



WORLD HEALTH ORGANIZATION

**Meeting on Statistical Methods for Enhancing
the Cross-Population Comparability of Survey Results**

Cambridge, Massachusetts, USA, October 1st–2nd, 2001

D R A F T

**REPORT ON WHO MEETING OF EXPERTS
ON
STATISTICAL METHODS
FOR
ENHANCING THE CROSS-POPULATION COMPARABILITY
OF
SURVEY RESULTS**

Introduction: This report is a summary of major conclusions and recommendations of a meeting of experts on statistical methods for enhancing the cross-population comparability of survey data organized by WHO and held in Cambridge, Massachusetts, USA, on October 1st-2nd, 2001. In addition to four WHO staff, the participants included psychometricians, statisticians, and social scientists who had substantial experience with survey data analysis. A list of participants and their affiliations, as well as details of the agenda, can be found in the Annex.

Background: The WHO Multi-Country Household Survey Study uses self-report data for assessing non-fatal health in populations as well as for assessing the responsiveness of health systems. These self-report data take the form of ordered categorical (ordinal) responses. One key analytical issue is that these self-report ordinal responses are not comparable across populations primarily because of response category cut-point shifts. Conceptualizing the observed responses as resulting from a mapping between an underlying unobserved latent variable (e.g., ability on the underlying domain of mobility) and a set of categorical responses, cut-points are threshold levels on the latent variable that characterize the transition from one observed categorical response to the next. If cut-points differ systematically across populations, or even across socio-demographic groups within a population, then the observed ordinal responses are not cross-population comparable since they will not imply the same level on the underlying latent variable that we are trying to measure (Figure 1).

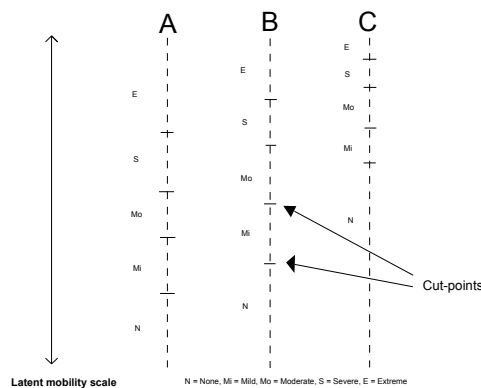


Figure 1. Mapping from unobserved latent variable to observed categorical response categories

Another way of characterizing this problem is that, for the same level of the latent variable on any given domain, the probability of an individual responding in any given response category is different across populations. The issue of cross-population comparability is not limited to health surveys: it is of equal relevance to self-report surveys on responsiveness of health systems, as well as to numerous other questions that rely on ordinal responses. In psychometric parlance, this is known as differential item functioning (DIF).

One example of self-report health data from the WHO Multi-Country Household Survey Study on Health and Responsiveness on the domain of mobility is: “Overall in the past 30 days, how much difficulty did you have with moving around?” Respondents are asked to classify themselves using one of five response categories: “1=Extreme/Cannot do; 2=Severe difficulty; 3=Moderate difficulty; 4=Mild difficulty; 5=No difficulty.”

Objectives and Agenda: There were two objectives for the meeting. The first was to obtain the opinions of a group of experts on the approach taken by WHO in enhancing the cross-population comparability of survey results. The second was to obtain advice and suggestions on future directions for this work.

The meeting began with an outline of standard statistical models used in the analysis of ordinal variables. The focus was on the ordered probit model (widely used by economists, political scientists, and other social scientists) and the partial credit model (widely used by psychometricians). A simulated data set was used where the observed response categories were generated from two hypothetical countries having different mean values of the latent variable as well as differences in response category cut-points. The country with the higher level of the latent variable was also assumed to have higher expectation for their health such that, in the end, the distributions of the observed categorical responses across three self-report questions did not look very different for the two countries. It was demonstrated that the use of standard techniques such as the ordered probit and the partial credit model – which do not allow for response

category cut-point differences in estimation – could lead to misleading inferences regarding the underlying latent variable when such differences were present in the data generating mechanism.

This was followed by a presentation of the work WHO has undertaken in terms of introducing methodological innovations to address the issue of cross-population comparability. These methodological innovations revolve around the use of vignettes to assess cut-point differences across socio-demographic groups. A vignette is a description of a concrete level of ability on a given domain that respondents are asked to evaluate in response to the main overall question for that domain and using the same categorical response scale. Vignettes are used to fix the level of ability such that any variations in responses are attributed to variations in response category cut-points.

More specifically, modifications to the ordered probit model and the partial credit model were introduced which utilize information from responses to vignettes to calibrate self-report responses to both the main and auxiliary questions (if any) so as to make estimates of the underlying latent variable cross-population comparable. These models, namely the hierarchical ordered probit (HOPIT) model and the hierarchical partial credit model (HPCM), modify the basic structure of the ordered probit model and the partial credit model so as to allow for cut-point shifts and “difficulty” parameter shifts, respectively, based on individual responses to vignettes.

This was followed by a presentation by Dr Jakob Bjorner from the National Institute of Occupational Health in Denmark. Dr Bjorner elaborated on models psychometricians use for analyzing ordinal data and their application to health surveys. He also talked about DIF and methods used to test for DIF in psychometric analysis. He highlighted the different sources of DIF, especially in the context of cross-language research. These included: (a) differences in translation/interpretation of response choices, (b) differences in translation/interpretation of items, (c) the fact that the relation between the items and the underlying construct may vary across populations, and (d) that the meaning of the underlying construct may vary. He elaborated on the basic structure of several models including: the basic Rasch model, the logistic item response theory model, the normal-ogive item response theory model, and several others. He discussed estimation of these models and presented a test of DIF using the conditional Rasch model. The basic idea behind the test of DIF is that, if the Rasch model were to be re-estimated in each sub-group, then the estimates should be the same as in the total population. A test can be constructed using the combined likelihood over sub-groups with the likelihood for the total population model.

Issues relating to unidimensionality as well as goodness-of-fit were also discussed. Results from using alternative calibration methods, such as the use of measured tests, were also presented. There was also a detailed discussion of the conditional estimation procedure implemented in the dichotomous Rasch model and the polytomous partial credit model. The meeting ended with a presentation of the results obtained from applying these methods to health and responsiveness data from the recent WHO Multi-Country Survey Study.

Main Conclusions and Recommendations: This section summarizes the discussions and conclusions for which there was general agreement.

- a) On the Need for Cross-Population Comparability:** There appeared to be general agreement for the need to correct survey results so as to make them cross-population comparable. The problem is especially pertinent since the WHO member states span a wide spectrum of levels in health status and socio-economic development.
- b) On Existing Methods for Cross-Population Comparability:** The problem is well-known and has been addressed before, but there appears not to be a satisfactory solution. For example, psychometric analysis is typically based on large “item” (i.e., question) banks, and one way to deal with DIF is to eliminate questions that exhibit DIF. It was acknowledged that this would not be applicable in the context of the WHO Survey Study given the small number of questions related to each of the domains in health and in responsiveness (typically ranging from one to five questions in each domain).
- c) On the Use of Vignettes:** There was general agreement that the methods proposed by WHO based on the use of vignettes were novel and interesting. Using the simulated data, these methods are demonstrably superior than standard statistical methods, such as the ordered probit and the partial credit models, in terms of recovering estimates of the underlying latent variable.

- d) On Goodness-of-Fit:** The problem of assessing goodness-of-fit for categorical response models was discussed. The methods used by WHO based on receiver operating characteristic (ROC) curves were presented. The ROC analysis was undertaken for all ordinal responses in one step. It was agreed that a better method would be to do the ROC analysis using success in predicting one categorical response at a time. It was also suggested that a detailed examination of existing methods for assessing goodness-of-fit was merited.
- e) On the Use of Response Categories:** One of the problems highlighted was the “stacking” of response categories in that most respondents in most WHO survey countries were answering “no difficulties” to the self-report questions, especially in the health domains. There was general agreement that perhaps the wordings of the questions be re-examined and adjusted so as to make better assessments of less-than-perfect states of health.

Other Discussion. This section summarizes suggestions made for developing the methods and analysis further.

- a) On Vignettes:** Suggestions were made on testing the assumption of fixed abilities for vignettes across countries. Both the HOPIT and the HPCM are premised on the assumption that vignettes are fixing ability on a given domain across countries, and that any differences are attributed to response category cut-point shifts. One way to test this assumption is to allow each vignette to vary in turn by country to assess the validity of this assumption using cross-country data. Another suggestion was to ask survey respondents to indicate which vignette they most resemble on a particular domain.
- b) On Latent Variable Differences:** The methods presented by WHO assume that both the vignettes and the self-report questions are based on the same underlying latent variable. It was suggested that this assumption could be relaxed by allowing a multi-dimensional latent variable formulation of the model that allowed for different latent variables for vignettes and for self-reports, but at the same time allowed for some degree of correlation between the two measures.

ANNEXAGENDA FOR WHO MEETING OF EXPERTS HELD IN CAMBRIDGE
OCTOBER 1st-2nd, 2001

Day 1

- 9:00-10:30 **Chris Murray, Ajay Tandon, Joshua Salomon, Bedirhan Ustun**
World Health Organization
Introduction, problem of cross-population comparability and Differential Item Functioning (DIF), and WHO's strategies for dealing with the problem: vignettes, HOPIT/CHOPIT, measured tests
- 10:30-11:00 Coffee Break
- 11:00-12:00 **Chris Murray, Ajay Tandon, Joshua Salomon, Bedirhan Ustun**
World Health Organization
Applications, "*in-silica*" experiment, Examples on domains of health from WHO survey programme.
- 12:00-12:30 **General Discussion and Reactions**
- 12:30-2:00 Lunch Break
- 2:00-3:30 **Jakob Bjorner**
National Institute of Occupational Health, Denmark
IRT and DIF presentation followed by discussion
- 3:30-4:00 Coffee Break
- 4:00-5:00 **General Discussion and Reactions**

Day 2

- 9:00-10:30 **Other Topics**
Unidimensionality of domains, Model fit, Applications to responsiveness
- 10:30-11:00 Coffee Break
- 11:00-12:30 **Closing Remarks and Discussion**

LIST OF PARTICIPANTS

1. Dr Betty Bergstrom
Vice President
Program Management and Psychometric Services
Computer Adaptive Technologies, Inc.
Evanston, IL
USA
2. Dr Juergen Rehm
Professor, Addiction Research Institute
University of Zurich
Zurich
SWITZERLAND
3. Dr Eugene Laska
Nathan S. Kline Institute for Psychiatric Research
Statistical Sciences and Epidemiological Division
Orangeburg, NY
USA

4. Dr Cees A. W. Glas
Faculty of Educational Science and Technology
Department of Educational Measurement and Data Analysis
University of Twente
Enschede
THE NETHERLANDS
5. Dr Gary King
Professor, Department of Government
Harvard University
Cambridge, MA
USA
6. Dr Jakob Bjorner
Senior Researcher
National Institute of Occupational Health
Copenhagen
DENMARK
7. Dr Rafael Di Tella
Harvard Business School
Boston, MA
USA
8. Dr Larry Ludlow
Educational Research, Measurement, and
Evaluation, Chair
Boston College
Lynch School of Education
Chestnut Hill, MA
USA
9. Dr Paul Gertler
Haas School of Business
University of California
Berkeley, CA
USA
10. Dr Jin Pihuan
Department of Health Statistics
Shanghai Medical University
Shanghai
CHINA
11. Dr Leo S. Morales
UCLA Assitant Professor
Division of General Internal Medicine and Health Services Research
School of Medicine
Los Angeles, CA
USA

WHO SECRETARIAT

1. Dr C.J.L. Murray
Executive Director, Evidence and Information for Policy Cluster
2. Dr Bedirhan Ustun
Co-ordinator, Classification, Assessment, Surveys, and Terminology Team

3. Dr Joshua Salomon
4. Dr Ajay Tandon