Methods of the World Mental Health Surveys

Authors: Kessler, R.C., Heeringa, S.G., Pennell, B.E, Mneimneh, Z., Chardoul. S.A., Zaslavsky, A.M.

WMH samples

The World Mental Health (WMH) Survey Initiative is a World Health Organization (WHO) initiative designed to help countries carry out and analyze epidemiological surveys of the burden of mental disorders in their populations (www.hcp.med.harvard.edu/wmh). This supplement provides a broad overview of the methods used in the surveys. A more detailed presentation can be found elsewhere.[1] We aim to obtain a representative sample of the household population in each country or region under study. This usually involves drawing a multi-stage clustered area probability sample of households in the population and then selecting one, or in some cases, two respondents from each sampled household using probability methods without replacement, and then carrying out face-to-face interviews in the homes of respondents. Unique opportunities available in individual countries are used to develop a sampling plan that meets WMH standards. Most WMH survey countries develop a similar sampling plan that features multi-stage area probability sampling, although several adopt an alternative plan, such as using a national registry or combining use of area probability methods and registry sampling to achieve the required probability sample of the designated target population. All these samples, though, are probability samples. No WMH survey uses a convenience sample, an interviewer-managed quota sample, or any other non-probability method of sample selection.

Probability sample surveys are designed to describe a target population.[2] These vary somewhat across WMH surveys based on several dimensions. One involves the age range. Although all WMH surveys focus on adults, the age that defines adulthood varies across countries (typically either 18 or 21). In addition, some countries decided to impose an upper age limit on the sample (usually 65). Other dimensions that define the survey population involve geographic scope limitations (typically excluding otherwise eligible people who live in remote areas of the country), language restrictions, citizenship requirements, and special populations such as persons living in military barracks and group quarters or persons who are institutionalized at the time of the survey (e.g., hospital patients, prison inmates). eTable 1 provides a summary of the survey populations for the 32 WMH surveys in 29 countries carried out up to now. The vast majority had a minimum age of 18 years. Most had an unrestricted upper age range, but a few had upper ages of 65 or 70. Turning to geographic scope, 24 of the 32 surveys defined it as the entire

country, and three others as all areas in the country other than very rural areas (Colombia, Mexico, and Peru). The five remaining surveys defined the geographic scope of the survey as specific regions (Murcia in Spain; five of the six regions in Nigeria, representing about 57% of the national population; and five metropolitan areas in Japan) or as one specific Metropolitan area (Sao Paulo in Brazil, Medellin in Colombia).

The WMH surveys share a set of common analysis objectives, primarily centered on the estimation of the population prevalence and correlates of mental disorders. Survey cost structures are highly variable from one country to another. Total funding for the surveys also varies widely. In many cases, funding restrictions limit not only the total size of the interviewed sample but also the scope of the survey populations or the use of costly sample design options. The individual WMH sample designs employ the full range of probability sampling techniques that survey statisticians can use to improve sample precision and reduce costs. Stratification is used to increase sample precision and control sample allocation. Multi-stage designs with modest clustering in the initial stages is as used to control travel time and expenses. The vast majority of the WMH surveys use a multi-stage area probability sampling method. The population registry approach is attractive when it is available because it avoids within-household selection and weighting, but this was not an option in most surveys. Most surveys added one or more intermediate stages of selecting electoral or postal districts before selecting eligible households and adults within households.

WMH Field procedures

Interviewer training: Although large-scale cross-national surveys have been undertaken for decades,[3] there is surprisingly little research on the practical aspects of training and supervising interviewers to achieve high-quality survey data. While cultural adaptation of survey methods is widely recognized as necessary to achieve equivalence in measurement across countries,[4-7] the literature contains few recommendations for how to achieve this equivalence across the many phases of a project's development. In the absence of standards of practice, many cross-national projects have accepted the research traditions of individual countries, which vary widely in methodological rigor. An approach at the other extreme is to implement a "one-size-fits-all" methodology, which naively imposes the same procedure and protocols across all countries and cultures, based on the assumption that good practice in one culture will invariably be good practice in other cultures.[8] The WMH Survey Initiative implements an approach between these two extremes by establishing guidelines that set minimum standards for each phase of project implementation but allows for country-specific adaptations.

A detailed description of the WMH data quality control standards and implementation is presented

elsewhere.[9] We only highlight some main points here. The WMH interview is a complex instrument. Successful implementation requires interviewers to be carefully trained. We consequently placed high importance on careful interviewer training and quality assurance monitoring. Before starting the interviewer recruitment and training process, each country sends at least two interviewer supervisors to a centralized "train-the-trainer" session presented by the WMH Data Collection Coordination Centre at the University of Michigan in the U.S. These sessions, which last an average of six days, are designed to prepare the interviewer supervisors to train and monitor interviewer performance as well as to manage data collection and data processing in their country. Through these sessions, research teams obtain all the information and materials necessary to train their own interviewing staff using consistent procedures. These trainers, who in almost all countries come from a pool of experienced interviewers, then train a team of supervisors to help in interviewer recruitment, training, and field quality control monitoring. Many countries are required to recruit and hire new interviewers for the WMH survey, while others use field staff from ongoing survey organizations. Careful centralized screening procedures is used in this hiring process.

Interviewer training is divided into two parts: general interviewing training (GIT) and training specific to the WMH interview. GIT is designed to introduce interviewers to the basic components of standardized questionnaire administration (e.g., question reading, appropriate techniques for probing and seeking clarification, providing feedback, accurate data recording). GIT sessions last two to three days in most countries. All interviewers are required to demonstrate competence with GIT concepts and procedures through a variety of tests before moving on to study-specific training. Study-specific training averages 30 hours across countries. The content is presented as a mix of lecture and round-robin practice sessions focused on general project background and importance, rules for obtaining informed consent, definitions of eligibility and respondent selection procedures, specifics of the precise interview procedures, and discussion of production requirements. Hands-on practice is stressed throughout. Trainers assess the skill level of each interviewer in small group exercises and often hold tailored "after-hours" special sessions to address areas where interviewers need additional assistance and practice. Most countries include in the training team a clinical consultant, typically a psychiatric social worker or a clinical psychologist, who provides interviewers with background information about the kinds of psychiatric symptoms they will encounter during production interviewing. This clinical contact person (CCP) is also a resource person for both interviewers and respondents during data collection to address the needs of respondents who might require a referral for follow-up and interviewers who might need to debrief after a particularly difficult interview. In most cases, the CCP is

available to interviewers 24 hours a day, seven days a week. Many countries develop protocols that allow interviewers to contact their CCP privately, without first going through a supervisor to provide interviewers a unique opportunity to speak freely about their own and their respondents' experiences.

Quality control monitoring: Interviewers must pass a certification test before being approved for production work. Interviewers who do not pass the certification test are either terminated from the project or receive additional retraining and another opportunity to obtain certification. Interviewer training often continues during the production phase of the project through periodic in-person seminars, telephone conference calls, and bulletins or newsletters. Special procedures are also developed to monitor interviewer performance during production. Systematic monitoring is critical to survey data quality assurance.[10,11] Consistent with best-practices guidelines for survey implementation,[12] four areas of performance are the main targets of quality assurance monitoring: detection and prevention of falsified information, compliance with the interviewing rules and guidelines set forth in the training manual, performance of non-interview tasks, and identification of interviewer-questionnaire interface problems. These areas were evaluated by supervisor re-interview of selected cases, supervisor verification of key survey elements through spot re-contact of respondents, direct observation of interviews, audio-recording, questionnaire review, analysis of performance and production measures, keystroke/trace file analysis (files that record keystrokes and movement of the interviewer through the computerized instrument), and mock interviews/tests of knowledge and practice. Interviewers are terminated from production interviewing if they are deemed unable to perform up to required standards. The recommended supervisor-to-interviewer ratio of one supervisor for every 8 to10 interviewers is used in surveys that use paper and pencil data collection, while lower ratios are used in countries with computer administration, based on the greater control over the data collection process afforded by computerized interviewing.[13,14]

The WMH interview

Overview: The WMH interviews administer the WHO Composite International Diagnostic Interview (CIDI) Version 3.0.[15] The CIDI is a fully structured research diagnostic interview designed for use by trained lay interviewers who do not have clinical experience. The version of the CIDI used in the initial WMH surveys generated diagnoses of mental disorders according to the criteria of both the ICD-10 and DSM-IV systems, but this has been updated in recent years also to include DSM-5 criteria, although only DSM-IV criteria are used in the current report. Consistent WHO translation, back-translation, and harmonization procedures are used to modify the

CIDI for use in each WMH survey.[16] The same interviewer training materials, training programs, and quality control monitoring procedures are used across all WMH surveys to guarantee cross-survey comparability of data.[9]

The two-part interview: In the vast majority of WMH surveys (the exceptions being Iraq, Israel, Romania, and South Africa, where all respondents were administered the full interview), the interview is divided into two parts, with the questions in Part I administered to all respondents and the questions in Part II administered to a probability sub-sample of respondents based on responses to the Part I questions. The reason for this is that the interview is long, and the interviewer sometimes must return to the respondent's household a second or third time to complete it. As most respondents do not have a history of mental disorder, we do not need all these non-cases to achieve maximum statistical power in comparing to cases. This means that we can realize considerable cost saving by terminating the interview for a probability sub-sample of non-cases as soon as we learn that they do not meet criteria for any mental disorder. This was accomplished with the two-part interview design. The interview begins with a series of basic descriptive warm-up questions and then evaluates lifetime presence of a wide range of core mental disorders. All (100%) of the respondents who meet criteria for any of these disorders are continued into Part II, which includes questions about a wide range of correlates of the core disorders and assesses mental disorders of secondary interest and those that take a great deal of time to assess. Most notable among the latter is lifetime post-traumatic stress disorder, which requires administration of a long question series about lifetime trauma exposure as well as questions about lifetime and current symptoms. In addition, a probability sub-sample of other Part I respondents (i.e., those who did not meet criteria for any core disorder) are also selected to complete Part II, while interviews with the remaining non-cases are ended after the completion of the Part I questions.

The Part II data are then weighted to adjust for the under-sampling of Part II cases. For example, if we included only a random one-fourth (.25) of all Part I non-cases in the Part II sample in a given country, each of those Part II non-cases would be assigned a weight of 4.0 (1.0/.25) to compensate for their under-sampling. This means that if the estimated prevalence of a given lifetime mental disorder was 10% in the Part I sample, it would still be 10% in the weighted Part II sample. In a similar way, estimates of the correlates of disorder would be expected to be unbiased in the Part II sample compared to the Part I sample. However, the precision of these estimates might be lower in the Part II sample because the denominator sample size on which estimates were based would be smaller. Four observations are relevant with regard to precision. First, the precision of estimates increases at a decreasing rate as the number of non-cases increases relative to cases, with precision generally not increasing meaningfully with

more than 4 controls per case.[17] The number of non-cases selected for the Part II sample is generally in the range 15% to 25%, yielding more non-cases than cases of every single disorder in every survey. Second, the Part II sample are weighted to adjust for any discrepancies that exists between the measured characteristics of the non-cases selected into Part II and those not selected into Part II. Third, the certainty selections into Part II include 100% of respondents not only with any lifetime core disorder, but also with any sub-threshold manifestation of the core disorders, increasing power to distinguish cases from near-cases. Fourth, probability of selection into Part II among other non-cases is made in proportion to the number of eligible respondents in the sample household in all surveys that used household sampling, thereby reducing the within-household probability of selection weight.

Translation: The CIDI was developed initially in English. One of the fundamental challenges in an undertaking such as the WMH Survey Initiative is to achieve both equivalence in meaning and consistency in measurement across surveys in multiple languages. Since the symptoms of mental disorders are described and interpreted differently in different cultures,[18,19] it is often necessary to use substantially different terms or questions in different countries to assess these symptoms. Another complexity is that the CIDI source language, English, has a larger lexicon (stock of vocabulary) than any other language. This can mean that distinctions made in English cannot be matched in one or more target languages. The opposite can also be true with respect to certain areas of lexical or grammatical distinctions, in which the source language may not specify enough detail necessary for translation into a given target language. WMH collaborators are given guidelines for translation and adaptation of the CIDI aimed at achieving both equivalence in meaning and consistency in measurement across surveys. These guidelines, which are discussed in detail by Harkness *et al.* (2008),[16] are modifications of longstanding WHO guidelines for translation, back-translations, and harmonization that were updated by the staff of the WHO Data Collection Coordination Centre.

Countries are instructed that the central aim of the translation process is to achieve target language versions of the English questionnaire that are conceptually equivalent in each of the countries/cultures rather than literally equivalent (i.e., word-for-word translations). It is emphasized that the translation should sound natural in each language (as far as that is possible in standardized instruments) and should perform in comparable fashion across the populations and languages. Independent assessors, who are experts on cross-national translation, have reviewed the CIDI translations and found that these aims are generally achieved, but that some translations are at times too close to the English questionnaire language formulation and structure to sound "natural." To achieve these aims, countries

are required to follow a six-step process, which includes: (1) forward translation; (2) expert panel review; (3) independent back-translation; (4) harmonization of vocabulary and formulation across different country versions of a shared language (if appropriate); (5) pretesting and cognitive interviewing; and (6) final revision, creation, and documentation of a final version of the translated questionnaire. A detailed description these six steps is presented elsewhere.[16]

Pretesting: Consistent with best-practices recommendations,[2,20-23] pretests of the instrument and procedures are carried out in each WMH survey prior to main study implementation. Pretesting is especially important in cross-national studies because of the challenges associated with working in many different languages and social contexts.[23] This should not be confused with the pretesting phase of translation described in the last subsection, as the latter focused only on translation, whereas the larger subsequent pretesting phase evaluates not only the instrument but all survey procedures. The pretests are designed to mirror the main study, but experienced interviewer supervisors do the pretest data collection rather than interviewers. Pretest interviews are evaluated by debriefing interviewers to identify potential problem areas with the instrument and survey procedures, checking the distributions of items for high rates of missing data and out-of-range values, and using behavior coding of audio-taped pretest interviews to pinpoint questions that are often misread or often elicited respondent requests for clarification.

Non-response surveys: WMH collaborators in all countries are encouraged to carry out systematic non-response surveys to evaluate and, to the extent possible, correct for the effects of systematic survey non-response. The basic design of the non-response survey is to select a stratified probability subsample of initial survey non-respondents who are approached one last time and asked to participate in a brief (typically 10 to 20 minutes) interview that provides the investigators with basic information about individuals who did not participate in the full survey. Respondents in these non-response surveys are typically offered a financial incentive to participate. The questions in the survey include a small number about basic socio-demographics (e.g., age, sex, education, marital status) and diagnostic stem questions for diagnoses of core mental and substance use disorders. Importantly, identical questions are asked in the main survey. Comparison of responses to these questions in the main sample and the non-respondent sample is used to make inferences about non-response bias, while weighting adjustments are used to adjust the main sample for these biases. Person-level analysis weights that incorporated factors for sample selection, non-response and calibration (i.e., weighting the sample distributions on one or more variables or

multivariate profiles to equal the distribution known to exist in the population, with the population targets typically coming from government Census data) are constructed for each WMH survey dataset. The case-specific analysis weights are used in computing estimates of descriptive statistics for the survey population and for estimating the descriptive statistics reported in this volume.

ACKNOWLEDGEMENTS

Portions of this eSupplement originally appeared in: (Kessler, R.C., Heeringa, S.G., Pennell, B.E, Zaslavsky, A.M. (2018). Methods of the World Mental Health Surveys. In E.J. Bromet, E.G. Karam, K.C. Koenen, D.J. Stein (Eds.), Trauma and Posttraumatic Stress Disorder: Global Perspectives from the WHO World Mental Health Surveys (pp. 13-42). New York: Cambridge University Press); Heeringa, S. G., Wells, J. E., Hubbard, F., *et al*. (2008). Sample designs and sampling procedures. In R. C. Kessler, & T. B. Üstün, eds., *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*. New York, NY: Cambridge University Press, pp. 14-32; Kessler, R. C., & Üstün, T. B. (2008). The World Health Organization Composite International Diagnostic Interview. In R. C. Kessler, & T. B. Üstün, eds., *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*. New York, NY: Cambridge University Press, pp. 58-90; Pennell, B.-E., Mneimneh, Z., Bowers, A., *et al*. (2008). Implementation of the World Mental Health surveys. In R. C. Kessler, & T. B. Üstün, eds., *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*. New York, NY: Cambridge University Press, pp. 33-57; Harkness, J., Pennell, B. E., Villar, A., *et al*. (2008). Translation procedures and translation assessment in the World Mental Health Survey Initiative. In R. C. Kessler, & T. B. Üstün, eds., *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*. New York, NY: Cambridge University Press, pp. 91-113; Haro, J. M., Arbabzadeh-Bouchez, S., Brugha, *et al.* (2008). Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health surveys. In R. C. Kessler, & T. B. Üstün, eds., *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*. New York, NY: Cambridge University Press, pp. 114-130; Kessler, R.C., Heeringa, S.G., Pennell, B.-E., *et al*. (In press). World Mental Health Survey Methods. In K.M. Scott, P. De Jonge, D.J. Stein, & R.C. Kessler., eds., *Mental Disorders Around the World: Facts and Figures from the World Mental Health Surveys*. New York, NY: Cambridge University Press.

References

1. Kessler RC, Heeringa SG, Pennell BE, Zaslavsky AM. Methods of the World Mental Health Surveys. In: Bromet EJ, Karam EG, Koenen KC, Stein DJ, eds. *Trauma and Posttraumatic Stress Disorder: Global Perspectives from the WHO World Mental Health Survey*. Cambridge University Press; 2018:13-42:chap Methods of the World Mental Health Surveys.

2. Groves RM, Jr. FJF, Couper MP, Lepkowski JM, Eleanor Singer RT. *Survey Methodology*. John Wiley & Sons; 2004.

3. Heath A, Fisher S, Smith S. The Globalization Of Public Opinion Research. *Annual Review of Political Science*. 2005;8(1):297-333. doi:10.1146/annurev.polisci.8.090203.103000

4. Bulmer M. Introduction:The Problem of Exporting Social Survey Research. *Am Behav Sci*. 1998;42(2):153-167. doi:10.1177/0002764298042002003

5. Jowell R. How Comparative Is Comparative Research? *Am Behav Sci*. 1998;42(2):168-177. doi:10.1177/0002764298042002004

6. Kuechler M. The Survey Method:An Indispensable Tool for Social Science Research Everywhere? *Am Behav Sci*. 1998;42(2):178-200. doi:10.1177/0002764298042002005

7. Lynn P. Developing quality standards for cross-national survey research: five approaches. *International Journal of Social Research Methodology*. 2003;6:323 - 336.

8. Harkness JA, Van de Vijver JR, Mohler PP. *Cross-Cultural Survey Methods*. John Wiley & Sons; 2002.

9. Pennell B-E, Mneimneh ZN, Bowers A, et al. Implementation of the World Mental Health Survey Initiative. 2008:pp 33-57.

10. Billiet J, Loosveldt G. Improvement of the quality of responses to factual survey questions by interviewer training. *Public Opin Q*. 1988;52:190-211. doi:10.1086/269094

11. Fowler FJ. *Standardized survey interviewing :minimizing interviewer-related error*. SAGE; 1990.

12. Biemer PP, Lyberg LE. Introduction to Survey Quality. *Technometrics*. 2003;45:277 - 277.

13. Lavrakas P. *Telephone Survey Methods*. 2 ed. 1993. Accessed 2023/02/16. https://methods.sagepub.com/book/telephone-survey-methods

14. Williams BJ. Suggestions for the application of advanced technology in Canadian collection operations. *Journal of official statistics*. 1986;2(4):555-60.

15. Kessler RC, Ustün TB. The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *Int J Methods Psychiatr Res*. 2004;13(2):93-121. doi:10.1002/mpr.168

16. Harkness J, Pennell B-E, Villar A, et al. Translation procedures and translation assessment in the World Mental Health Survey Initiative. 2008:

17. Schlesselman JJ. *Case control studies : design, conduct, analysis*. Oxford University Press; 1982.

18. Cheng AT. Case definition and culture: Are people all the same? *Br J Psychiatry*. Jul 2001;179:1-3. doi:10.1192/bjp.179.1.1

19. Prince R, Tcheng-Laroche F. Culture-bound syndromes and international disease classifications. *Cult Med Psychiatry*. Mar 1987;11(1):3-52. doi:10.1007/bf00055003

20. Converse JM. *Survey questions : handcrafting the standardized questionnaire*. Sage Publications; 1986.

21. Harkness J, Pennell B-E, Schoua-Glusberg A. Survey Questionnaire Translation and Assessment. *Methods for Testing and Evaluating Survey Questionnaires*. 2004:453-473.

22. Sheatsley PB. Chapter 6 - Questionnaire Construction and Item Writing. Elsevier Inc; 1983:195-230.

23. Smith TW. Developing and Evaluating Cross-National Survey Instruments. *Methods for Testing and Evaluating Survey Questionnaires*. 2004:431-452.