: 10.1002/mpr

All surveys were based on multi-stage geographically clustered area probability household samples. Interviews were carried out face-to-face by trained lay interviewers. Surveys in 10 countries were based on nationally representative samples, while two others were based on nationally representative samples of urbanized areas (Colombia, Mexico), and the other two on regional samples (Brazil, PRC). Respondents had to be at least 18 years of age in most countries (20 in Japan). Colombia and Mexico were the only countries with an upper age limit (65). Informed consent was obtained using procedures approved by local Institutional Review Boards. Detailed descriptions of WMH sampling, recruitment, and consent procedures are presented elsewhere (Heeringa *et al.*, 2008).

Other than in Romania and South Africa, where all respondents were administered the full interview, internal sub-sampling was used to reduce respondent burden by dividing the interview into two parts. Part 1 assessed core mental disorders and was administered to all respondents. Part 2 included additional disorders and correlates and was administered to all Part 1 respondents who met criteria for any lifetime Part 1 disorder plus a probability sub-sample of other respondents. The K6 was included in Part 2. Part 1 data were weighted to adjust for differential probabilities of selection and to match population distributions on census socio-demographic and geographic distributions. Part 2 data were additionally weighted for the under-sampling of Part 1 respondents without core disorders. WMH weighting procedures are discussed elsewhere (Heeringa *et al.*, 2008).

## Measures

### Diagnostic assessment

Lifetime and 12-month prevalence of DSM-IV anxiety, mood, behavioral, and substance disorders were assessed using Version 3.0 of the WHO CIDI (Kessler and Üstün, 2004), a fully structured lay-administered interview. The English source version of the CIDI was translated into other languages using standardized WHO protocols (Harkness *et al.*, 2008). Rigorous interviewer training and quality control monitoring were used to guarantee consistent administration (Pennell *et al.*, 2008). CIDI diagnoses were compared with blinded clinical diagnoses using the Structured Clinical Interview for DSM-IV (SCID) (Spitzer *et al.*, 1994) in probability sub-samples of WMH respondents from France, Italy, Spain, and the USA. As detailed elsewhere, good CIDI-SCID diagnostic concordance was found for most DM-IV/CIDI disorders (Haro *et al.*, 2006).

### SMI

Respondents were classified as having SMI in the 12 months before interview if they met criteria for one or more 12-month DSM-IV/CIDI mental disorders and also had any of a number of indicators of severity, which included a 12-month suicide attempt with serious lethality of intent, work disability, or substantial limitation as the result of a mental disorder, bipolar I disorder, a behavioral disorder with associated serious violence or criminal behavior, or any disorder that resulted in 30+ days out of role in the year. A more detailed description of the SMI coding scheme is presented elsewhere (Demyttenaere *et al.*, 2004).

### The K6

The K6 consists of six questions that ask subjects to rate how often they felt

1 nervous
2 hopeless
3 restless or fidgety
4 so depressed that nothing could cheer you up
5 that everything was an effort
6 worthless over one of two recall periods: the past-month (respondents were asked to rate how often the symptoms occurred in the 30 days before the survey) and the worst-month (respondents were asked about the 30-day period during the past 12 months when they had the most severe psychological distress).

Some WMH surveys used only one of these recall periods while others used both. The decision about which recall period to used hinged on whether the investigators were interested in calibrating SMI point prevalence (most useful for screening in clinical settings), 12-month prevalence (most useful for estimating prevalence in surveys used for health-policy planning purposes, as the year is the usual health-policy planning period), or both.

The surveys that used both recall periods began by administering the past-month questions and then asked respondents a single question about whether there was any other 30-day period in the past 12 months when they had had these symptoms more frequently than in the past 30 days. If not, then the past-month responses were also used as the worst-month responses. However, if the respondents reported that there was a worst month, they were asked to think about that time in answering the six questions a second time. The six K6 questions had to be repeated for about 20% of respondents when this two-part approach was used. That is, about 80% of the time respondents reported that there was no other 30-day

period in the past 12 months that was worse than the last 30 days. The response options, which were identical in the two recall periods, were all of the time, most of the time, some of the time, a little of the time, and none of the time. These were coded 4 to 0, which means that the unweighted summary scale has a 0–24 range. However, it is also possible to weight either the scale items (as in a factor analysis factor-weighted scale; Kim, 1993) or to weight the item responses within each item [as in an analysis of nested dichotomous items in an item response theory (IRT) modeling approach; Embretson and Reise, 2000].

### Socio-demographics

We considered three dichotomously scored socio-demographic variables in the analysis: gender, age (18–38, 39+), and education (completed no more than secondary education, completed more than secondary education). All three were used to predict SMI both alone and in interaction with the others. Interactions of K6 scores were evaluated with each socio-demographic alone as well as with the cross-classification of the full set of socio-demographics.

## Statistical analysis

Six-by-six matrices of Pearson correlations among the K6 item responses were created for each recall period in each country. Principal axis factor analysis using the FACTOR procedure in SAS 9.1 (SAS Institute Inc., 1999) was carried out to determine if the unidimensionality found in the original US psychometric studies was confirmed in each country in the WMH data. Parallel factor analyses were then carried out based on matrices of polychoric correlations, which allow for non-linear monotonic relationships between pairs of variables. The factor analysis results showed that the scale is unidimensional (i.e. has a large first unrotated eigenvalue and a second unrotated eigenvalue less than 1.0) and has factor loadings on the first factor that are quite similar across items in both Pearson and polychoric matrices. This last result means that it would be unproductive to create a factor-weighted scale rather than a simple 0–24 unweighted scale.

However, polychoric correlations were generally somewhat larger than Pearson correlations, suggesting that meaningful non-linearities might exist in the associations between items. The implications of this were investigated by estimating IRT models using the BILOG-MG program (Scientific Software International, 1996) based on nested dichotomous versions of the K6 items. By the term 'nested dichotomies' we mean that the 0–4 responses to each item were converted into four dichotomies (0 versus 1–4, 0–1 versus 2–4, 0–2 versus 3–4, 0–3 versus 4). Unlike classical psychometric test theory models, the IRT models allowed us to capture information about the contribution of each item to the sensitivity of the total scale using conventional one-parameter and two-parameter IRT logistic regression models for binary scale items (van der Linden and Hambleton, 1997). The two-parameter model is given by:

$$P_{ij}(T) = \left[1 + e^{-a_j(\mathrm{TPD}_i - b_j)}\right]^{-1} \quad (1)$$

Where the outcome variable $P_{ij}(T_i)$ is the probability that respondent i will endorse binary item j as a function of his or her underlying true score ($T$). The slope $a_j$ measures the steepness of the logistic curve at the point where the probability of endorsing item j is 0.5. A steep curve means that the item has strong discriminating ability at the point on the curve where it has maximum information value. The intercept $b_j$ is the point on the $T$ distribution at which the probability of endorsing item j is 0.5, therefore representing the severity of the item. The one-parameter model differs from the two-parameter model in that the parameter for the steepness of the slope is constrained to be constant across items. Inspection of a wide range of items in such models was used in the initial development phase to select the final set of items to include in the K6 (Kessler *et al.*, 2002). In the current application, the models were used to generate an optimally weighted version of the scale in each country by summing the item slopes for each endorsed item. When the item parameters are fixed, as they would be when results in a benchmarking survey are used to define the metric of the scale in later surveys, this score is a sufficient statistic for the person parameter ($T$).

A series of nested logistic regression equations was then estimated to predict SMI in each time frame in each country using either the unweighted or IRT-weighted versions of the K6 scale along with controls for age, sex, and education. The equations explored the existence of non-linearities in the association of K6 scores with log-odds of SMI by including not only linear but also quadratic and third-degree forms of the K6 as predictors (i.e. K6, K6-squared, and K6-cubed all as predictors in the same equation). We also evaluated the significant of interactions between K6 scores and the socio-demographic variables. Model fit was evaluated using the Akaike Information Criterion and the Bayesian Information Criterion, two commonly used methods to select best-fitting models (Burnham and Anderson, 2004).

Once a best-fitting model was determined, parameters for that model were used to generate an estimate of the predicted probability of SMI for each respondent in the


Int. J. Methods Psychiatr. Res. 19(Supplement 1): 4–22 (2010). DOI: 10.1002/mpr

WMH survey. That variable was then compared with observed SMI scores using receiver operating characteristic curve analysis (Margolis *et al.*, 2002) and the area under the receiver operating characteristic curve (AUC) (Pepe, 2003) was calculated as a measure of concordance between predicted and observed SMI scores. The AUC can be interpreted as the probability of correctly identifying a case of SMI in a series of paired comparison tests in which scores on the K6-tranformed predicted probability scale are compared between one randomly selected respondent with SMI and one randomly selected respondent without SMI and the respondent with the higher score is estimated to be the one with SMI. In cases where the predicted probabilities of the two respondents are identical, the estimate of which one has SMI is based on random assignment. The AUC has an expected value of 0.50 when the predicted probability is completely unrelated to the true SMI and an expected value of 1.0 when the predicted probability is perfectly related to true SMI. Scores between these two extremes are often interpreted in parallel with the interpretation of Kappa (Landis and Koch, 1977) as slight (0.5–0.6), fair (0.6–0.7), moderate (0.7–0.8), substantial (0.8–0.9), or almost perfect (0.9+).

Once the best-fitting model was selected in each survey, the parameters of the best-fitting model were estimated again in each of 10 pseudo-samples selected with replacement from the sample in the WMH survey for the country. The parameter values for these 10 equations are presented for each recall period for each country in appendix tables (available at http://www.hcp.med.harvard.edu/wmh/publications.php) that can be used by other investigators to convert K6 scores in their samples into predicted probabilities of SMI. Ten different estimates are presented to allow researchers to generate 10 different estimates of SMI in their data for use in MI analysis. A brief exposition of appropriate analysis methods in using these MI estimates is presented below in the Discussion section.

## Results

### Dimensionality and consistency of factor loadings

Exploratory factor analysis showed a strong unidimensional structure in both the Pearson and polychoric correlation matrices in all countries for both past-month and worst-month recall periods. In the case of the Pearson correlation matrices, the unrotated eigenvalues were in the range 2.0–3.8 for the first factor, but were also generally greater than 1.0 for the second factor (1.2–1.5). However, promax rotation found no consistently interpretable second factor in the Pearson data. Furthermore, once we adjusted for non-linearities in the polychoric matrices, the unrotated eigenvalues of the first factor consistently increased (3.7–5.0) and with one exception the unrotated eigenvalues of the second factor became less than 1.0 (0.4–0.8) in both time frames (Table 2). The exception was in India, where the eigenvalues for the first two factors were 4.0 and 1.2 in the past-month time frame and 4.2 and 1.1 in the worst-month time frame. Inspection of factor loadings for the promax rotated second factor found this to be a unique factor for the 'everything was an effort' item. This result was due to a very low correlation (0.06 in the worst-month time frame) of the 'everything was an effort' item with the 'restless-fidgety' item and lower correlations of the 'everything was an effort' item than the 'restless- fidgety' item with the other four items (0.32–0.61 versus 0.75–0.89).

The low correlation between the 'everything was an effort' item and the 'restless-fidgety' item might be

**Table 2** Eigenvalues for first two unrotated factors from principal axis factor analysis of the polychoric correlation matrix of K6 items separately for past-month and worst-month recall periods[1]

| | Past-month | | Worst-month | |
|---|---|---|---|---|
| | I | II | I | II |
| Brazil | 4.1 | 0.6 | 4.4 | 0.6 |
| Bulgaria | 4.5 | 0.7 | 4.7 | 0.7 |
| Colombia | 4.2 | 0.6 | 4.3 | 0.6 |
| India | 4.0 | 1.2 | 4.2 | 1.1 |
| Japan | 3.9 | 0.7 | 4.4 | 0.6 |
| Lebanon | 3.9 | 0.7 | – | – |
| Mexico | – | – | 4.8 | 0.5 |
| New Zealand | 4.3 | 0.6 | 4.5 | 0.5 |
| Nigeria | 4.2 | 0.6 | – | – |
| People's Republic of China | | | | |
| Beijing/Shanghai | 3.9 | 0.8 | – | – |
| Shenzhen | 3.9 | 0.8 | 4.0 | 0.8 |
| Romania | 4.9 | 0.4 | 5.0 | 0.4 |
| South Africa | 4.1 | 0.7 | – | – |
| Ukraine | 3.7 | 0.7 | – | – |
| USA | – | – | 4.2 | 0.6 |
| Total | 4.2 | 0.6 | 4.4 | 0.6 |

[1] Some countries used only one recall period. This is why there are missing values in some cells of the table. The results in the total row are based on analysis of pooled within-country matrices that weight countries by the number of respondents in their samples rather than by their population sizes.

**Table 3** Unrotated factor loadings for the first factor from principal axis factor analysis of the polychoric correlation matrix of K6 items separately for past-month (P) and worst-month (W) recall periods[1]

| | Nervous | | Hopeless | | Restless | | Depressed | | Effort | | Worthless | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | W | P | W | P | W | P | W | P | W | P | W |
| Brazil | 0.80 | 0.82 | 0.85 | 0.88 | 0.78 | 0.81 | 0.90 | 0.92 | 0.81 | 0.85 | 0.84 | 0.85 |
| Bulgaria | 0.75 | 0.80 | 0.91 | 0.93 | 0.81 | 0.82 | 0.93 | 0.94 | 0.91 | 0.92 | 0.87 | 0.87 |
| Colombia | 0.75 | 0.76 | 0.83 | 0.85 | 0.81 | 0.82 | 0.87 | 0.88 | 0.87 | 0.88 | 0.90 | 0.91 |
| India | 0.86 | 0.93 | 0.97 | 0.96 | 0.82 | 0.81 | 0.88 | 0.86 | 0.43 | 0.50 | 0.84 | 0.89 |
| Japan | 0.72 | 0.80 | 0.87 | 0.92 | 0.76 | 0.82 | 0.90 | 0.93 | 0.78 | 0.85 | 0.78 | 0.82 |
| Lebanon | 0.71 | – | 0.85 | – | 0.81 | – | 0.90 | – | 0.70 | – | 0.85 | – |
| Mexico | – | 0.82 | – | 0.92 | – | 0.90 | – | 0.93 | – | 0.92 | – | 0.90 |
| New Zealand | 0.75 | 0.80 | 0.92 | 0.90 | 0.78 | 0.79 | 0.91 | 0.92 | 0.82 | 0.86 | 0.90 | 0.92 |
| Nigeria | 0.79 | – | 0.87 | – | 0.85 | – | 0.88 | – | 0.78 | – | 0.84 | – |
| People's Republic of China | | | | | | | | | | | | |
| Beijing/Shanghai | 0.68 | – | 0.85 | – | 0.79 | – | 0.92 | – | 0.78 | – | 0.82 | – |
| Shenzhen | 0.80 | 0.79 | 0.87 | 0.87 | 0.76 | 0.77 | 0.85 | 0.86 | 0.73 | 0.78 | 0.79 | 0.80 |
| Romania | 0.82 | 0.83 | 0.93 | 0.94 | 0.90 | 0.90 | 0.95 | 0.95 | 0.89 | 0.89 | 0.93 | 0.94 |
| South Africa | 0.77 | – | 0.87 | – | 0.82 | – | 0.86 | – | 0.76 | – | 0.85 | – |
| Ukraine | 0.71 | – | 0.84 | – | 0.77 | – | 0.86 | – | 0.68 | – | 0.85 | – |
| USA | – | 0.76 | – | 0.89 | – | 0.78 | – | 0.90 | – | 0.82 | – | 0.89 |
| Total | 0.76 | 0.80 | 0.88 | 0.89 | 0.80 | 0.82 | 0.89 | 0.91 | 0.80 | 0.84 | 0.86 | 0.88 |

[1] The K6 asked respondents to rate how often they felt (1) *nervous*, (2) *hopeless*, (3) *restless or fidgety*, (4) *so depressed that nothing could cheer you up*, (5) *that everything was an effort*, and (6) *worthless* over one of two recall periods: the *past-month* (the P columns in the table; respondents were asked to rate how often the symptoms occurred in the 30 days before the survey) and the *worst-month* (the W columns in the table; respondents are asked about the 30-day period during the past 12 months when they had the most severe psychological distress). The response options, which were identical in the two recall periods, were *all of the time*, *most of the time*, *some of the time*, *a little of the time*, and *none of the time*. Some WMH surveys used only one of these recall periods while others used both. This is why there are missing values in some cells of the table. The results in the total row are based on analysis of pooled within-country matrices that weight countries by the number of respondents in their samples rather than by their population sizes.

expected because the first is associated with retarded symptoms and the second with agitated symptoms. Indeed, this correlation is generally the lowest one in both the Pearson and polychoric matrices across countries. Despite this fact, though, the factor loadings of both these items are acceptable in both the past-month and worst-month time frames in all countries, including India (Table 3). Even in India, where 'everything was an effort' formed a separate factor, the factor loading of this item in the one-factor solution was acceptable (0.50). It is also noteworthy that variation in factor loadings was relatively small in all countries other than India in both the past-month and worst-month time frames, with within-country ranges of 0.10–0.24 (past-month) and 0.10–0.15 (worst-month). This implies that factor-weighted scales would be very highly correlated with unit-weighted scales. We found this to be the case empirically, with Pearson correlations above 0.95 between the two kinds of scales in each country, leading us to focus on the unit-weighted (i.e. 0–24) scale in the remainder of the analysis.

## The IRT analysis

It was noted above that the polychoric correlations were generally larger than the Pearson correlation, indicating that meaningful non-linearities exist in the associations among K6 items. Consistent with this observation, the IRT severity parameters were in most cases found to be non-linear across the range of item responses. This was true both in one-parameter IRT models (i.e. models in which slopes were constrained to be equal across items) and in two-parameter IRT models (i.e. where slopes were estimated separately for each item). Indeed, severity parameters were very similar in one-parameter and two-parameter models because all slopes were excellent. In saying this, we note that good IRT slopes are generally considered ones

**Table 4** Severity parameters for the two-parameter worst-month Item Response Theory (IRT) model based on nested dichotomous scoring of the K6 in all countries combined

| | Frequency of symptom (How much of the time . . .)[1] | | | |
|---|---|---|---|---|
| | All | Most | Some | A little |
| Nervous | 2.4* | 1.8* | 1.1* | 0.3* |
| Hopeless | 2.3* | 1.9* | 1.4* | 0.9* |
| Restless or fidgety | 2.4* | 1.9* | 1.2* | 0.5* |
| So depressed that nothing could cheer you up | 2.3* | 1.9* | 1.4* | 1.0* |
| Everything was an effort | 2.4* | 1.8* | 1.2* | 0.6* |
| Worthless | 2.3* | 1.9* | 1.5* | 1.1* |

*Significant at the 0.05 level, two-sided test. Standard errors of parameter estimates are not reported, as each rounds to 0.0.

[1]Coefficients are for dummy variables for the four response categories in comparison to the *none of the time* response category.

that are greater than 1.0 (van der Linden and Hambleton, 1997). The K6 slopes were in the range 1.1–3.0 with a median of 1.8 and an inter-quartile range (25th to 75th percentiles) of 1.6–2.7. This means that scales based on one-parameter and two-parameter IRT models are very highly correlated (over 0.9 in each country).

The general pattern of IRT severity parameters can be seen by inspecting these parameters based on the two-parameter IRT model for the worst month estimated in all the countries combined (Table 4). We see there that the severity estimates for responses of 'none of the time' (benchmarked at odds of 1.0 by construction) differ little from those for responses of 'a little of the time' (odds-ratios of 0.3–1.1 across items) or 'some of the time' (1.1–1.5), whereas the severity estimates for responses of 'most of the time' are considerably higher (1.8–1.9) and those for 'all of the time' are higher still (2.3–2.5). This means that a scoring scheme that gave especially high values to the highest two responses (e.g. 0,0,0,5,10) would do better than the unit-scoring scheme (i.e. 0,1,2,3,4) in maximizing inter-correlations among the six K6 items. It is not clear from this result, though, whether an alternative scoring scheme would be superior to the unit-weighting scheme in predicting SMI. We investigated this question next by comparing the strength of predictions based on versions of the K6 scale using unit weight versus IRT-based weighting.

## The regression analysis

A number of multiple regression equations were estimated to predict DSM-IV/CIDI diagnoses of SMI in each country. Although we originally considered separate models for unweighted (i.e. unit weighting) and weighted (i.e. IRT-based weighting) K6 scores, these models turned out to be virtually identical because the weighted and unweighted K6 scores were found to be very highly correlated to each other in all countries. These correlations were in the range 0.96–0.99 for past-month K6 for all countries. The correlations were all 0.99 for worst-month K6. In countries where both past-month and worst-month K6 were assessed, the models based on these two different recall periods were also very similar because of high correlations between past-month and worst-month K6 scores (0.78 in Japan, 0.87 in Brazil, and 0.92–0.97 in the other countries that assessed the K6 in both recall periods).

A number of the total of 93 multiple regression equations were estimated to predict DSM-IV/CIDI diagnoses of SMI for each recall period assessed in each country. (As a result of the high correlation between unweighted and weighted K6 scores, all results reported below refer to the unweighted version that used 0–24 scoring.) The first three models considered the K6 alone with either a linear, quadratic, or third-degree functional form (e.g. K6, K6-squared, and K6-cubed all in the same equation) in predicting SMI. A series of seven additional models for each of the three K6 functional forms then added socio-demographic controls either one at a time, two at a time, or all three at once. Eight more complex models were then estimated that included two-way interactions among the socio-demographic variables either one at a time, two at a time, or all three at once and then added the three-way interaction among all the socio-demographics. The remaining models then added interactions of the K6 (with and without its successive polynomials) with the socio-

demographics in each of the 15 lower-order socio-demographic models.

As noted above in the section on analysis methods, comparative model fit across these 93 equations was evaluated with the Akaike Information and Bayesian Information criteria. In cases where these two measures led to different conclusions, we selected the less complex of the two preferred models. This resulted in a variety of different models being selected across countries (Table 5). It is noteworthy that the quadratic term of the K6 was significant in about half the surveys. The coefficient associated with the K6 was always positive and the coefficient associated with the K6-squared was always negative in these cases, indicating that the log-odds of SMI increased at a decreasing rate as the values of the K6 increased. Socio-demographics were significant in 11 surveys and interactions among socio-demographics in eight surveys. These significant coefficients indicate that The K6 does not explain socio-demographic differences in SMI in most countries, a pattern that could be due either to a lack of reliability of the K6 or to systematic differences in the extent to which the K6 detects the kinds of disorders responsible for SMI in these different segments of the population.

Differential sensitivity would be indicated by interactions between K6 scores and the socio-demographics in predicting SMI. These interactions were significant in only five surveys (in Colombia, Lebanon, Nigeria, India, and the USA). Significant interactions were found with gender in four of these five countries, all involving stronger associations between K6 scores and SMI among women than men. Significant interactions were also found with age and education, but they were inconsistent in sign across countries. Inclusion of main effects and interactions involving these socio-demographic variables with K6 scores corrected for bias that would otherwise have occurred in estimating individual-level predicted probabilities of SMI. With these corrections in the models, the AUC values were generally substantial, with the median value of AUC across countries equal to 0.83, the

**Table 5** Significant model parameters of best-fitting regression models to predict DSM-IV/CIDI SMI from K6 scores[1]

| | K6 | $K6^2$ | $K6^3$ | SD | SD × SD | K6 × SD |
|---|---|---|---|---|---|---|
| Brazil | X | X | X | | | |
| Bulgaria | X | X | | X | X | |
| Colombia | X | X | | X | X | X |
| India | X | | | X | | X |
| Japan | X | | | | | |
| Lebanon | X | X | | X | X | X |
| Mexico | X | | | X | X | |
| New Zealand | X | X | | X | X | |
| Nigeria | X | | | X | | X |
| People's Republic of China | | | | | | |
| Beijing/Shanghai | X | | | X | X | |
| Shenzhen | X | | | | | |
| Romania | X | X | | X | | |
| South Africa | X | | | | | |
| Ukraine | X | X | | X | X | |
| USA | X | X | | X | X | X |
| Total | | | | | | |

[1] Significant parameters are indicated by an X. Only a general summary is presented here. See the appendix tables (available at http://www.hcp.med.harvard.edu/wmh/publications.php) for a description of exact parameters and parameter values. Due to the high correlations between past-month and worst-month K6 scores in countries that used both recall periods, the coefficients found to be significant were identical for past-month and worst-month models in these countries. The entries for K6, $K6^2$ and $K6^3$ represent the significance of the main effects of these three transformations of the K6 scale. The entry for SD represents the significance of one or more main effects of socio-demographic (SD) variables. The entry for SD × SD represents the significance of one or more interactions among the socio-demographic variables. The entry for K6 × SD represents the significance of one or more interactions between the K6 (or its polynomials) and one or more socio-demographic variables (or interactions among socio-demographic variables).

**Table 6** Area under the receiver operating characteristic curve (AUC) for the best-fitting model of the association between the K6 and DSM-IV/CIDI Serious Mental Illness (SMI)[1]

| | Past-month | Worst-month |
|---|---|---|
| Brazil | 0.83 | 0.82 |
| Bulgaria | 0.83 | 0.83 |
| Colombia | 0.81 | 0.83 |
| India | 0.85 | 0.85 |
| Japan | 0.79 | 0.86 |
| Lebanon | 0.86 | – |
| Mexico | – | 0.85 |
| New Zealand | 0.81 | 0.88 |
| Nigeria | 0.82 | – |
| People's Republic of China | | |
| Beijing/Shanghai | 0.85 | – |
| Shenzhen | 0.84 | 0.86 |
| Romania | 0.78 | 0.80 |
| South Africa | 0.76 | – |
| Ukraine | 0.83 | – |
| USA | – | 0.89 |

[1] Missing values are because some surveys used only one of the two recall periods.

range 0.76–0.89, and the inter-quartile range 0.81–0.85 (Table 6). These high AUC support the use of the equations to generate individual-level estimates of predicted probability of SMI in all countries in both time frames.

## Discussion

The results presented here show that relatively simple optimal scoring rules can be developed for the K6 in all the countries studied that generate predicted probabilities of SMI having substantial concordance with observed ratings of DSM-IV SMI based on the CIDI. It is striking that responses to a simple set of six questions, which take no more than 2 minutes to administer, can reproduce with such good accuracy diagnostic ratings based on a fully structured research diagnostic interview that takes an average of more than 1 hour to administer. This result argues that the K6 can be a valuable screening scale for SMI in general-purpose epidemiological surveys where it is not feasible to include a long assessment of mental disorders like the CIDI. Other validated screening scales exist to screen for specific mental disorders, such as the PHQ-9 screening scale of major depression (Kroenke *et al.*, 2001) and the ASRS screening scale of adult attention-deficit/hyperactivity disorder (Kessler *et al.*, 2005a). When the purpose of a particular research study is to investigate patterns and correlates of specific disorders, more specific screening scales such as these should be preferred to the K6. However, SMI is associated with a wide range of different mental disorders and it is important to have a more general screening scale for SMI such as the K6 when the researcher is more interested in screening broadly for SMI than for a particular kind of disorder.

It is important in this context to recognize that while the value of the K6 is as a broad screener rather than a specific screener for any one mental disorder, a limitation of the K6, as of the concept of SMI itself, is that the specific policy implications for treatment planning purposes of documented trends or correlates can be determined only by carrying out further analyses of component disorders. To some extent, of course, the same criticism can be made even of screening scales for more specific disorders, as it might be that some sub-types of specific disorders are more strongly related than others with correlates. However, as treatment approaches are much more similar within than between mental disorders, it remains true that policy implications of results regarding correlates of SMI are less clear than those regarding correlates of specific disorders.

Another limitation of the K6 is that, despite showing substantial concordance with an independent measure of SMI based on research diagnostic interviews, the number of items in the scale is so small that they might not span the full conceptual space that defines SMI in the population, leading to less sensitivity in detecting some types of SMI than others. This would not be a concern if AUC was perfect, but it is not. An AUC of 0.85, while very good, still means that 15% of true cases of SMI are not detected by the screening scale. If this under-detection is systematic (i.e. concentrated in a particular type of mental disorder in a specific segment of the population) rather than random, then even a dramatic increase in the component of SMI systematically missed by the scale will not be detected in trend surveys. Because of this possibility, it is important to carry out a second generation of methodological studies of the K6 now that its overall validity has been documented. These second-generation studies should search for evidence of systematic bias. We know from the analyses carried out here that biases with respect to age, gender, and education are minimal and that the scale has good properties across a wide range of countries, but we are aware of no comparable attempt to study bias with respect to other socio-demographic variables or with respect to specific types of mental disorders.

A final noteworthy limitation is that the optimal scaling rules developed here and reported in the appendix tables (available at http://www.hcp.med.harvard.edu/wmh/publications.php) are known to be optimal only with respect to the WMH surveys in which the rules were developed. This point was made in the Introduction, but needs to be reiterated here: that it is always preferable to base scaling rules whenever possible on clinical calibration studies embedded in the very same data collection that is used to administer the screening scale. This is the only way to guarantee that the sensitivity, specificity, and positive predictive value of the screening scale in the population under study are identical to those in the calibration sample. As described elsewhere is this issue (Colpe *et al.*, 2010), the US Substance Abuse and Mental Health Services Administration (SAMHSA) is addressing this problem by initiating an ongoing K6 calibration component in its annual National Household Survey on Drug Use and Health (NHSDUH). In this approach, a probability sub-sample of NHSDUH respondents that oversamples those with high K6 scores is administered a clinical re-interview to assess the presence of SMI. Clinical interviewers are blinded to the K6 responses in the main survey. This allows the SAMHSA investigators to calibrate K6 scores to predicted probabilities of SMI on an ongoing basis so as to protect against the possibility that the sensitivity and positive predictive value of the K6 decrease over time. Such decreases could be the result either of an increase in the relative importance of types of mental disorders that are not sensitively detected in the K6 in making up the total number of people with SMI or of secular changes in the words people use to describe their mental disorders. In any case, where the K6 is used as a screening scale in some other country in a large ongoing survey like the NHSDUH, it could be of considerable valuable to include an ongoing clinical calibration component of this sort. Not only will this allow the formula to estimate predicted probability of SMI from K6 scores to be modified over time to correct for secular changes in concordance, but the accumulation of more and more clinical cases over time will make it possible to refine calibration rules successively to capture subtle differences in concordance among respondents who differ in other characteristics assessed in the survey.

The question arises how to make best use of predicted-probability data in analysing transformed K6 scores. Several methods exist to do this. One is to treat the mean predicted probability as the variable of interest in linear or restricted linear (e.g. Tobit) prediction equations. Another possibility is to generate actual yes–no classifications of SMI for each respondent based on their predicted probabilities and then to analyse the data as one would with any other dichotomous diagnostic measure. There are two ways to do this that yield equivalent results. One is to use the predicted probabilities as weights. Under this approach, a respondent assigned a 0.25 predicted probability of SMI is treated as two people, one a person with a weight of 0.25 who has SMI and the other a person with a weight of 0.75 who does not have SMI. More generally, the observational record for each respondent is reproduced so that the sample is treated as having twice as many observations as it actually has, one for each respondent coded as having SMI and the other as not, with the pair of observations for each respondent having a sum of weights of 1.0 (or, in the case of otherwise weighted data, with the sum of weights equal to whatever the sample weight would otherwise have been for the respondent). The relative weighting of the two data records in each pair varies across respondents depending on the respondent's predicted probability of SMI. The MI estimation method described earlier in the paper can be used here by applying this weighting approach separately to each of the separate MI datasets and then pooling coefficients across pseudo-samples using standard MI methods (Rubin, 1987).

A more parsimonious alternative to the above approach that does not require the number of records to be doubled is to classify each respondent either as having or not having SMI by using a random number generator from a binomial distribution that is defined by the respondent's own predicted probability of SMI from the imputation equation. For example, if a random number was selected between 1 and 100 for a respondent with a predicted probability of SMI of 0.25, the respondent would be classified as having SMI if the random number was in the range 1–25 and as not having SMI if the random number was in the range 26–100. A separate random number would then be selected independently for each other respondent such that the respondent would be classified as having SMI if the random number was less than or equal to 100 times the respondent's predicted probability of SMI and classified otherwise as not having SMI. The dichotomous outcomes generated in this way can be analysed in the same way as any other dichotomous diagnostic variable. In addition, MI can again be implemented by repeating the entire process for each of the individual's SMI prevalence estimates (importantly, selecting a new random number for each respondent for the estimated probability of SMI based on each separate pseudo-sample, *not* using a single predicted probability for a given individual across all pseudo-samples) and using conventional MI methods to analyse and combine results across the separate pseudo-samples.

## Acknowledgments

The analysis is carried out in conjunction with the WHO WMH Survey Initiative. We thank the WMH staff for assistance with instrumentation, fieldwork, and data analysis. These activities were supported by the United States 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 R03-TW006481), the Pan American Health Organization, the Eli Lilly & Company Foundation, Ortho-McNeil Pharmaceutical, Inc., GlaxoSmithKline, Bristol-Myers Squibb, and Shire. A complete list of WMH publications can be found at http://www.hcp.med.harvard.edu/wmh/. The São Paulo Megacity Mental Health Survey is supported by the State of São Paulo Research Foundation (FAPESP) Thematic Project Grant 03/00204-3. The Bulgarian Epidemiological Study of common mental disorders EPIBUL is supported by the Ministry of Health and the National Center for Public Health Protection. The Chinese World Mental Health Survey Initiative is supported by the Pfizer Foundation. The Shenzhen Mental Health Survey is supported by the Shenzhen Bureau of Health and the Shenzhen Bureau of Science, Technology, and Information. The Colombian National Study of Mental Health (NSMH) is supported by the Ministry of Social Protection. The WMHI was funded by WHO (India) and helped by Dr. R Chandrasekaran, JIPMER. The World Mental Health Japan (WMHJ) Survey 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, Labour, and Welfare. The Lebanese National Mental Health Survey (LEBANON) is supported by the Lebanese Ministry of Public Health, the WHO (Lebanon), Fogarty International, Act for Lebanon, anonymous private donations to IDRAAC, Lebanon, and unrestricted grants from Janssen Cilag, Eli Lilly, GlaxoSmithKline, Roche, and Novartis. The Mexican National Comorbidity Survey (MNCS) is supported by The National Institute of Psychiatry Ramon de la Fuente (INPRFMDIES 4280) and by the National Council on Science and Technology (CONACyT-G30544- H), with supplemental support from the Pan American Health Organization (PAHO). Te Rau Hinengaro: The New Zealand Mental Health Survey (NZMHS) is supported by the New Zealand Ministry of Health, Alcohol Advisory Council, and the Health Research Council. The Nigerian Survey of Mental Health and Wellbeing (NSMHW) is supported by the WHO (Geneva), the WHO (Nigeria), and the Federal Ministry of Health, Abuja, Nigeria. The Romania WMH study projects 'Policies in Mental Health Area' and 'National Study regarding Mental Health and Services Use' were carried out by the National School of Public Health & Health Services Management (former National Institute for Research & Development in Health), with technical support of Metro Media Transilvania, the National Institute of Statistics-National Centre for Training in Statistics, SC. Cheyenne Services SRL, Statistics Netherlands and were funded by the Ministry of Public Health (former Ministry of Health) with supplemental support of Eli Lilly Romania SRL. The South Africa Stress and Health Study (SASH) is supported by the US National Institute of Mental Health (R01-MH059575) and National Institute of Drug Abuse with supplemental funding from the South African Department of Health and the University of Michigan. The Ukraine Comorbid Mental Disorders during Periods of Social Disruption (CMDPSD) study is funded by the US National Institute of Mental Health (RO1-MH61905). The US National Comorbidity Survey Replication (NCS-R) is supported by the National Institute of Mental Health (NIMH; U01-MH60220) with supplemental support from the National Institute of Drug Abuse (NIDA), the Substance Abuse and Mental Health Services Administration (SAMHSA), the Robert Wood Johnson Foundation (RWJF; Grant 044708), and the John W. Alden Trust.

## Declaration of interests

Dr. Kessler has been a consultant for GlaxoSmithKline Inc., Kaiser Permanente, Pfizer Inc., Sanofi-Aventis, Shire Pharmaceuticals, and Wyeth-Ayerst; has served on advisory boards for Eli Lilly & Company and Wyeth-Ayerst; and has had research support for his epidemiological studies from Bristol-Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Johnson & Johnson Pharmaceuticals, Ortho-McNeil Pharmaceuticals Inc., Pfizer Inc., and Sanofi-Aventis. The remaining authors report no conflicts of interest.

## References

Baggaley R.F., Ganaba R., Filippi V., Kere M., Marshall T., Sombie I., Storeng K.T., Patel V. (2007). Detecting depression after pregnancy: the validity of the K10 and K6 in Burkina Faso. *Trop Med Int Health*, **12**, 1225–1229, DOI: 10.1111/j.1365-3156.2007.01906.x

Baillie A.J. (2005). Predictive gender and education bias in Kessler's psychological distress Scale (k10). *Social Psychiatry Psychiatr Epidemiol*, **40**, 743–748, DOI: 10.1007/s00127-005-0935-9

Burnham K.B., Anderson D.R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociol Methods Res*, **33**, 261–304, DOI: 10.1177/0049124104268644

Colpe L.J., Barker P.G., Karg R.S., Batts K.R., Morton K.B., Gfroerer J.C., Stolzenberg S.J., Cunningham D.B., First M.B., Aldworth J. (2010). The National Survey on Drug Use and Health Mental Health Surveillance Study: calibration study design and field procedures. *Int J Methods Psychiatr Res*, **19**(Suppl. 1), 36–48.

Coyne J.C., Thompson R., Racioppo M.W. (2001). Validity and efficiency of screening for history of depression by self-report. *Psychol Assessment*, **13**, 163–170, DOI: 10.1037/I040-3590.13.2.163

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., Ustun T.B., Chatterji S. (2004). Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *J Am Med Assoc*, **291**, 2581–2590, DOI: 10.1001/jama.291.21.2581

Derogatis L.R. (1983). *SCL-90-R Revised Manual*, Johns Hopkins School of Medicine.

Dohrenwend B.P., Shrout P.E., Egri G., Mendelsohn F.S. (1980). Nonspecific psychological distress and other dimensions of psychopathology. Measures for use in the general population. *Arch Gen Psychiatry*, **37**, 1229–1236.

Embretson S.E., Reise S.P. (2000). *Item Response Theory for Psychologists*, Erlbaum.

Fassaert T., De Wit M.A., Tuinebreijer W.C., Wouters H., Verhoeff A.P., Beekman A.T., Dekker J. (2009). Psychometric properties of an interviewer-administered version of the Kessler Psychological Distress scale (K10) among Dutch, Moroccan and Turkish respondents. *Int J Methods Psychiatr Res*, **18**, 159–168, DOI: 10.1002/mpr.288

Fayyad J., De Graaf R., Kessler R., Alonso J., Angermeyer M., Demyttenaere K., De Girolamo G., Haro J.M., Karam E.G., Lara C., Lepine J.P., Ormel J., Posada-Villa J., Zaslavsky A.M., Jin R. (2007). Cross-national prevalence and correlates of adult attention-deficit hyperactivity disorder. *Br J Psychiatry*, **190**, 402–409, DOI: 10.1192/bjp.bp.106.034389

Furukawa T.A., Andrews G., Goldberg D.P. (2002). Stratum-specific likelihood ratios of the general health questionnaire in the community: help-seeking and physical co-morbidity affect the test characteristics. *Psychol Med*, **32**, 743–748, DOI: 10.1017/S0033291702005494

Furukawa T.A., Goldberg D.P., Rabe-Hesketh S., Ustun T.B. (2001). Stratum-specific likelihood ratios of two versions of the general health questionnaire. *Psychol Med*, **31**, 519–529, DOI: 10.1017/S0033291701003713

Furukawa T.A., Kawakami N., Saitoh M., Ono Y., Nakane Y., Nakamura Y., Tachimori H., Iwata N., Uda H., Nakane H., Watanabe M., Naganuma Y., Hata Y., Kobayashi M., Miyake Y., Takeshima T., Kikkawa T. (2008). The performance of the Japanese version of the K6 and K10 in the World Mental Health Survey Japan. *Int J Methods Psychiatr Res*, **17**, 152–158, DOI: 10.1002/mpr.257

Furukawa T.A., Kessler R.C., 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**, 357–362, DOI: 10.1017/S0033291702006700

Gill S.C., Butterworth P., Rodgers B., Mackinnon A. (2007). Validity of the mental health component scale of the 12-item Short-Form Health Survey (MCS-12) as measure of common mental disorders in the general population. *Psychiatry Res*, **152**, 63–71, DOI: 10.1016/j.psychres.2006.11.005

Gurin G., Veroff J., Feld S.C. (1960). *Americans View Their Mental Health*, Basic Books Inc.

Guyatt G., Rennie D. (2001). *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*, AMA Press.

Haller D.M., Sanci L.A., Sawyer S.M., Patton G.C. (2009). The identification of young people's emotional distress: a study in primary care. *Br J Gen Practice*, **59**, e61–e70.

Harkness J., Pennell B.E., Villar A., Gebler N., Aguilar-Gaxiola S., Bilgen I. (2008). Translation procedures and translation assessment in the World Mental Health Survey Initiative. In: *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders* (eds Kessler RC, Üstün TB), pp. 91–113, Cambridge University Press.

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. *Int J Methods Psychiatr Res*, **15**, 167–180, DOI: 10.1002/mpr.196

Heeringa S.G., Wells E.J., Hubbard F., Mneimneh Z.N., Chiu W.T., Sampson N.A., Berglund P.A. (2008). Sample designs and sampling procedures. In: *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders* (eds Kessler RC, Üstün TB), pp. 14–32, Cambridge University Press.

Hides L., Lubman D.I., Devlin H., Cotton S., Aitken C., Gibbie T., Hellard M. (2007). Reliability and validity of the Kessler 10 and Patient Health Questionnaire among injecting drug users. *Aust N Z J Psychiatry*, **41**, 166–168, DOI: 10.1080/00048670601109949

Huang Y., Kotov R., de Girolamo G., Preti A., Angermeyer M., Benjet C., Demyttenaere K., de Graaf R., Gureje O., Karam A.N., Lee S., Lepine J.P., Matschinger H., Posada-Villa J., Suliman S., Vilagut G., Kessler R.C. (2009). DSM-IV personality disorders in the WHO World Mental Health Surveys. *Br J Psychiatry*, **195**, 46–53, DOI: 10.1192/bjp.bp.108.058552

Kessler R.C., Adler L., Ames M., Demler O., Faraone S., Hiripi E., Howes M.J., Jin R., Secnik K., Spencer T., Ustun T.B., Walters E.E. (2005a). The World Health Organization Adult ADHD Self-Report Scale (ASRS): a short screening scale for use in the general population. *Psychol Med*, **35**, 245–256, DOI: 10.1017/S0033291704002892

Kessler R.C., Andrews G., Colpe L.J., Hiripi E., Mroczek D.K., Normand S.L., Walters E.E., Zaslavsky A.M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychol Med*, **32**, 959–976, DOI: 10.1017/S0033291702006074

Kessler R.C., Angermeyer M., Anthony J.C., R D.E.G., Demyttenaere K., Gasquet I., G D.E.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., Ustun 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**, 168–176.

Kessler R.C., Barker P.R., Colpe L.J., Epstein J.F., Gfroerer J.C., Hiripi E., Howes M.J., Normand S.L., Manderscheid R.W., Walters E.E., Zaslavsky A.M. (2003). Screening for serious mental illness in the general population. *Arch Gen Psychiatry*, **60**, 184–189.

Kessler R.C., Berglund P., Demler O., Jin R., Merikangas K.R., Walters E.E. (2005b). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry*, **62**, 593–602.

Kessler R.C., Berglund P.A., Zhao S., Leaf P.J., Kouzis A.C., Bruce M.L., Friedman R.M., Grosser R.C., Kennedy C., Kuehnel T.G., Laska E.M., Manderscheid R.W., Narrow W.E., Rosenheck R.A., Santoni T.W., Schneier M. (1996). The 12-month prevalence and correlates of Serious Mental Illness (SMI). In: *Mental Health, United States 1996* (eds Manderscheid RW, Sonnenschein MA), pp. 59–70, U.S. Government Printing Office.

Kessler R.C., Birnbaum H., Demler O., Falloon I.R., Gagnon E., Guyer M., Howes M.J., Kendler K.S., Shi L., Walters E., Wu E.Q. (2005c). The prevalence and correlates of nonaffective psychosis in the National Comorbidity Survey Replication (NCS-R). *Biol Psychiatry*, **58**, 668–676, DOI: 10.1016/j.biopsych.2005.04.034

Kessler R.C., Chiu W.T., Demler O., Merikangas K.R., Walters E.E. (2005d). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry*, **62**, 617–627.

Kessler R.C., Üstün T.B. (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: 10.1002/mpr.168

Kessler R.C., Üstün T.B. *et al.* (2008). *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*, Cambridge University Press.

Kessler R.C., Wittchen H.-U., Abelson J.M., McGonagle K., Schwarz N., Kendler K.S., Knäuper B., Zhao S. (1998). Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey. *Int J Methods Psychiatr Res*, **7**, 33–55, DOI: 10.1002/mpr.33

Kim K.O. (1993). *Factor Analysis: Statistical Methods and Practical Issues*, Sage.

Kroenke K., Spitzer R.L., Williams J.B. (2001). The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*, **16**, 606–613, DOI: 10.1046/j.1525-1497.2001.016009606.x

Landis J.R., Koch G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Langner T.S. (1962). A twenty-two item screening score of psychiatric symptoms indicating impairment. *J Health Hum Behav*, **3**, 269–276.

Leighton A.H. (1975). *My Name is Legion*. Vol. 1 of the Stirling County Study, Basic Books.

Link B.G., Dohrenwend B.P. (1980). Formulation of hypotheses about the true relevance of demoralization in the United States. In: *Mental Illness in the United States: Epidemiological Estimates* (eds Dohrenwend BP, Dohrenwend BS, Gould MS, Link B, Neugebauer R, Wunsch-Hitzig R), pp. 114–132, Praeger.

MacMillan A.M. (1957). The Health Opinion Survey: techniques for estimating prevalence of psychoneurotic and related types of disorder in communities. *Psychol Rep*, **3**, 325–339.

Margolis D.J., Bilker W., Boston R., Localio R., Berlin J.A. (2002). Statistical characteristics of area under the receiver operating characteristic curve for a simple prognostic model using traditional and bootstrapped approaches. *J Clin Epidemiol*, **55**, 518–524, DOI: 10.1016/S0895-4356(01)00512-1

Myers J.K., Lindenthal J.J., Peper M.P. (1975). Life events, social integration and psychiatric symptomatology. *J Health Soc Behav*, **16**, 421–427.

National Advisory Mental Health Council. (1993). Health care reform for Americans with severe mental illnesses. *Am J Psychiatry*, **150**, 1447–1465.

Patel V., Araya R., Chowdhary N., King M., Kirkwood B., Nayak S., Simon G., Weiss H.A. (2008). Detecting common mental disorders in primary care in India: a comparison of five screening questionnaires. *Psychol Med*, **38**, 221–228, DOI: 10.1017/S0033291707002334

Pearlin L.I., Lieberman M.A ., Menaghan E.G., Mullan J.T. (1981). The stress process. *J Health Soc Behav*, **22**, 337–356.

Pennell B.-E., Mneimneh Z., Bowers A., Chardoul S., Wells J.E., Viana M.C., Dinkelmann K., Gebler N., Florescu S., He Y., Huang Y., Tomov T., Vilagut G. (2008). Implementation of the World Mental Health Surveys. In: *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders* (eds Kessler RC, Üstün TB), Cambridge University Press.

Pepe M.S. (2003). *Statistical Analysis of Medical Tests for Classification and Prediction*, Oxford University Press.

Rao G. (2006). *Rational Medical Decision-Making: A Case-Based Approach*, McGraw-Hill.

Regier D., Narrow W., Rupp A., Rae D., Kaelber C. (2000). The epidemiology of mental disorder treatment need: community estimates of medical necessity. In: *Unmet Need in Psychiatry* (eds Andrews G, Henderson S), pp. 41–58, Cambridge University Press.

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. *Arch Gen Psychiatry*, **38**, 381–389.

Robins L.N., Regier D.A. (1991). *Psychiatric Disorders in America: The Epidemiologic Catchment Area Study*, The Free Press.

Robins L.N., Wing J., Wittchen H.U., Helzer J.E., Babor T.F., Burke J., Farmer A., Jablenski A., Pickens R., Regier D.A. (1988). The Composite International Diagnostic Interview. An epidemiologic Instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Arch Gen Psychiatry*, **45**, 1069–1077.

Rubin D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons.

Rush J., Carmody T., Reimitz P.-E. (2000). The Inventory of Depressive Symptomatology (IDS): Clinician (IDS-C) and Self-Report (IDS-SR) Ratings of Depressive Symptoms. *Int J Methods Psychiatr Res*, **9**, 45–59, DOI: 10.1002/mpr.79

SAS Institute Inc. (1999). *SAS User's Guide, release 8*, SAS Institute, Inc.

Scientific Software International I. (1996). *BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items*, Scientific Software International, Inc.

Seiler L.H. (1973). The 22-item scale used in field studies of mental illness: a question of method, a question of substance, and a question of theory. *J Health Soc Behav*, **14**, 252–264.

Sheehan D.V., Lecrubier Y., Sheehan K.H., Amorim P., Janavs J., Weiller E., Hergueta T., Baker R., Dunbar G.C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry*, **59**(Suppl 20), 22–33;quiz 34–57.

Spitzer R.L., Williams J.B., Kroenke K., Linzer M., deGruy F.V., 3rd, Hahn S.R., Brody D., Johnson J.G. (1994). Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. *J Am Med Assoc*, **272**, 1749–1756.

Srole L., Langner T.S., Michael S.T., Opler M.K., Rennie T.A. (1962). *Mental Health in the Metropolis: The Midtown Manhattan Study*, McGraw-Hill.

Swartz J.A., Lurigio A.J. (2006). Screening for serious mental illness in populations with co-occurring substance use disorders: Performance of the K6 scale. *J Substance Abuse Treatment*, **31**, 287–296, DOI: 10.1016/j.jsat.2006.04.009

Tesfaye M., Hanlon C., Wondimagegn D., Alem A. (2009). Detecting postnatal common mental disorders in Addis Ababa, Ethiopia: Validation of the Edinburgh Postnatal Depression Scale and Kessler Scales. *J Affect Disord* (published online), DOI: 10.1016/j.jad.2009.06.020

van der Linden W.J., Hambleton R.K. (1997). *Handbook of Modern Item Response Theory*, Springer-Verlag.

Ware J.E., Sherbourne C.D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care*, **30**, 473–483.