

XIII. CROSS-POPULATION COMPARABILITY

1. WHR 2000

In making the estimates for WHR 2000, corrections were made for major known biases in available measurements to improve cross-population comparability – for example, for under-reporting of mortality data in vital registration systems. The concept of internal consistency was used as a tool to improve the validity of epidemiological assessments.

2. Main commentaries and criticisms

Data criticisms of WHR 2000 were rather severe but this section deals only with the question of cross-country comparability. There has been little public debate and discussion on this issue beyond recognizing it as a problem with self-report data.

3. WHO responses and proposals

In examining self-assessed morbidity from survey data across the states of India, Murray and Chen (1992) reported the following findings: Kerala has the highest self-reported morbidity, and Bihar the lowest, across the Indian states. On the other hand, an objective measure of health – such as mortality – reveals that Kerala has a much higher life expectancy than Bihar. Next, a comparison between the US and Kerala shows that self-assessed morbidity in the US is much greater than in Kerala, despite life expectancy in the US being higher than that in Kerala.

What is going on? Are there features of the environment – educational, medical (e.g., frequency of exposure to the health system), income, etc. – that can explain these apparently inconsistent findings? Amartya Sen (1992) in an article in *Philosophy and Public Affairs* tried to understand these results in terms of what he called 'positional objectivity': the 'position' of the individual (in terms of education, income, etc.) matters in the response that is given – but all individuals in the same position will give the same response – hence 'positional objectivity'. In a more recent editorial in the *British Medical Journal*, he again emphasizes the fact that self-reported morbidity data have limitations that can make its use extremely misleading for policy purposes (Sen 2002).

WHO is seeking to make the responses of individuals comparable (whether they live in different states of India or in the US) by *correcting* for the 'positions' of the individuals in the different states of India and the US. This is obviously a very important exercise in obtaining health-status information from survey data that is comparable across countries. Moreover, self-reported data on health are still by far the most common source of such information around the world.

As a response to the paucity of representative population-based information on two key variables in the HSPA exercise in WHR 2000, the WHO launched the Multi-Country Survey Study on Health and Responsiveness. For the purposes of HSPA, these survey data are utilized to construct measures of: (i) health-adjusted life expectancy (HALE), and (ii) the level of responsiveness of the health system in a country. For example, the measurement of HALE includes estimates of non-fatal health that are, in part, derived from survey data on the different domains of health (e.g., mobility, cognition, affect, etc.). Similarly, the level of responsiveness of a country's health system is also based on such survey data. Respondents are asked to evaluate their experiences relating to different domains of responsiveness of the health system (e.g., autonomy, dignity, prompt attention, etc.).

There are two characteristics of these survey data that lead to the problem of cross-population comparability. First, the information on the domains is obtained on the basis of self-reporting. Respondents are asked to evaluate their own experience (or perception) with respect to various domains of health and of health-system responsiveness. Secondly, these self-report responses are categorical and ranked ordinally.

One example from the WHO Multi-Country Survey Study for the health domain of mobility illustrates the characteristics of the data. The main self-report question asks respondents how much difficulty they have had in moving around in the past 30 days. Respondents are asked to characterize their mobility using a 5-category ordinal response scale ranging from 1 to 5, where 1 is "Extreme/Cannot do", 2 is "Severe difficulty", 3 is "Moderate difficulty", 4 is "Mild difficulty", and 5 is "No difficulty".

This is where the issue of cross-population comparability arises. The problem with using these self-report data from the domains of health and responsiveness is that the responses are not comparable across countries, or even across different socio-demographic groups within countries. As Figure 1 illustrates, the categorical responses can be conceptualized as a mapping from the true level of the domain (here the line labelled "latent mobility scale") to the categorical responses for three different populations A, B, and C. As the figure shows, someone answering "No difficulty" in population A maps to a different interval on the true scale as someone answering "No difficulty" in populations B and C. Obversely, the same level of true mobility could be self-reported by a person in population A as representing "no difficulty", by a person in population B as representing "mild" difficulty, and

by a person in population C as representing “moderate” difficulty. The reasons could be due to differing norms, expectations, and experiences of respondents from different populations.

This problem has been previously identified in the psychometrics literature on ability (IQ) testing and, more generally, in educational testing through standardized tests (e.g., GRE, SAT, GMAT, etc.). Certain groups, conditional on ability or knowledge, systematically do better on certain types of question than other groups. This problem is known as “differential item functioning” in the educational testing and psychometrics literature (Holland and Wainer 1993). For instance, in the item response theory literature, the partial credit model (which is akin to the ordered probit model) specifies the probability of responding in one of two ordered (adjacent) categories as an increasing function of a respondent’s ability and a decreasing functioning of the category difficulty. For the same level of ability, the difficulties may be systematically different for different population groups, which will lead to a bias in measured ability. Although this problem is similar to the problem of cut-point shifts in measuring health or health-system responsiveness, the solution methods are somewhat different.

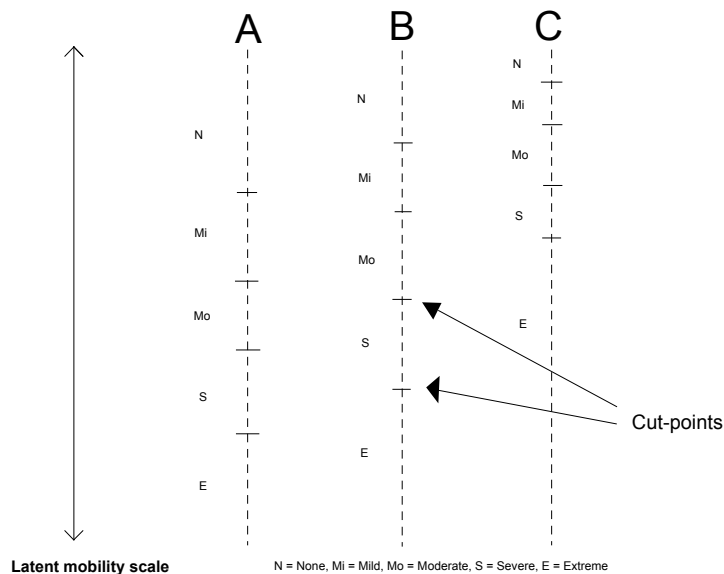


Figure 1

There are basically two strategies that WHO has developed to adjust survey responses for systematic differences in people’s attitudes. Both strategies involve the use of a statistical model – the hierarchical ordered probit (HOPIT) model. The first strategy is to use the HOPIT model with ‘vignettes’. The second strategy is to use the HOPIT model with measured tests. These are described in turn.

A vignette is a description of a level of ability on a given domain that respondents are asked to evaluate with respect to the same question and the on the same categorical response scale as the main self-report question. A vignette depicts a fixed level of ability on a given domain, so that for that vignette, differences in responses across countries or socio-demographic groups may be attributed to differences in cut-points for the response categories. The response category cut-points are estimated by use of the HOPIT model through a maximum likelihood procedure. These cut-point estimates are used to calibrate the respondent's own self-report in order to make it cross-population comparable. If, for example, respondents from a certain population group systematically give higher categorical responses to the vignettes than respondents from another group, this will show up as a lower cut-point for the first group in the HOPIT estimation.

A second strategy is to calibrate self-report responses using measured tests (instead of vignettes) in conjunction with the HOPIT model. Measured tests are tests of the level of ability of the underlying latent variable for a domain of health. Examples include the posturo-locomotion-manual (PLM) test for mobility, and the Snellen eye chart exam for the domain of vision. Such measured tests are used to estimate cut-point differences across population groups for calibration of self-report responses that are cross-population comparable. The set-up of the HOPIT model with measured tests is quite straightforward. The model assumes that the measured test is correlated with the underlying latent variable for a domain, and the cut-points for a particular categorical response are allowed to differ by population group.

For a variety of reasons including those related to measurement error, it appears that vignettes are a superior mechanism than measured tests for the calibration of self-report responses. Hence, current WHO estimates of outcome measures of health and responsiveness are based primarily on the use of vignettes as a calibration strategy.

4. SPRG comments and recommendations

SPRG welcomed WHO's work in this area and recognized the importance of ensuring that the data used in the HSPA exercise are comparable across populations.

- (i) The HOPIT methodology depends crucially on the assumption that the categorical responses derive from a single dimension (or attribute), which can be ordered on a unilinear scale. SPRG noted the responses should not be based on mappings by individuals that involve comparisons in two or more dimensions (of planar or higher-dimensional regions corresponding to the five categories). Application

of this methodology requires that the domain of each self-report question is narrowly and unambiguously specified.

- (ii) A promising avenue for future research would be to develop statistical methods that combine the information from both vignettes and measured tests in a joint estimation procedure. These methods, akin to the multiple-indicator multiple-cause models in the statistical literature, have the advantage that they take full account of all available information on a given individual, in this case the multiple sources being the individual's self-report (calibrated using vignettes) as well as his/her measured test. These types of methods can also allow for different statistical errors in information that is self-reported and information that is obtained from measured tests.
- (iii) The HOPIT model depends critically on the cross-cultural reliability and consistency of the vignettes – e.g., translation problems or errors do not change the meaning of a question so that a different latent variable is being measured. SPRG recommends that the vignettes be tested further in different settings, including through back translation.
- (iv) It may be possible to explore some of the problems related to (iii) above through a random coefficients version of the HOPIT model. Unlike the current version of the model, a random coefficients model allows the latent variable associated with each vignette to have its own variance (rather than the variance being the same for all vignettes). This method allows one to take account of the possibility that some vignettes may be inherently 'noisier' than others. This may be of particular relevance for vignettes referring to the middle range(s) of a domain, i.e., for vignettes that are not at either extreme of a domain.
- (v) SPRG noted that the HOPIT model not only addressed the problem of cross-population comparability, but also converted the discrete (categorical) information on each domain of health and responsiveness into a *continuous variable*. For each individual the aggregation of these variables across the appropriate domains generates the continuous distribution from which the mean level of, and inequality in, health (or responsiveness) is estimated. Hence, the HOPIT model yields much more than cross-population comparability: it forms the basis for estimating four of the five intrinsic indicators used in HSPA.
- (vi) SPRG members made several technical comments on the HOPIT methodology. Some of these are noted below.
 - (a) The estimates of the cut-points for a population group (e.g., country) will depend on the universe of groups included in the cut-point estimation. For example, suppose the cut-points for group A are estimated from data for groups A and B. Now, data on group C become available and the cut-points for A are re-estimated from data

for all three groups A, B, and C. In general, the cut-points (and other parameter estimates) for group A will change. This could make the relative ranking between, say, groups A and B depend on the precise other groups included in the estimation (especially when considering the aggregates across domains). Hence, caution will need to be exercised in making judgements about the relative ranking between countries, which could be universe-dependent.

(b) The assumption made in the HOPIT model is that the latent variable (e.g., mobility) is unbounded (as the normal distribution is used for the error term). SPRG recommends that the WHO Secretariat check the robustness of their results to restricting the latent variable to a finite interval (e.g., through the assumption of a truncated normal distribution for the error term), as this would seem a more realistic assumption for the domains considered.

(c) SPRG members commented that it would be valuable to estimate non-linear functional forms for the latent variable equation (e.g., health production function), which might also to some extent address the problem noted in (b). A log-linear form for the health production function seems more realistic as it allows for diminishing returns to the factors that determine health (e.g., age, education, etc.), which may be more reasonable than assuming constant returns to each factor. In any case, it would be valuable to check the sensitivity of the present HOPIT results to the assumptions made about the functional form.

5. References

Holland, P. W. and H. Wainer (1993): *Differential Item Functioning*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Murray, C. J. L. and L. C. Chen (1992): Understanding morbidity change. *Population and Development Review*, 18(3): 481-503.

Sen, A. K. (2002): Health: Perception versus observation. *British Medical Journal*, 324: 860-861.

Sen, A. K. (1992): Positional objectivity. *Philosophy and Public Affairs*, 22: 126-145.

