Comparability in Population Based Surveys
Concepts, Methods and Analysis

The purpose of this paper is to:

1) **Examine the problem of cross-population comparability in population based health surveys and review a series of empirical examples**
2) **Introduce a conceptual framework for understanding the comparability problem mathematically in terms of response category cut point shifts across populations, across subgroups within a population, or within the same population over time**
3) **Outline and describe a strategy for enhancing the cross-population comparability using both new measurement instruments that incorporate vignettes and novel statistical methods**
4) **Use World Health Survey (WHS) data for the Mobility domain in 71 countries to examine factors that predict cut-point shifts, to evaluate anchoring vignette methodology, and to compare adjusted "comparable" results with the unadjusted results.**

Data collected through household surveys have several advantages such as population representativeness and direct information from individuals on their own experiences, rather than from professionals or surveillance. A major area of research in the survey field has focused on the fact that since questions, response scales and samples differ greatly across surveys and across time, there is reduced comparability within and across surveys particularly in an international context. In recognition of these differences, researchers have come to the conclusion that standard data collection and analysis procedures should be used to overcome these avoidable methodological differences. However recent research has shown that even when identical or equivalent items have been used, the results across individuals, groups or populations may not be comparable (1). Survey results may be reliable and valid within each population but the results cannot be compared across populations without adjustment. This arises from the finding that in survey items that elicit responses to health questions on a categorical scale, if the meaning of response categories differs systematically across populations, or even across socio-demographic groups within a population, unrelated to health status, then the observed ordinal responses are not cross-population comparable since they will not imply the same level (*2*).

Reviewing the problem of cross-population comparability using a series of empirical examples

National health surveys rely heavily on self-reported health measures, but interpretation of these measures is complicated by comparability problems that arise when different persons understand and respond to a given question in different ways. These data are essential for the purposes of measurement, monitoring, and evaluation of health systems performance and are critical components of the evidence base for clinical practice and

health policy. In the broadest sense, comparability is required not only across countries, but also within countries over time, or across different subpopulations delineated by age, sex, education, income or other characteristics. We will use reported findings from the literature on 4 national health surveys as well as a critical review and re-analysis of 64 datasets, covering self-reported health status from population-based surveys in 46 countries, to illustrate the non-comparability of survey data across populations and show that self-reported health measures may give misleading results in the absence of adjustments.

In Australian national health surveys comparing the self-reported health status of Aboriginals with that of the general population, only around 12% of the Aboriginal population characterized their own health status as "fair" or "poor", while more than 20% of the general population rated their health in these low categories. By any other major indicator of mortality and morbidity, the Aboriginal population fares much worse than the general population (*3*). Residents of the state of Kerala in India, which has the lowest rates of infant and child mortality and the highest rates of literacy in the country, consistently report the highest incidences of morbidity in the country (*4*). A series of studies from the Living Standards Measurement Surveys has examined the gradient of reported illness as a function of income and found that individuals in higher income quintiles consistently report more illness than those with lower income levels (*51*). The 1994 European Community Household Panel (ECHP) collected data on the self-reported question "How is your health in general?" on a five point Likert response scale "very good," "good," "good," "fair," "bad," "very bad" in 12 countries of the European Union, based on the same survey and methods in all countries. The fraction of respondents reporting "very bad" or "bad" health varies from a high of 19% of the Portuguese to as little as 5% of the Irish population (*6*). Such divergent levels of health are implausible, given other major health indicators, even when differences in the underlying true level of health status, language, or measurement error are taken into account.

A conceptual framework for understanding the comparability problem in terms of differences in response category cut-points

The mainstay of health status measurement, regardless of the instrument used, is self-reported responses on health status in survey interviews. Because of issues of cost and feasibility, even if self-reported data are supplemented by measured tests, self-responses will likely remain one of the major data collection methods for population health status assessment. These self-response data typically take the form of ordered categorical (ordinal) responses, such as "excellent"/ "very good"/ "good"/ "poor"/ "bad" or "none"/ "mild"/ "moderate"/ "severe"/ "extreme." One critical issue that has been debated extensively is the degree to which self-responses on these items are comparable across individuals, socioeconomic subgroups or populations. The challenge of comparability is central to the future of health status instrument.

If health is understood as consisting of multiple dimensions (for example, mobility, cognition, vision, and affect, among others), then we may conceptualize the level on any

given dimension as a continuous but latent (unobserved) scale value, with higher values corresponding to better health levels. Each of the available response choices for a categorical self-report question corresponds to a certain range of values on the latent scale, which may differ across individuals. Thus, the influence of varying expectations for health may be expressed in terms of individual differences in the levels on a given dimension at which a person transitions from using one response category to the next. These response category boundaries are labeled ***cut-points*** and the problem of comparability is conceptualized in terms of response category cut-point shifts across populations, across subgroups within a population, or within the same population over time.

Different populations and subgroups may attach different meanings to the response categories in a survey. The issue can be conceptualized in terms of response category cut-point shifts across populations and subgroups. This is highly problematic when such data are used as a basis for comparison across countries, populations, or subgroups.

Strategy for enhancing the cross-population comparability of self-reported health data using Vignettes & the Categorical Hierarchical Ordered Probit (CHOPIT) model

The challenge then, is to correct responses so as to make the results useful for analyses where real differences matter. An instrument should have a common metric in different populations; the same response should correspond to the same level of health in the domain measured. We propose to do his through the use of new survey instruments that incorporate anchoring vignettes and a new statistical model for data analysis.

One strategy for establishing cross-population comparability is to fix the level of health on a domain and assess variation in the response categories across individuals, groups, and populations. For example, if the level of mobility is fixed but one group says that maps to a response category of "no difficulty" and another says it maps to the category "some difficulty", that information can be used to assess response category cut-points. Anchoring vignettes have been developed as a new component of survey instruments that may be used to position self-reported responses on a common, interpersonally comparable scale(7). An anchoring vignette is a description of a concrete level on a given health domain that respondents are asked to evaluate with the same questions and response scales applied to self-assessments on that domain. Vignettes fix the level of ability on a domain so that variation in categorical responses is attributable to variation in response category cut-points. The key objective underlying the anchoring vignette strategy is to elicit responses from subjects for hypothetical levels on a given domain, which reflect individual norms and expectations for health in approximately the same way the self-ratings do for the subjects' own health levels.

Standard statistical models for ordinal data, such as the ordered probit model, cannot allow for variation in response category cut-points. We propose to use an adaptation of standard statistical models for ordinal data, called the categorical hierarchical ordered probit (CHOPIT) (2) model, which uses vignette data to incorporate systematic

differences in response category use, to generate comparable scores in each domain. In order to incorporate information on vignette ratings, the expanded model has two components to the likelihood function: the first component refers to estimation of cut-points in relation to some defined set of covariates using responses to vignettes. The second component utilizes responses on the self-report questions in each domain. This formulation is slightly different from the standard ordered probit model; since we are allowing the vignettes to drive the cut-point estimation, this second component of the likelihood function has more in common with an interval regression model (i.e. an ordered probit model with known cutpoints).

In formal terms, the first component of the likelihood function assumes there is an unobserved latent variable Y_{ij}^{v*} distributed with mean u_{ij}^{v*} and variance $\sigma^{2\,v*}$. Here, *i* refers to the respondent, *j* refers to the vignette number, and the *v* superscript indicates that this refers to the vignette component of the model. t_i^k refers to the k^{th} cutpoint of respondent *i*.

$$y_i^{*v} = f(\text{dummy variables for vignettes})$$

$$y_{ij}^{v} = 1 \text{ if } -\infty < y^{*v} \leq t_i^1$$
$$y_{ij}^{v} = 2 \text{ if } t_i^1 < y^{*v} \leq t_i^2$$
$$y_{ij}^{v} = 3 \text{ if } t_i^2 < y^{*v} \leq t_i^3$$
$$y_{ij}^{v} = 4 \text{ if } t_i^3 < y^{*v} \leq t_i^4$$
$$y_{ij}^{v} = 5 \text{ if } t_i^4 < y^{*v} \leq +\infty$$

Where $t_i^k = f(\text{covariates})$

The second component of the likelihood function assumes there is an unobserved latent variable Y_{ij}^{s*} distributed with mean u_{ij}^{s*} and variance $\sigma^{2\,s*}$. Here, *i* refers to the respondent, *j* refers to the vignette number, and the *s* superscript indicates that this refers to the self-report component of the model. Since vignettes are driving the cut-point estimation and the scale is set by the first estimation component, we are now able to obtain estimates of the variance of the latent variable

$$y_i^{*s} = f(\text{covariates})$$

$$y_i^{s} = 1 \text{ if } -\infty < y^{*s} \leq t_i^1$$
$$y_i^{s} = 2 \text{ if } t_i^1 < y^{*s} \leq t_i^2$$
$$y_i^{s} = 3 \text{ if } t_i^2 < y^{*s} \leq t_i^3$$
$$y_i^{s} = 4 \text{ if } t_i^3 < y^{*s} \leq t_i^4$$
$$y_i^{s} = 5 \text{ if } t_i^4 < y^{*s} \leq +\infty$$

Where $t_i^k = f(\text{covariates})$

For multiple self-report questions, each tied to vignettes, the observation mechanism for each self-report question is the same, with the same y_i^{*s} and differing cutpoints. A key feature of the model is that each question has its own cutpoints and the cutpoints for any question for vignettes and for self-report is the same. There is explicit parametric dependence between the different components of the likelihood function. The two components of the likelihood function are additive in logs and can be jointly maximized to yield the parameter estimates.

Analysis & Results

The World Health Survey (WHS) was initiated to strengthen national capacity to monitor critical health outputs and outcomes and to improve the methodological and empirical basis for the measurement of population health through the fielding a standardized instrument together with new statistical methods for adjusting self-reported health measures to comparable scales. The 2002 WHS was conducted in 71 countries on 6 continents, using a probabilistic, nationally representative sampling strategy, and consisted of 54 household long, 13 household short and 4 CATI surveys. Sample sizes ranged from 700 – 38,746 respondents, the average being around 4022 respondents
In the health module of the WHS, self-assessed health levels and anchoring vignette levels are elicited for each of eight domains of health —mobility, self- care, pain and discomfort, cognition, interpersonal activities, vision, sleep and energy, and affect. For each domain, two items are included to reduce measurement error and improve the efficiency of statistical models used to analyze these data.

We use data from the Mobility domain to:

1) Evaluate the implementation of vignette methodology
2) Determine cutpoints by covariate subgroups and to assess the pattern of cutpoints between countries and age, sex, and education subgroups
3) Generate vignette adjusted scores in all countries and compare these to the unadjusted scores

We have defined two key requirements for the use of anchoring vignettes: *response consistency*, which states that an individual will use the response categories for a particular question in a similar way when evaluating hypothetical scenarios as when providing a self-assessment; and *vignette equivalence*, which states that the underlying domain levels represented in each vignette are understood in approximately the same way by all respondents, irrespective of their age, sex, income, education, country of residence, or other characteristics.

To assess the success to vignette implementation in the WHS we examine vignette rating patterns for one question in each country for the mobility domain. Deviations from global vignette ordering are assessed for each respondent by determining the rank order correlation coefficient (ROCC) of their ranks resulting from their categorical vignette ratings. The classic Spearman's coefficient deals with rank ties by assigning partial

scores. Since this unfairly penalizes those with cutpoints shifted to either extreme, we devised two variations of the classical correlation measure called the Benefit-of-Doubt-Rank Order Correlation Coefficient (BODROCC), which treats all ties as being perfectly ordered and the Victim-of-Doubt-Rank Order Correlation Coefficient (VODROCC), which treats all ties as being order-inversed. For both measures, average correlations and proportion of respondents below a correlation of 0.8 are compared across countries. A regression analysis where the dependent variable is the correlation and predictor variables are country, mode, domain, age, sex, education is conducted to examine factors that relate to vignette quality in household surveys. Secondly, a ratings analysis is performed for two indicators that measure problems in vignette responses: proportion of respondents rating the 'worst' vignette better than the 'best' vignette and proportion of respondents rating all vignettes the same. The former points to problems in implementation, translation while the latter points to respondent not understanding the task.

We run CHOPIT analysis, outlined above, on the mobility domain in each country. We allow cutpoints to be a function of country, age, sex, and education. The latent variable on mobility is allowed to be a function of the same covariates and also includes interaction terms between country and age and country and education. First we look at cutpoint shifts by covariate for each cutpoint. We hypothesize that cutpoints will shift down with age, meaning lower levels of health result in higher categorical ratings and that cutpoints will shift up with education. Patterns by country may yield interesting data regarding cultural 'optimism' and 'pessimism'. Finally, we look at latent-variable mobility scores by age, sex and country and compare these vignette-adjusted results to the raw unadjusted results. Figures one and two provide examples of the type of comparative outputs that will be presented. Fig 1. Illustrates unadjusted raw and post-CHOPIT results by age for the mobility domain in China. Fig 2. shows adjusted and unadjusted mean scores by country in the 2001 Multi Country Survey Study (MCSS).

Fig1. Mobility Mean Self-Report vs. Mean Posterior by Age in China

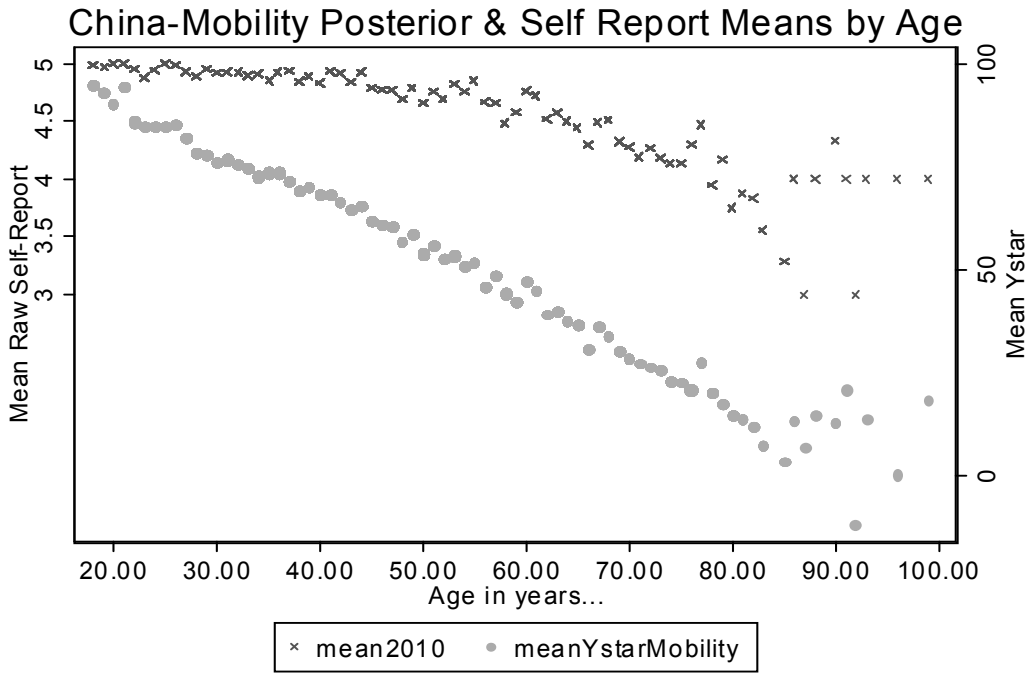
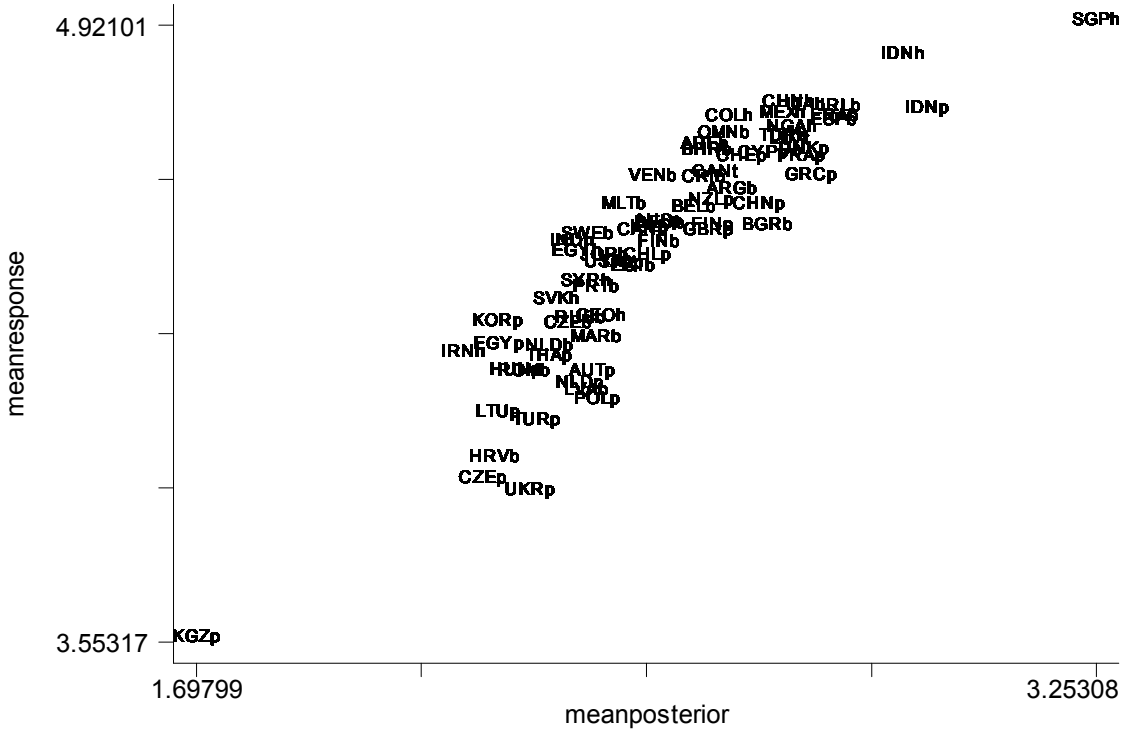


Fig2. All Country Mean Self-Report vs. Mean Posterior from 2001 MCSS Data



This paper provides a comprehensive look at the problem of comparability of data in national health surveys and provides a framework for correction. The application of the outlined methodology in 71 countries from the 2002 WHS is analyzed.

REFERENCES

(1) Sadana R et al. Comparative analyses of more than 50 household surveys on health status. In: Murray CJL et al., eds. *Summary measures of population health: concepts, ethics, measurement and applications.* Geneva, World Health Organization, 2002:369–386.

(2) Tandon A et al. Statistical models for enhancing crosspopulation comparability. In: Murray CJL, Evans DB, eds. *Health systems performance assessment: debates, methods and empiricism*. Geneva, World Health Organization, 2003.

(3) Mathers CD, Douglas RM. Measuring progress in population health and well-being. In: Eckersley R, ed. *Measuring progress: is life getting better?* Collingwood, CSIRO Publishing, 1998:125–155.

(4) Murray CJL. Epidemiology and morbidity transitions in India. In: Dasgupta M, Chen LC, Krishnan TN, eds. *Health, poverty and development in India*. Delhi, Oxford University Press, 1996:122–147.

(5) Murray CJL, Chen LC. Understanding morbidity change. *Population and development Review*, 1992, 18(3):481–503.

(6) Eurostat. Self-reported health in the European Community. *Statistics in focus, population and social conditions*.ISSN 1024–4352. Eurostat, 1997.

(7) King G et. al. Enhancing the validity of cross-cultural comparability of measurement in survey research. *American Political Science Review*. February 1994.