

**THE GLOBAL BURDEN OF DISEASE 2000 IN AGING POPULATIONS**

**Research Paper No. 01 .20**

**Cross-Population Comparability of Physician-Assessed and Self-Reported Measures of Health**

**Kim Moesgaard Iburg  
Joshua A Salomon  
Ajay Tandon  
Christopher JL Murray**

ISSN 0000 0000

**HARVARD BURDEN OF DISEASE UNIT  
NATIONAL INSTITUTE ON AGING GRANT 1-P01-AG17625**

## **THE GLOBAL BURDEN OF DISEASE 2000 IN AGING POPULATIONS**

This working paper series reports on research supported by the National Institute on Aging program grant entitled The Global Burden of Disease 2000 in Aging Populations (1-P01-AG17625). The purpose of the grant is to strengthen the methodological and empirical bases for undertaking comparative assessments of health problems, their determinants and consequences in aging populations.

Since the publication of the Global Burden of Disease Study 1990, there has been increasing interest in comparative analyses of health outcomes, determinants and consequences. A major revision of the Global Burden of Disease Study has been launched for the year 2000 with the full commitment of the World Health Organization (WHO). The Global Programme on Evidence for Health Policy at WHO has developed a Global Burden of Disease Network, which operates in parallel to the research conducted as part of the program project. The program project will strengthen the scientific basis for the large-scale undertaking led by WHO at the global, regional and national level.

The purpose of this working paper series is to present original research that emerges from the various project components of this program grant. The views expressed in these working papers are those of the author(s) and do not necessarily reflect the views of the Harvard Burden of Disease Unit, the World Health Organization nor the National Institute on Aging.

## **THE HARVARD BURDEN OF DISEASE UNIT**

The Harvard Burden of Disease Unit was established to design, test, and implement methodologies to aid in the effective allocation of health resources. To achieve this end, the Unit conducts research in collaboration with national governments, international agencies and other researchers and policy-makers. The Unit's research has two main foci:

- to forge the theory, design, and implementation of approaches to the combined measurement of mortality and non-fatal health outcomes, in order to develop valid, reliable, comparable and comprehensive measures of population health and comparative assessments of the burden of diseases, injuries and risk factors; and
- to investigate the costs, efficacy and effectiveness of major health interventions applied in diverse settings, toward the goal of establishing a broad database on cost-effectiveness.

Harvard Burden of Disease Unit  
Center for Population and Development Studies  
9 Bow Street  
Cambridge, MA 02138  
[www.hsph.harvard.edu/organizations/bdu](http://www.hsph.harvard.edu/organizations/bdu)

## **Cross- Population Comparability of Physician-Assessed and Self-Reported Measures of Health<sup>1</sup>**

Kim Moesgaard Iburg,<sup>2</sup> Joshua A Salomon,<sup>2</sup> Ajay Tandon,<sup>2</sup> Christopher JL Murray<sup>3</sup>

---

<sup>1</sup> This research paper is also available as GPE discussion paper No. 14 (World Health Organization). This work has been supported by the National Institute on Aging Grant 1-P01-AG17625.

<sup>2</sup> Global Programme on Evidence for Health Policy, World Health Organization

<sup>3</sup> Executive Director, Evidence and Information for Policy, World Health Organization and Director, Harvard Burden of Disease Unit

# 1 Introduction

Assessing levels of health on various domains is a key component of measuring population health, evaluating the impact of health interventions and monitoring individual health levels. Meaningful comparisons across countries are useful in setting goals for the improvement of population health and charting progress towards attaining these goals. Comparisons are also useful within countries in order to understand differences in health levels across subpopulations and to measure health inequalities.

Efforts to compare self-reported health status across population subgroups often have indicated major differences between males and females, rich and poor, between different ethnic groups, or across various other demographic and socio-economic variables. One of the challenges in the measurement of individual health, however, has been in interpreting self-reported health data in a way that allows meaningful comparisons. Empirical evidence pointing to concerns about the comparability of self-reported data on health abounds. Substantial evidence shows that disability rates have been falling rapidly over the last two decades while trends in self-reported health have been more mixed. In some studies, higher income groups report higher morbidity than lower income groups, even though observed disability declines rapidly with income (Murray 1996). The challenge is to ascertain how much of the difference is determined by real differences in health and how much is due to differences in the way individuals report on their health relative to different norms and expectations. Previous studies describing comparisons of self-reported health have raised concerns about the face validity of some of these measures, highlighting the fundamental challenge of cross-population comparability.

A number of different approaches have been taken to improve comparability of self-reported data both across and within countries. One approach has been to ensure that specific survey items have the same meaning across languages and different cultural settings. Variations among questions used in health surveys, such as recall periods, definitions of terms, and response categories are documented for 16 surveys conducted in 11 EU countries (Rasmussen et al. 1999) and another 30 surveys from 23 OECD countries (Gudex & Lafortune 2000). The main findings were that variations in question content are prevalent in nearly all health surveys, with the exception of those that use standardized health status instruments such as the SF-36 or EuroQol. The best example of attempts to minimize differences in items in health surveys is probably the adaptation of the standardised SF-36 questionnaire to more than 40 countries (Ware 2000). But while this may help to remove one important barrier to comparisons, it does not account for the fact that individuals may have different expectations for health that are unrelated to linguistic differences in the phrasing of questions.

A key obstacle to cross-population comparability that remains even after reliability and within-population validity of instruments is established relates to the fact that survey questionnaires most commonly elicit categorical responses, which do not provide cardinal values for levels of health, *i.e.*, distances between response categories are not equal, and are unknown. Comparisons are complicated by the fact that different individuals use the

categorical response scales in different ways. We may conceptualise these differences in terms of variation in individual response category cutpoints, which mark the boundaries between categories in reference to an unobserved, continuous latent scale. Attempts to establish equivalent scale endpoints across different questions may offer some benefits in enhancing comparisons, but they cannot account for cutpoint shifts. A recent study by the World Health Organization (Sadana et al. 2000) described a confirmatory factor analytic approach to fix the endpoints of self-reported data in order to improve the comparability of estimated health levels from household interview surveys in 64 countries. Despite efforts to improve comparability of endpoints, the study concluded that a valid and meaningful comparison of existing data on non-fatal health from household interview surveys across countries was limited.

Even in cases where cutpoint differences would seem unlikely, as in binary questions about clearly defined physical phenomena, surprising results have been reported. For example, a study from Ghana (Belcher 1976) showed that missing body parts very rarely were self-reported. Other studies have found large differences between self-reported morbidity and clinical examinations (Krueger 1957), and cross-cultural differences in people's experiences of illness severity and norms for when to seek health services are well-documented (Tsuji et al. 1994, Bletzer et al. 1993, Hunt et al. 1981). When are people feeling sick enough to label themselves as sick persons – or healthy enough to say that they have an excellent health? Clearly these differences may depend not only on cultural influences but also on age, sex, race and socio-economic status. In order to gauge these differences in scale references, exogenous information is required in order to translate categorical responses into comparable cardinal measures.

It is worth noting that problems stemming from the non-comparability of categorical response data apply not only to self-reported information but to any source of data that uses categorical responses. Thus, while it may be useful to distinguish self-reported data from physician assessments, both forms of information are subject to the constraints of the question format that they employ. In both cases, responses to categorical questions will depend critically on the individual cutpoints for a particular question.

Murray et al. (2001) have outlined a series of different strategies for enhancing cross-population comparability of survey results through the formal analysis of systematic cutpoint shifts. One way to address this problem, whether it arises in self-reported or physician-assessed data, is by fixing the levels of the unobserved latent variable of interest in order to isolate cutpoint differences as the source of variation in assessments of these levels. There are several ways of fixing the scales, including the use of vignettes or the inclusion of measured tests. In combination with the new statistical models described in Tandon et al. (2001), the incorporation of this exogenous information allows estimation of variation in cutpoints attributable to socio-demographic or other factors (Salomon et al. 2001).

In this paper, we describe the application of this new approach to the publically-available National Health and Nutrition Examination Survey dataset. Using the results of performance tests on the domain of mobility from this survey, we estimate differences in cutpoints for various sub-groups in the United States population on a range of self-reported

and physician-assessed items relating to this domain. The objective of this paper is to examine whether sex, race/ethnicity and income affect self-reports and physician-assessments of mobility through predictable differences in the use of categorical responses.

## **2 Materials and methods**

### *Data*

Data originate from the third National Health and Nutrition Examination Survey (NHANES III), conducted in the United States from 1988 to 1994 (NCHS 1999). NHANES is a periodic survey conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention, designed to obtain health information on the non-institutionalised civilian population through interviews, physical examinations and laboratory tests. The full data set includes information on approximately 40,000 respondents aged 2 months and older, based on a complex stratified sampling design.

NHANES III includes items relating to a large number of different health domains. For this study, the objective was to identify one domain to illustrate a new approach to estimating cutpoint differences in both self-reported descriptions and physician assessments of health levels. The approach relies on the availability of measured tests that can be used to fix the scale of the unobserved latent variable (level of ability on a particular domain) in order to understand systematic differences in response category cutpoints across different sub-populations.

Our choice of domain for this analysis was guided by a review of existing health status instruments in order to develop an extensive list of candidate domains, followed by a systematic inventory of the number of self-reported, physician assessed and measured items available on various domains. This inventory is summarized in Table 1 using a listing of domains from the Health Utilities Index Mark III (Feeny et al. 1995) as a parsimonious catalogue of key health domains, as well as other components of health captured by various items in the survey. Across different domains, there are varying numbers of items on self-reports, physician assessments and measured performance tests. The domain in NHANES III with the largest number of available items across the three assessment types is physical ability.

**Table 1. Numbers of items in NHANES III by domain and mode of assessment**

<b>Domains</b>	<b>Self-reported</b>	<b>Physician-assessed</b>	<b>Measured tests</b>
<i>Key domains</i>			
Vision	9	3	
Hearing	6	1	2
Speech	1	1	
<b>Getting around (physical ability)</b>	<b>15</b>	<b>5</b>	<b>5</b>
Hands and fingers (dexterity)	5	4	1
Feelings (emotional function)	12	5	
Memory and thinking (cognitive function)	9	1	2
Pain and discomfort	22	2	
<i>Other components of health</i>			
General health	5	9	
Social functioning	12		
BMI	2	1	1
Height	1		1
Weight	2	1	1
Dental status	3	1	
Blood pressure	1		1
Pulse			1
Lung function	3	1	1
Allergy	2	1	1
Blood	2		1
Urine	1		1
Gall bladder	1		1
Ocular photo	1		1

The choice of the domain of physical ability leads to a more narrow focus on those examinees aged 60 years and older, as the physical performance tests in NHANES III are confined to this sub-sample. In total, the final study population includes 5,724 respondents who completed the battery of physical performance measures. Within the domain of physical ability, we have identified the questions of interest more precisely as those pertaining specifically to mobility or ambulation, which resulted in the selection of seven self-reported items, four physician-assessed items, and eight measured tests (Table 2). Self-reported items concern ability to walk, climb stairs, bend down, carry heavy objects, do household chores, and get in and out of bed. Physician assessments refer to abilities to walk, run, bend down, and perform heavy housework and exercise. Selected physical tests relate to shoulder movements, hip and knee flexibility and timed performance measurements, such as walking eight feet and rising repeatedly from an armless chair.

**Table 2. Mobility items and response codes from NHANES III.**

<b>Item</b>	<b>Response codes</b>
<i>Self-reported</i>	
Difficulty walking for a quarter of a mile (about 2 or 3 blocks)	(Note <i>a</i> )
Difficulty walking up 10 steps without resting	(Note <i>a</i> )
Difficulty stooping, crouching or kneeling	(Note <i>a</i> )
Difficulty lifting or carrying something as heavy as 10 pounds (like a sack of potatoes or rice)	(Note <i>a</i> )
Difficulty doing chores around the house (like vacuuming, sweeping, dusting or straightening up)	(Note <i>a</i> )
Difficulty walking from one room to another on the same level	(Note <i>a</i> )
Difficulty getting in or out of bed	(Note <i>a</i> )
<i>Physician-assessed</i>	
Estimated level of difficulty: walking 1/4 mile	(Note <i>b</i> )
Estimated level of difficulty: running 100 yards	(Note <i>b</i> )
Estimated level of difficulty: stooping, crouching, or kneeling	(Note <i>b</i> )
Estimated level of difficulty: doing heavy housework, gardening, exercise or play	(Note <i>b</i> )
<i>Measured tests</i>	
Right shoulder external rotation	(1) full, (2) partial, (3) unable
Left shoulder external rotation	(1) full, (2) partial, (3) unable
Right hip and knee flexion	(1) full, (2) partial, (3) unable
Left hip and knee flexion	(1) full, (2) partial, (3) unable
Time to complete 8-foot walk (mean time from 2 trials)	2 – 60 seconds
Time tandem stand held	(1) 10 or more seconds, (2) 1 – 9 sec, (3) not able
Time to complete five stands (from an armless chair)	2 – 93 seconds

*a* Response codes for self-reports were (1) unable to do, (2) much difficulty, (3) some difficulty, and (4) no difficulty.

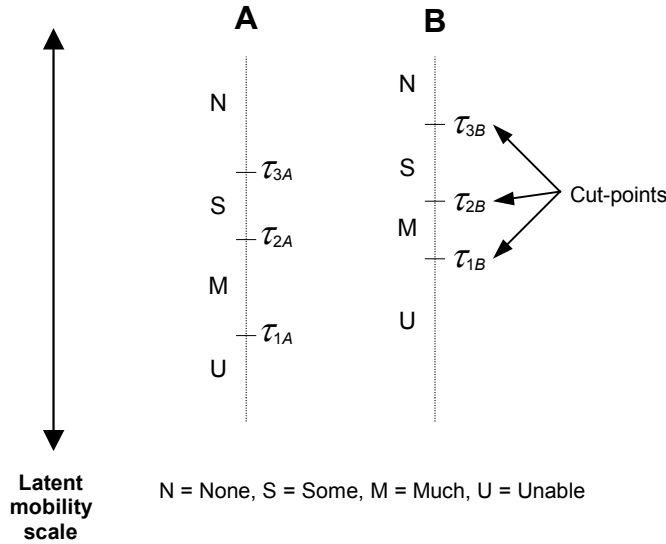
*b* Response codes for physician assessments were (1) could not be done, (2) moderate difficulty, (3) some difficulty, and (4) no difficulty.

As independent variables, we include sex, race/ethnicity, and median family income per capita in the last 12 months, all defined as dichotomous variables. Thus, the combinations of these three variables delineate eight different population sub-groups for examination of differences in response category cutpoints for each of the 11 self-reports and physician-assessments.

### *Statistical analysis*

Missing data. Across the variables included in our analysis, only approximately 64% of the observations include complete information on all variables. In order to address the problem of missing data, we adopt a multiple imputation approach as described by King et al. (1999), using the software program Amelia (Honaker et al. 1999). Five different completed data sets are imputed in order to reflect the uncertainty around the missing values, and all analyses are run separately on each dataset. The results of the five sets of analyses are then combined using standard methods (King et al. 1999).

HOPIT model. The goal of the analysis is to estimate differences in cutpoints across different population sub-groups for either self-reported or physician-assessed categorical questions. The conceptual basis for the statistical model is illustrated in Figure 1. Consider as an example the following question: “How much difficulty do you have in walking for a quarter of a mile?” with response categories “unable to do”, “much difficulty,” “some difficulty” and “no difficulty.” If we assume that there is an unobserved latent variable that represents an individual’s true mobility level, then each individual’s response to this question will depend on his or her cutpoints, which are the threshold levels on the latent scale at which an individual will transition from one category to the next. In Figure 1, we imagine two hypothetical individuals (*A* and *B*) who have different response category cutpoints for this question. At the same true mobility level, individual *A* may respond that she has no difficulty, while individual *B* reports some difficulty in walking for a quarter of a mile. With some knowledge of true mobility levels, it is possible to understand differences in responses to a particular question in terms of cutpoint shifts. In this study, we use results from measured performance tests as a source of information on the unobserved mobility levels of individuals in order to quantify these cutpoint shifts.



**Figure 1. Illustration of cut-point differences for physical ability at different sex, race and income combinations**

In order to estimate cutpoint differences, we apply the hierarchical ordered probit model, an extension of the ordered probit model described in more detail elsewhere (Tandon et al. 2001). The HOPIT model, like the standard ordered probit model, assumes that there is an unobserved latent variable  $Y_i^*$  (e.g., level of mobility) that is distributed normally with mean  $\mu_i$  and variance 1, where  $i$  is an indicator for the respondent.<sup>1</sup> The mean level of the latent variable is described by a function of some set of covariates, in this case a vector of measured test results  $X_i$ ,

$$Y_i^* \sim N(\mu_i, 1)$$

$$\mu_i = X_i' \beta$$

If we define  $y_i$  as the observed categorical response of individual  $i$  to the question of interest (either a self-report or physician assessment), the HOPIT model stipulates an observation mechanism such that, for questions with four response categories:

$$y_i = 1 \text{ if } -\infty \leq Y_i^* < \tau_{1i}$$

$$y_i = 2 \text{ if } \tau_{1i} \leq Y_i^* < \tau_{2i}$$

$$y_i = 3 \text{ if } \tau_{2i} \leq Y_i^* < \tau_{3i}$$

<sup>1</sup> Since the latent variable is unobserved, the variance of the latent variable conditional on determinants is arbitrarily set to 1 in the ordered probit model. In addition, in order to identify the model, the constant term is set to 0. These conventions produce a scale that is unique up to any positive affine transformation, *i.e.*, the latent scale has so-called *interval* properties.

$$y_i = 4 \text{ if } \tau_{3i} \leq Y_i^* < \infty$$

where  $\tau_{1i}$ ,  $\tau_{2i}$  and  $\tau_{3i}$  are the response category cutpoints for individual  $i$ . The key difference between the HOPIT model and the standard ordered probit model is that these cutpoints are allowed to vary as a function of covariates such as sex, race and income:

$$\tau_{ki} = Z_i' \gamma_k$$

Maximum likelihood methods are used to derive estimates of the  $\beta$  and  $\gamma$  coefficients, along with the variance-covariance matrix for these estimators. In this study, the HOPIT model is run on all seven self-reports and all four physician-assessed questions simultaneously, in order to fix the scale of the estimates across questions while allowing cutpoints to vary across questions.

Uncertainty analysis. After the model is run, numerical simulation methods are used in order to compute estimated cutpoints by question for the different sub-populations delineated by sex, race and income, as well as ranges around these estimates. Simulation allows the combination of the results from the five different imputed data sets in a way that reflects the uncertainty of the estimated coefficients both within and across data sets. For each separate data set, ten different draws of the vector of coefficients are generated by sampling from the joint distribution defined by the maximum likelihood estimates of the parameters and their variance-covariance matrix. For each draw, we calculate the predicted cutpoints for the eight different sub-populations defined by the three covariates included in the model (sex, race and income), for each of the self-reported or physician assessed questions. The final distributions for these cutpoints are produced by combining the draws from the five different analyses (creating a total of fifty draws), and we report the median value and the confidence intervals defined by values at the the 2.5<sup>th</sup> percentile and 97.5<sup>th</sup> percentile of this distribution. These estimated cutpoints have been rescaled such that 0 corresponds to the point on the latent scale defined by the worst possible scores on all measured performance tests, and 1 corresponds to the point on the latent scale defined by the best possible scores on all measured tests.

### 3 Results

Characteristics of the study population appear in Table 3. The study includes nearly equal numbers of men and women. The median reported income in the sample is between \$16,000 and \$17,000. Non-hispanic whites make up approximately 59% of the study population. For purposes of sub-group analyses of cutpoint differences, we have defined both race and income dichotomously: race as either white (non-hispanic) or non-white, and annual family income as either above or below \$17,000 per year.

**Table 3. Characteristics of the study population (N = 5,724)**

	Number	Percent
<i>Sex</i>		
Males	2,756	48.2
Females	2,968	51.9
<i>Race-ethnicity</i>		
Non-hispanic white	3,364	58.8
Non-hispanic black	1,125	19.7
Mexican-American	1,067	18.6
Other	168	2.9
<i>Family income (last 12 months)</i>		
Less than \$10,000	1,397	24.4
\$10,000 to 16,999	1,191	20.8
\$17,000 to 29,999	1,290	22.5
\$30,000 to 49,999	706	12.3
\$50,000 and over	438	7.7
Unknown	702	12.3

Table 4 presents a summary of the HOPIT results on cutpoint differences by sex, race and income for both self-reported and physician-assessed items. For each question, the direction and magnitude of shifts are reported for the three response category cutpoints:  $\tau_3$  marks the transition from “no difficulty” to “some difficulty”;  $\tau_2$  the transition from “some” to “moderate / much difficulty”; and  $\tau_1$  the transition from “moderate / much difficulty” to “unable to do”. Of most interest in a general health survey like NHANES III is perhaps  $\tau_3$  because the largest proportion of respondents is often found in the mildest category of many questions – a phenomenon characterised as a “ceiling effect” in population surveys.

**Table 4. Results from HOPIT analysis of self-reported and physician-assessed mobility: cutpoint differences by sex, race and income.**

Item	Sex (male=1)		Race (non-white=1)		Income (high=1)	
	Coef.	p-value	Coef.	p-value	Coef.	p-value
<i>Self-report</i>						
$\tau_3$						
Walking 1/4 mile	<b>-0.171</b>	<0.001	-0.016	0.682	<b>-0.287</b>	<0.001
Walk up 10 steps	<b>-0.290</b>	<0.001	<b>0.187</b>	<0.001	<b>-0.275</b>	<0.001
Stooping, crouching, kneeling	<b>-0.219</b>	<0.001	<b>-0.203</b>	<0.001	<b>-0.179</b>	<0.001
Carrying 10 pounds	<b>-0.476</b>	<0.001	<b>-0.189</b>	<0.001	<b>-0.253</b>	<0.001
Chores around the house	<b>-0.287</b>	<0.001	-0.031	0.438	<b>-0.182</b>	<0.001
Walking room to room	-0.081	0.123	<b>0.165</b>	0.002	<b>-0.152</b>	0.004
Getting in or out of bed	-0.075	0.084	-0.064	0.147	<b>-0.121</b>	0.005
$\tau_2$						
Walking 1/4 mile	<b>-0.175</b>	<0.001	-0.077	0.079	<b>-0.232</b>	<0.001
Walk up 10 steps	<b>-0.227</b>	<0.001	<b>0.108</b>	0.017	<b>-0.235</b>	<0.001
Stooping, crouching, kneeling	<b>-0.253</b>	<0.001	<b>-0.180</b>	<0.001	<b>-0.187</b>	<0.001
Carrying 10 pounds	<b>-0.312</b>	<0.001	<b>-0.159</b>	<0.001	<b>-0.192</b>	<0.001
Chores around the house	<b>-0.127</b>	0.011	-0.068	0.184	<b>-0.154</b>	0.002
Walking room to room	0.050	0.522	0.068	0.388	-0.084	0.282
Getting in or out of bed	-0.079	0.268	0.112	0.111	-0.032	0.649
$\tau_1$						
Walking 1/4 mile	<b>-0.135</b>	0.006	<b>-0.316</b>	<0.001	<b>-0.231</b>	<0.001
Walk up 10 steps	<b>-0.211</b>	<0.001	-0.019	0.741	<b>-0.246</b>	<0.001
Stooping, crouching, kneeling	<b>-0.208</b>	<0.001	<b>-0.063</b>	<0.001	<b>-0.175</b>	<0.001
Carrying 10 pounds	<b>-0.231</b>	<0.001	-0.082	0.242	-0.093	0.079
Chores around the house	0.002	0.976	0.062	0.185	0.009	0.879
Walking room to room	0.141	0.189	0.082	0.567	-0.044	0.683
Getting in or out of bed	0.009	0.947	-0.238	0.526	-0.010	0.940
<i>Physician-assessment</i>						
$\tau_3$						
Walking 1/4 mile	<b>-0.114</b>	0.002	<b>-0.171</b>	<0.001	<b>-0.381</b>	<0.001
Running 100 yards	<b>-0.197</b>	<0.001	<b>-0.148</b>	0.001	<b>-0.422</b>	<0.001
Stooping, crouching, kneeling	<b>-0.208</b>	<0.001	<b>-0.129</b>	<0.001	<b>-0.386</b>	<0.001
Heavy housework, exercise, etc.	<b>-0.190</b>	<0.001	<b>-0.153</b>	<0.001	<b>-0.394</b>	<0.001
$\tau_2$						
Walking 1/4 mile	<b>-0.092</b>	0.019	<b>-0.185</b>	<0.001	<b>-0.369</b>	<0.001
Running 100 yards	<b>-0.170</b>	<0.001	<b>-0.163</b>	<0.001	<b>-0.343</b>	<0.001
Stooping, crouching, kneeling	<b>-0.210</b>	<0.001	<b>-0.171</b>	<0.001	<b>-0.354</b>	<0.001
Heavy housework, exercise, etc.	<b>-0.112</b>	0.002	<b>-0.131</b>	<0.001	<b>-0.397</b>	<0.001
$\tau_1$						
Walking 1/4 mile	0.063	0.206	<b>-0.238</b>	<0.001	<b>-0.296</b>	<0.001
Running 100 yards	<b>-0.142</b>	<0.001	<b>-0.136</b>	<0.001	<b>-0.300</b>	<0.001
Stooping, crouching, kneeling	-0.059	0.258	<b>-0.192</b>	<0.001	<b>-0.244</b>	<0.001
Heavy housework, exercise, etc.	0.012	0.774	<b>-0.135</b>	0.002	<b>-0.343</b>	<0.001

The HOPIT regression results show that there are significant differences by sex, race and income in individual cutpoints separating “no difficulty” from “some difficulty” ( $\tau_3$ ) for all physician-assessed questions. In all cases, the direction of the effects are the same, with lower cutpoints for males compared to females, for nonwhites compared to whites, and for high income respondents compared to low income respondents. A lower cutpoint may be interpreted as a lower standard for defining excellent mobility levels; in other words, given the same level of mobility, an individual with a lower cutpoint will be more likely to characterise this level of mobility favourably than an individual with a higher cutpoint.

For self-reported items, only income is a statistically significant predictor of differences in  $\tau_3$  for all questions. Sex is statistically significant for five out of seven questions and race for four out of seven questions. Where coefficients are significant, the directions of the effects are the same as in the physician assessments for all cases except for race in the items relating to walking 10 steps and walking from room to room.

Similar patterns emerge with respect to systematic differences in  $\tau_2$  for both physician assessments and self-reports, although the overall magnitude of the differences tends to be slightly smaller. A notable exception is the effect of race on physician assessments, where the size of the differences is greater for  $\tau_2$  than  $\tau_3$  on all questions. There are fewer significant effects on  $\tau_1$  in both self-reported and physician-assessed items. Nevertheless, there is remarkable consistency in the direction of the effects on nearly all of the significant results on all cutpoints and questions.

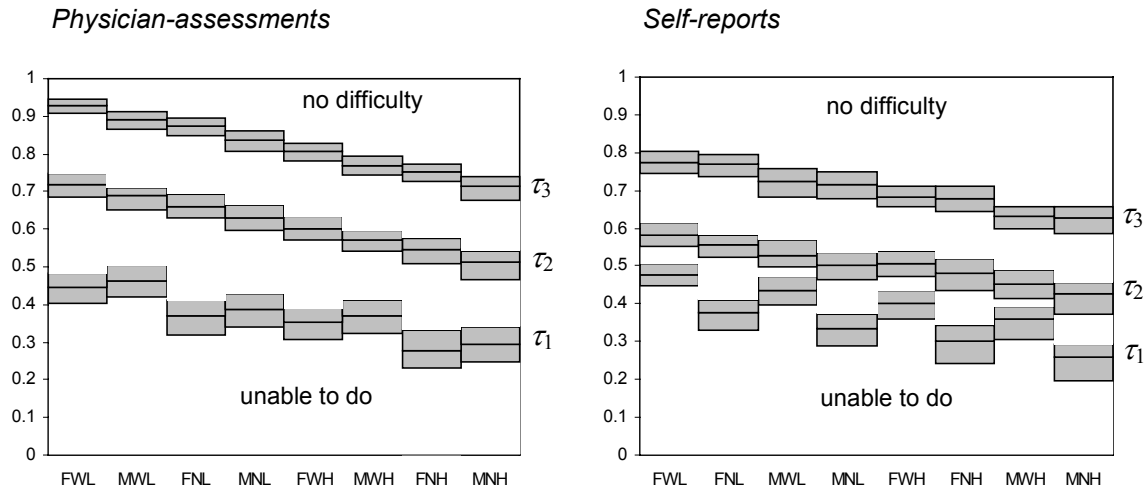
Based on the results from the HOPIT regression, we can estimate predicted cutpoint values on each question for different subgroups in the sample, as well as ranges around these estimates (Figure 2). The pattern for the estimated cut-points for population groups is quite similar for all seven self-reported and four physician-assessed items. The example in Figure 2 depicts cutpoints on two parallel questions relating to difficulties in walking one-quarter mile. The cutpoints are located on the latent mobility scale that emerges from HOPIT, after rescaling based on anchors defined by the best and worst results on the range of performance tests, as described above. While the 95% confidence intervals show that there is some overlap between estimated cutpoints in different groups, significant differences remain between various subgroups in the study population. Thus, for example, at the same level of health, a white female with low income will report some difficulties in walking one-quarter mile, while a non-white male with high income will report no difficulties.

This example highlights two key findings:

(1) that the ordering of cutpoints across the eight subgroups is almost identical in physician assessments as compared to self-reports. In other words, the variation in norms and expectations associated with sex, race and income may apply not only to individuals assessing their own health levels, but also to medical professionals making these assessments for their patients. In fact, judging by the range in cutpoints across subgroups,

differences in norms and expectations may even be larger for physicians than for self-reporting individuals.

(2) that physician cutpoints tend to be higher overall for the threshold between “no difficulty” and “some difficulty,” but slightly lower overall for the threshold between “much / moderate difficulty” and “unable to do” ; in other words, given a relatively high level of health, physicians will be more likely to characterize individuals as having difficulties on this domain than the individuals themselves, whereas physician appraisals of low levels are more generous.



**Figure 2. Estimated cutpoints and ranges by population subgroup for self-reports and physician assessments of difficulties in walking one-quarter mile. Bars indicate 95% confidence intervals. Abbreviations for population groups are as follows: F = female, M = male, W = white, N = non-white, L = low income, H = high income.**

## 4 Discussion

Self-reported health is a complex function of observed morbidity, health expectations, contact with health services or other sources of health knowledge, and social and cultural context. A number of studies have illustrated the difficulties of comparing responses to self-reported health questions across individuals who differ in terms of various socio-economic, demographic or cultural factors. A classic example is from Kerala state in India, where rates of self-reported morbidity have been found to be higher than in anywhere else in India, despite the fact that Kerala has the lowest mortality rates in India (Murray & Chen 1992). Another noteworthy example comes from a study in Australia (Mathers & Douglas 1998) in which the Aboriginal population describe their health levels much more favourably than the general population, even though the opposite would be expected given incidence rates of major health problems and other key health indicators. Survey results such as these ones suggest that there are major problems of comparability that make meaningful cross-population analyses difficult.

One important strategy to improve cross-population comparability of health surveys has been to improve the standardization of questionnaires across countries. Even if cross-cultural standardization in questions could be achieved, however, a major challenge to comparability remains in the reliance on categorical responses, which will be subject to individual variation in the use of the available response categories, even for instruments with established reliability and validity. An individual's use of categories such as "no difficulties," "some difficulties," *etc.* will depend on their norms and expectations for health along a particular domain. Johansson (1992) has pointed to a "cultural inflation of morbidity," in which expectations for health rise in countries as they undergo the health transition to lower rates of mortality, but even within populations, it is easy to imagine that expectations will vary according to age, sex, education, income, and a host of other variables. It is therefore essential to adjust categorical questions to a comparable cardinal scale before attempting comparisons between countries or across population groups within countries.

In this paper, we describe the application of a new method for addressing the problem of comparability in categorical responses to health questions, using measured tests to capture fixed levels on a latent health domain in order to quantify individual differences in response category cutpoints. We apply the hierarchical ordered probit (HOPIT) model described in Tandon et al. (2001) to information from the Third National Health and Nutrition Examination Survey in the United States to develop empirical estimates of cutpoint differences in population sub-groups defined by sex, race and income. This study focuses on the domain of mobility, using information from almost six thousand adults aged 60 years and older, but any data set containing both self-reports or physician-assessments and measured performance tests relating to the same domain of health could be analysed using the same approach.

The results of this study point to significant differences in individual response category cutpoints as a function of sex, race and income, on several different questions relating to mobility levels. Across different questions, the nature of the differences are largely consistent. There is a strong tendency for males to have lower cutpoints on mobility questions than females, which implies that males are less likely to report difficulties given the same levels of mobility. The frequently observed pattern in many health surveys in which women report worse health than men may therefore be understood not simply as an indicator of lower health levels, but of higher expectations for health (*i.e.*, higher cutpoints). Our findings also point to lower cutpoints for non-white respondents relative to white respondents, which again may suggest important differences in expectations for health that lead to different characterizations of the same fixed levels on a particular domain.

A somewhat surprising finding is that individuals with higher income levels tend to have lower cutpoints for mobility than individuals with lower income levels. This result runs counter to the notion that standards for excellent health increase with rising income, but might be understood in terms of a "wishful thinking" scenario in which wealthier individuals have a belief that they *should* be in excellent health and therefore use liberal standards for excellence in reporting on their own health. These findings may have

important implications for the measurement of social inequalities in health where these measures rely on categorical self-reported data. If wealthier respondents use more lenient standards in reporting on their own health levels, this could exaggerate disparities in health between the rich and poor.

It is interesting to observe that both self-reported and physician-assessed health measures are subject to the same variations in cutpoints relating to socio-demographic factors such as sex, race and income. In this study, the differences were even more marked for physician assessments than for individual self-reports. Thus, the way physicians characterize the mobility levels of different individuals with the same performance levels on measured tests will depend on whether the examinee is a man or woman, white or non-white, and of higher or lower economic status. The results from our study add to previous work that has examined potential biases in physician evaluations according to certain patient characteristics. For example, a study from Norway found that general practitioners' awareness of their patients' psychosocial problems were dependent on the age and sex of both the doctor and the patient, as well as the patient's educational level and living conditions (Guldbrandsen et al. 1997). Our study also finds that physicians and patients, when asked the same question (such as the one relating to difficulties in walking one-quarter mile), will characterize the same fixed mobility levels in different ways. A previous study has noted that patients with multiple sclerosis appear less concerned than their clinicians about physical disabilities caused by their illness (Rothwell et al. 1997), and significant differences have also been found between patients and geriatric experts in Europe and United States when comparing the importance of functional status items (Kane et al. 1998). Understanding how patient characteristics influence physician evaluations, and how these evaluations differ from those of the patients themselves, remain important topics for further study.

## 5 References

1. Belcher DW, Neumann AK, Wurapa FK, Lourie IM. Comparison of morbidity interviews with a health examination survey in rural Africa. *American Journal of Tropical Medicine and Hygiene* 1976; 25: 751-758
2. Bletzer KV. Perceived severity: Do they experience illness severity as we conceive it? *Human Organization* 1993;52:1:68-75
3. Feeny DH, Furlone W, Boyle MH, Torrance GW. Multiple-attribute health status classification systems: Health Utility Index. *Pharmacoeconomics* 1995;7:490-502
4. Gudex C, Lafortune G. An inventory of health and disability-related surveys in OECD countries. Directorate for Education, Employment, Labour and Social Affairs. OECD, Paris, 2000

5. Guldbrandsen P, Hjortdahl P, Fugelli P. General practitioners' knowledge of their patients' psychosocial problems: multipractice questionnaire survey. *BMJ* 1997; 314:1014-1018
6. Honaker J, Joseph A, King G, Scheve K, Naunihal S. *Amelia: A programme for missing data*. Harvard University, 2001.
7. Hunt SM, McKenna SP, McEwen J, Williams J, Papp E. The Nottingham Health Profile: Subjective health and medical consultations. *Soc Sci Med* 1981;15A;221-229
8. Johansson SR. The health transition: The cultural inflation of morbidity during the decline of mortality. *Health Transition Review* 1992;2:78-89
9. Kane RL, Rockwood T, Philip I, Finch M. Differences in valuation of functional status components among consumers and professionals in Europe and the United States. *Journal of Clinical Epidemiology* 1998; 51:657-666
10. King G, Honaker J, Joseph A, Scheve K. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review* 2001; 95(1): 49-69.
11. Krueger DE. Measurement of prevalence of chronic disease by household interviews and clinical evaluations. *American Journal of Public Health* 1957; 47: 953-960
12. Mathers CD, Douglas RM. Measuring progress in population health and wellbeing. In: Eckersley R (ed.). *Measuring progress: Is life getting better?* CSIRO Publishing, Collingwood Vic, pp 125-155, 1998
13. Murray CJL, Chen LC. Understanding morbidity change. *Population and Development Review* 18, No. 3, 1992
14. Murray CJL, Tandon A, Salomon J, Mathers CD. Enhancing cross-population comparability of survey data. *GPE Discussion Paper No. 35*. Geneva, World Health Organization, 2000.
15. NCHS. *The third US National Health and Nutrition Examination Survey (NHANES III) 1988-94*. National Centres for Health Statistics 1999
16. Rasmussen N, Gudex C, Christensen S. Survey data on disability. *Eurostat Working Papers, Population and Social Conditions 3/1999/En 29*, European Commission, Luxembourg.
17. Rothwell PM, McDowell Z, Wong CK, Dorman PJ. Doctors and patients don't agree: cross sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. *BMJ* 1997; 314:1580-1583
18. Sadana R, Mathers CD, Lopez AD, Murray CJL, Iburg K. Comparative analyses of more than 50 household surveys on health status. *GPE Discussion Paper No. 15*. Geneva, World Health Organization, 2000.

19. Salomon JA, Tandon A, Murray CJL. Using vignettes to improve cross-population comparability of health surveys: concepts, design and evaluation techniques. Global Programme on Evidence for Health Policy Discussion Paper No. 41. Geneva, World Health Organization, 2001.
20. Tandon A, Murray CJL, Salomon JA. Statistical methods for enhancing cross-population comparability. Global Programme on Evidence for Health Policy Discussion Paper No. 42. Geneva, World Health Organization, 2001.
21. Tsuji I, Minami Y, Keyl PM, Hisamichi S, Asano H, Sato M, Shinoda K. The predictive power of self-rated health, activities of daily living, and ambulatory activity for cause-specific mortality among elderly: A three-year follow-up in urban Japan. JAGS 1994; 153-156
22. Ware J. SF-36 Health survey update. SPINE 2000;25 (24):3130-3139