Ethnic Differentials in Responses to Self-reported Health: An Application of Item Response Theory (IRT)

Rania Tfaily[*] **Beth J. Soldo**[†]

September 23, 2005

Abstract

Despite its utility as an overall summary indicator of health, self-reported health (SRH) has limited value in comparative studies of health inequalities. The SRH variable suffers from problems in measurement comparability (due to linguistic and cultural factors) and/or differences in interpretation of the question and its categories. Item response theory (IRT) offers a method to test the meaningfulness of the self-reported health measure and to compare its categorical cut-points across different groups. In this paper, we apply the Graded Response Model (GRM), a generalization of the two-parameter logistic IRT model to study the measurement of self-reported health using data from the 1998 Health and Retirement Study (HRS) and the 2001 Mexican Health and Aging Study (MHAS). Category response curves for each response category of self-reported health are drawn for the following ethnic group: White-Americans, African-Americans, Mexican-Americans and Mexicans with and without controls for age, gender and education.

Introduction

A large number of studies showed that self-reported health is a good summary indicator of overall health in different settings (Benyamini and Idler 1999; Idler and Benyamini 1997; Kuhn et al. 2004; Rahman and Barsky, 2003; Frankenberg and Jones 2003; Zimmer et al. 2000). However, the use of self-reported health has many drawbacks in studies of health inequalities and cross-cultural comparisons (Franks et al., 2003; Humphries and van Doorslaer, 2000; Liu and Zhang, 2004). In contrast to non-Hispanic White Americans, for example, Mexican Americans tend to rate their health less favorably than the physicians' assessments. This is even more pronounced among Spanish speakers (Angel and Cleary 1984; Angel and Guarnaccia 1989; Shetterly et al. 1996). Differences in language and cultural norms affect the comparability of self-reported health in cross-national studies (Angel and Guarnaccia 1989; Jylhä et al. 1998; Wagner et al. 1998). In certain settings there is aversion to positively assess

[*] Ph.D. student, University of Pennsylvania, 239 McNeil Building, 3718 Locust Walk, Philadelphia, PA 19104-6298. Tel: +1-(215)-898-9641, Fax: +1-(215)-898-2124, Email: rania2@ssc.upenn.edu.

[†] *Professor of Sociology(adj) and DistinguishedRresearchSscholar, Population Studies Center, University of Pennsylvania, 239 McNeil Building, 3718 Locust Walk, Philadelphia, PA 19104-6298. Tel: +1-(215)-898-1535, Fax: +1-(215)-898-2124, Email: bsoldo@pop.upenn.edu.*

health even in the absence of diseases and pain out of modesty or fear of invoking illnesses (Frankenberg and Jones 2003; Rahman and Barsky 2003; Kuhn et al. 2004; Wagner et al. 1998).

Cross-cultural or group variations in the latent trait –health- could in fact be due to two factors: differences in degree (e.g. differences in distribution of illnesses across groups) and differences in kind[1] (differences in the interpretation of self-reported health). For group comparisons to be meaningful there should be no differences in kind. In other words, the measurement of the items related to health should be invariant across groups. Otherwise, conclusions would be artifactual and misleading (Andrich 1988; Reise et al. 1993).

Item Response Theory offers a method for comparing the measurement of items across groups. The theory underlying IRT is the assumption that the response to an item is a function of the respondent's latent trait level and the characteristics of the item in question (Andrich, 1988; Embretson and Reise, 2000). IRT also allows testing for items that display differential item functioning (DIF) (Embretson and Reise, 2000). IRT has been used mostly in research in education and psychology, but there have been increasing applications of this method in demography and social sciences (see for example, Ghuman et al., 2004; MacIntosh, 1998; Smith and Furstenberg, 1994; Smith and Morgan, 1994).Using data from the 1998 HRS and 2001 MHAS, we use IRT to examine how the measurement of self-reported health and its category cut points differ across ethnic groups.

ITEM-RESPONSE THEORY

Item Response Theory is designed to test whether the measuring device is invariant across groups. IRT uses the same scale to measure question items and trait levels of the respondents. The scale of measurement generally has an arbitrary midpoint of zero, a unit measurement of one and values that range from -3 to +3. The IRT function shows how changes in the probability of a given response are related to changes in the trait level (Baker, 2001; Embretson and Reise, 2000). "In other words, the function, describes, in probabilistic terms, how a person with a higher standing on a trait (i.e., more of the trait) is likely to provide a response in a different response category to a person with a low standing on the trait" (Ostini and Nering, 2006: 2). As a function of the trait, the probability curve of a binary item has a smooth S-shaped curve (unless the item has a low discrimination index). This is known as item characteristic curve and it is item-specific (Baker, 2001; Embretson and Reise, 2000).

The item characteristic curve has 2 characteristics: location (also called difficulty in education research) and slope (also called discrimination). Location describes the position of the item along the trait continuum. The inflection point on the curve of a

[1] An item is considered unbiased if respondents who belong to different groups but otherwise have the same level of the latent trait have the same probability of choosing each of the categories of the item in question (Andrich, 1988).

binary item corresponds to the probability of 0.5, at which the respondent has an equal chance of agreeing or rejecting the statement in the item. The inflection point of an easy item occurs at lower trait level than that of a difficult item (i.e. the item characteristic curves of difficult items are lower than those of easier items). The second characteristic of the item – slope – is a measure of the steepness of the curve and indicates how well the item distinguishes among respondents with traits above the item inflection point from those below it. Discrimination is also a measure of reliability. Items with more discrimination have steeper[2] and more reliable curves than easier items. The discrimination index of the item is independent of its location (Baker, 2001; Embretson and Reise, 2000; Schaeffer, 1988). An important feature of IRT is that the estimates of the item parameters are independent of the distribution of the trait in the sample (Bond and Fox, 2001).

We describe, below, the IRT function for binary items and then discuss how it differs for polytomous variables. The two-parameter Logistic model describes the probabilistic relationship between trait level and the two characteristics of the item:

$P\,(X_{is} = 1 \mid \theta_s, \beta_i) = \exp\,(\alpha_i\,(\theta_s - \beta_i))/\ 1 + \exp\,(\alpha_i\,(\theta_s - \beta_i))$
Where:
X_{is} is the response of respondent s to item i
θ_s is the trait level of respondent s
β_i is the difficulty value of item i
α_i is the discrimination value of item i

There are two other IRT Logistic models. The simplest model is the one-parameter Logistic model or Rasch model in which items are assumed to be equally discriminating (discrimination parameter is set at 1). In this case, only the location parameter varies from one item to another. In Rasch model, all respondents who have the same total number of items answered affirmatively have the same estimated trait level. The third IRT model allows for the possibility that a person could report the “true” answer by guessing. The third model is more prevalent in education research in which responses are classified as right or wrong. In the above three models, unidimensionality and local independence are assumed. Unidimensionality specifies that the items measure one attribute or dimension of the trait, while local independence indicates that controlling for the respondent’s trait level, the probability of agreeing with an item is independent of the probability of agreeing with another. These assumptions can be checked (Baker, 2001; Bond and Fox, 2001; Embretson and Reise, 2000).

In the case of polytomous items, each category function is modeled separately. Each item is characterized by one slope (or discrimination) parameter and k -1 between category threshold parameters; where k is the number of item response categories. For example, an item such as SRH is characterized by having one slope parameter and four threshold parameter. Multiple dichotomizations of item response take place (e.g. categories 1 vs. 2, 3, 4, 5; 1, 2 vs. 3, 4, 5; 1, 2, 3 vs. 4, 5; & 1, 2, 3, 4 vs. 5) in order to draw the category

[2] The probability of affirmative response of a binary item increases rapidly with the increase in trait

response function for each item category. The category response functions are no longer exclusively monotonic as in binary items except for the two categories at the extreme such as poor and excellent in the case of SRH (Ostini and Nering, 2006).

METHODOLOGY

Data

The Health and Retirement Study (HRS) is a prospective panel study of the U.S. population of the birth cohorts born <1953. The 1998 HRS is representative of the U.S. non-institutionalized population aged 50 and over at the two time point – 1998 and 2004—when younger cohorts were aged-in to the study from the bottom. The HRS collects extensive information on socio-demographic conditions, marital history, completed fertility, living arrangements, and various health domains: physical, functional and cognitive, and affective health, chronic conditions, and symptoms, lifestyle behavior (smoking, drinking), and migratory history.

The Mexican Health and Aging Study (MHAS)/Encuesta Nacional Sobre Salud y Envejecimiento en Mexico (ENSEM) is also a prospective panel study, similar in design, content, and coverage to the HRS. All MHAS respondents were aged 50 and over at the time of the 2001 baseline. Like HRS, MHAS provides information on the various health domains listed above. The first wave of MHAS (2001) is representative of the 13 million Mexicans born prior to 1951.The MHAS sample was selected from households participating in the 4th Quarter 2000 National Employment Study/Encuesta Nacional de Empleo (ENE), nationally representative study. The 2001 MHAS response rate was 90.1% and 15,186 eligible respondents and their spouses/partners were successfully interviewed.

Measures and Analyses

In addition to self-reported health, the following items are used in the analyses: bodily pain, symptoms such as frequent swelling in feet and ankles; difficulty breathing; fainting spells; and persistent wheezing cough or bringing up phlegm) and Activities in Daily Living (difficulty in dressing, bathing, eating, getting in and out of bed, and using the toilet). As these items do not have the same number of response categories, we use the Graded-Response Model (GRM), which is a generalization of the 2PL model (Embretson and Reise, 2000; Ostini and Nering, 2006). A number of control variables are used: ethnicity and immigrant status, gender, age and education.

We carry-out the analyses using software Multilog 7 as it could provide estimates of the 2 PL model and polytomous items, as well (Embretson and Reise, 2000). We use differential item functioning (DIF) to examine whether self-reported health have measurement properties that are invariant across groups. DIF tests if the threshold parameters of SRH are systemically different for different groups (Ellis and Kimmel, 1992). Category response curves of self-reported health are then drawn (and compared) for the following ethnic groups: White Americans, African Americans, Mexican Americans and Mexicans) with and without controls for age, gender and education.

REFERENCES

Andrich, David. 1988. *Rasch Models for Measurement. Sage University Paper series on Quantitative Applications in the Social Sciences*, series no. 07-068. Beverly Hills, Sage Pubns.

Angel, R. and P. Cleary.1984. The effects of social structure and culture on reported health. *Social Science Quarterly* 65*:* 814-828.

Angel, R. and P. J. Guarnaccia. 1989. Mind, body, and culture: somatization among Hispanics. *Social Science & Medicine 28*: 1229-1238.

Baker, Frank B. 2001. *The Basics of Item Response Theory.* U.S.: ERIC Clearinghouse on Assessment and Evaluation.

Benyamini, Y. and E. L. Idler.1999. Community studies reporting associations between self-rated health and mortality. *Research on Aging* 21: 392-401.

Bond, Trevor G., and Christine M. Fox. 2001. *Applying the Rasch Model. Fundamental Measurement in the Human Sciences.* New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Ellis, Barbara B., and Herbert D. Kimmel. 1992. "Identification of Unique Cultural Response Patterns by Means of Item Response Theory." *Journal of Applied Psychology* 77, no. 2: 177-184.

Embretson, Susan E., and Steven P. Reise. 2000. *Item Response Theory for Psychologists.* New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Frankenberg, E. and N. R. Jones. 2003. Self-rated health and mortality: Does the relationship extend to a low income setting? California Center for Population Research. On-Line Working Paper Series. Paper CCPR-021-03.

Franks, Peter, Marthe R. Gold, and Kevin Fiscella. 2003. "Sociodemographics, self-rated health, and mortality in the U.S." *Social Science & Medicine* 56: 2505-2514.

Ghuman, Sharon J., Helen J. Lee, and Herbert L. Smith. 2004. "Measurement of Women's Autonomy according to Women and Their Husbands: Results from Five Asian Countries." PSC Research Report 04-556.

Humphries, Karin H. and Eddy van Doorslaer. 2000. "Income-related health inequality in Canada." *Social Science & Medicine* 50: 663-671.

Idler, E. L. and Y. Benyamini. 1997. Self-rated health and mortality: a review of twenty-seven community studies. *Journal of Health and Social Behavior* 38: 21-37.

Jylhä, M., J. M. Guralnik, L. Ferrucci, J. Jokela, and E. Heikkinen. 1998. Is self-rated health comparable across cultures and genders? *Journal of Gerontology* 58B(3): S144-S152.

Kuhn, Randall, Omar Rahman, and Jane Menken. 2004. "relating Self-Reported and Objective Health Indicators to Adult Mortality in Bangladesh." Paper presented at Population Association of America (PAA) annual meeting, April 1-3, Boston, MA.

Liu, Guiping, and Zhen Zhang. 2004. "socioeconomic differentials of the self-rated health of the oldest-old Chinese" *Population Research and Policy Review* 23: 117-133.

MacIntosh, Randall. 1998. “Globe Attitude Measurement: An Assessment of the World Values Survey Postmaterialism Scale.” *American Sociological Review* 63: 452-464.

Ostini, Remo, and Michael L. Nering. 2006. Polytomous Item Response Theory Models. *Sage University Paper series on Quantitative Applications in the Social Sciences*, series no. 07-144. Thousand Oaks, Sage Pubns.

Rahman, M. Omar, and Arthur J. Barsky. 2003. “Self-Reported Health Among Older Bangladeshis: How Good a Health Indicator Is It.” *The Gerontologist* 43: 856-878.

Reise Steven P., Keith F. Widaman, and Robin H. Pugh. 1993. “Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance.” *Psychological Bulletin* 114, no. 3: 552-566.

Schaeffer, Nora Cate. 1988. “An Application of Item Response Theory to the Measurement of Depression.” *Sociological Methodology* 18: 271-307.

Shetterly, S. M., J. Baxter, L. Mason, and R. Hamman.1996. Self-rated health among Hispanics vs. non-Hispanic White adults: the San Luis Valley Health and Aging Study. *American Journal of Public Health* 86: 1798-1801.

Smith, Herbert L., and Frank F. Furstenberg, Jr. 1994. “Application of a Response Model for Mother-Daughter Agreement by Race.” *Social Science Research* 23: 136-166.

Smith, Herbert L., and S. Philip Morgan. 1994. “Children’s Closeness to Father as Reported by Mothers, Sons and Daughters: Evaluating Subjective Assessments with the Rasch Model.” *Journal of Family Issues* 15: 3-29.

Wagner, A., B. Gandek, N. Aaronson, C. Acquadro, J. Alonso, G. Apolone, M. Bullinger, J. Bjorner, S. Fukuhara, S. Kaasa, A. Leplége, M. Sullivan, S. Wood-Dauphinee, et al. 1998. Cross-cultural comparisons of the content of SF-36 translations across 10 countries: Results from the IQOLA Project. *Journal of Clinical Epidemiology* 51: 925-932.

Zimmer, Z., J., J. Nativida, H. Lin, and N. Chayovan. 2000. A cross-national examination of the determinants of self-assessed health. *Journal of Health and Social Behavior* 41: 465-481.