Chan, K.S., Kasper, J.D., Brandt, J., & Pezzin, L.E. (2012). Measurement equivalence in ADL and IADL difficulty across international surveys of aging: findings from the HRS, SHARE, and ELSA. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 67(1)*, 121–132, doi:10.1093/geronb/gbr133. Advance Access published on December 12, 2011

Measurement Equivalence in ADL and IADL Difficulty Across International Surveys of Aging: Findings From the HRS, SHARE, and ELSA

Kitty S. Chan,[1] Judith D. Kasper,[1] Jason Brandt,[2,3] and Liliana E. Pezzin[4]

[1]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland.
[2]Department of Psychiatry and Behavioral Sciences.
[3]The Copper Ridge Institute, Sykesville, Maryland.
[4]Department of Medicine and Health Policy Institute, Medical College of Wisconsin, Milwaukee.

Objective. To examine the measurement equivalence of items on disability across three international surveys of aging.

Method. Data for persons aged 65 and older were drawn from the Health and Retirement Survey (HRS, n = 10,905), English Longitudinal Study of Aging (ELSA, n = 5,437), and Survey of Health, Ageing and Retirement in Europe (SHARE, n = 13,408). Differential item functioning (DIF) was assessed using item response theory (IRT) methods for activities of daily living (ADL) and instrumental activities of daily living (IADL) items.

Results. HRS and SHARE exhibited measurement equivalence, but 6 of 11 items in ELSA demonstrated meaningful DIF. At the scale level, this item-level DIF affected scores reflecting greater disability. IRT methods also spread out score distributions and shifted scores higher (toward greater disability). Results for mean disability differences by demographic characteristics, using original and DIF-adjusted scores, were the same overall but differed for some subgroup comparisons involving ELSA.

Discussion. Testing and adjusting for DIF is one means of minimizing measurement error in cross-national survey comparisons. IRT methods were used to evaluate potential measurement bias in disability comparisons across three international surveys of aging. The analysis also suggested DIF was mitigated for scales including both ADL and IADL and that summary indexes (counts of limitations) likely underestimate mean disability in these international populations.

Key Words: Activities of daily living—Differential item functioning—Disability.

LIMITATIONS in self-care and ability to do activities necessary for independent community living are commonly used indicators of disability and a central focus of research and policy aimed at reducing disability in older people. Disability has individual and societal consequences that include greater use of health and social services (Ferrucci, Guralnik, Pahor, Corti, & Havlik, 1997), poorer subjective well-being (George, 2010), and elevated mortality (Guralnik, LaCroix, Branch, Kasl, & Wallace, 1991). The phenomenon of global population aging (Kinsella & He, 2009; National Institute on Aging, 2007) heightens the importance of examining disability and strategies and interventions with the potential to reduce it, from an international perspective. Cross-national comparisons highlight contextual influences on disability including differences in societal and familial responses that can be important in identifying appropriate policy options and developing programs or interventions to address needs for assistance. In addition, international clinical trials of interventions for conditions prevalent among elderly people often include functional disability as a key outcome (Black et al., 2003; Carpenter et al., 2004). One challenge in cross-cultural and cross-national research, however, is demonstrating the equivalence of measures that are the focus of comparisons and conclusions regarding the efficacy and effectiveness of interventions or policies (McHorney & Fleishman, 2006). Measurement equivalence across countries has been examined for measures such as depression and well-being (Ploubidis & Grundy, 2009), self-rated general health (Jürges, 2007), and work disability (Kapteyn, Smith, & Van Soest, 2007).

Measures that assess ability and limitations in self-care and independent living activities are widely used in evaluating disability prevalence and trends (Freedman, Martin, & Schoeni, 2002; Freedman, Martin, Cornman, Agree, & Schoeni, 2009). Three large population-based surveys of aging in the United States (the Health and Retirement Survey or HRS), the United Kingdom (the English Longitudinal Study of Aging or ELSA), and Europe (the 12-country Survey of Health, Ageing and Retirement in Europe or SHARE) incorporate such measures in the form of identically worded questions about difficulty in doing routine daily activities. Although attention to question wording is critical, identical wording alone does not guarantee that no measurement bias exists in cross-survey comparisons, particularly comparisons across countries or cultures.

Differential item functioning (DIF) is the broad term used in measurement theory to indicate items that demonstrate

Received December 1, 2010; Accepted October 18, 2011
Decision Editor: Merril Silverstein, PhD

differences in response across groups whose members have the same underlying abilities (or levels of a trait or condition being measured; Hambleton, Swaminathan, & Rogers, 1991; Holland & Wainer, 1993 are classic texts on DIF and item response theory [IRT]). There are numerous sources of potential DIF ranging from demographic characteristics such as age and gender (Fleishman, Spector, & Altman, 2002; McHorney & Fleishman, 2006; Perkins, Stump, Monahan, & McHorney, 2006; Teresi, Cross, & Golden, 1989) to attributes of the survey process such as social desirability or interview mode (Chan, Orlando, Ghosh-Dastidar, Duan, & Sherbourne, 2004; Johnson & van de Vijver, 2003). DIF in functional disability measures has been observed. Fleishman and colleagues (2002) found age and gender DIF for activities of daily living (ADL) and instrumental activities of daily living (IADL) using data from elderly and nonelderly adult respondents. Among items examined, shopping and managing money (both IADL) were observed to have the largest response bias. LaPlante (2010) also identified DIF by age in items measuring receipt of help in ADL and IADL. The overall impact of the observed age DIF was stronger for help with ADL items alone than for a score using help with both ADL and IADL items.

Multiple strategies are employed in the design and conduct of surveys to reduce or eliminate measurement error, but the extent of success can be difficult to determine. This can be especially challenging in cross-national research where despite the use of identical questions, numerous cultural differences may influence responses. Regression modeling of sociodemographic variables, although it does adjust for differential distribution of these factors across countries or regions, does not specifically address these response biases. To the extent that disability comparisons between countries reflect differences in how groups respond to questions about disability as opposed to real differences in disability level, comparisons will be flawed.

One approach used in a number of recent studies to identify and correct for DIF in cross-national comparisons is anchoring vignettes (King & Wand, 2007; King, Murray, Salomon, & Tandon, 2004). Vignettes are often used for questions with multiple ordinal responses that are structured as Likert scales and have been used to assess DIF and adjust for it in cross-national comparisons of perceived work disability (Kapteyn et al., 2007) and self-rated health (Salomon, Tandon, & Murray, 2004). However, this approach requires that vignette assessment data for the target question be available or that a new survey effort is planned to collect this information, making it challenging to use for secondary data analyses. Furthermore, this method focuses more on response category use in individual items and generally requires multiple vignette assessments for accurate anchoring of each target item (King et al., 2004). These characteristics make vignettes anchoring less feasible as a strategy for addressing DIF in multi-item scales. Confirmatory Factor Analysis (CFA) and IRT methods can also be used to identify measurement nonequivalence. These two methods share conceptual parallels, including the loadings in CFA, which are comparable to discrimination parameters in IRT. Unlike vignettes, both methods can make use of existing data and are readily applied to multi-item scales. However, IRT appears to be superior to CFA, particularly linear CFA models, for identifying item location DIF (Meade & Lautenschlager, 2004; Reise, Widaman, & Pugh, 1993). IRT also offers useful graphical detail on identified DIF at the item and scale level, and identified DIF are readily modeled to produce DIF-adjusted scores. Overall, IRT offers the best combination of feasibility, validity, and applicability for investigating and addressing DIF in our study.

Demonstrating that it is valid to pool data across surveys on disability will greatly expand opportunities for cross-country and cross-cultural research. This study uses IRT methods to examine DIF in individual measures of difficulty in routine daily activities and their aggregate effect for the 11-item scale using data from three major ongoing international surveys, the HRS, ELSA, and SHARE. The overall objective is to evaluate the need for DIF adjustment in comparative studies of disability using these surveys.

Method

Data

Data are drawn from three international surveys of aging: the HRS, the ELSA, and the SHARE. Details concerning the survey design can be found on the websites of each survey (http://hrsonline.isr.umich.edu/ for the HRS; http://www.esds.ac.uk/longitudinal/access/elsa/l5050.asp for ELSA; http://www.share-project.org/ for SHARE which includes 12 countries in Europe). SHARE data are for calendar year 2004; ELSA data are from March 2002 to March 2003 (Wave 1; Release 2); HRS data are from 2002. These years were selected to maximize the number of items common across surveys, including measures of cognitive functioning (which are being used in other analyses and will be reported elsewhere).

Our study was restricted to respondents who were aged 65 years or older. A small number of persons (4 in HRS, 106 in ELSA, and 60 in SHARE) with missing information across all ADL and IADL items were excluded. To minimize nonresponse bias, we used proxy responses if they were available to retain respondents in our analysis who were likely to be the most disabled and/or cognitively impaired. Proxy respondents varied by survey representing 13.2% of HRS respondents, 9.6% of SHARE respondents (proxy only or proxy and self-report), and <1% in ELSA (more recently ELSA has allowed for increased use of proxy respondents; Weir, Faul, & Langa, 2011). We were able to include proxy responses because the same questions on ADL/IADL difficulty were asked of proxy and self-respondents. Although differences between proxy and

self-reported information have been demonstrated, overall agreement between self- and proxy reports appears satisfactory for functional status measures (Epstein, Hall, Tognetti, Son, & Conant, 1989), particularly ADL (Østbye, Tyas, McDowell, & Koval, 1997). Final sample sizes for analyses were as follows: 10,905 in HRS, 5,437 in ELSA, and 13,408 in SHARE.

Measures

ADL and IADL disability.—All three surveys asked whether, "because of physical, mental, emotional, or memory problems," the sample person had "any difficulty" (yes/no) with ADL. Respondents were asked to exclude any difficulties expected to last less than 3 months. ADL were as follows: (a) dressing (including putting on shoes and socks), (b) eating (such as cutting up your food), (c) using the toilet (including getting up and down), (d) bathing and showering, (e) getting in and out of bed, and (f) walking across a room. IADL were as follows: (a) preparing a hot meal, (b) shopping for groceries, (c) making telephone calls, (d) taking medications, and (e) managing your money, such as paying your bills and keeping track of expenses. A scale ranging from 0 to 11 (number of items with reported difficulty) was constructed. Some studies suggest a composite ADL/IADL scale can be considered to represent a single underlying dimension of disability (LaPlante, 2010; Spector & Fleishman, 1998).

Demographic characteristics.—Age, gender, and education are used in comparisons of mean disability scores across surveys to assess the impact of adjusting for DIF in generating scores. A dichotomous variable indicating "secondary/high school or less" or "beyond secondary/high school" was created from variables provided in each survey. SHARE used the 1997 International Standard Classification of Education ISCED-97 (Classifying Educational Programmes: Manual for ISCED-97 Implementation in Organisation for Economic Co-operation and Development Countries; 1999 edition) to implement a standard coding with six levels ranging from preprimary through second stage tertiary education across all countries. The HRS provides items on completed education and degrees. ELSA provides a 7-level categorical variable; 500 individuals who were classified as "foreign/other" were coded as missing.

Analysis

As noted earlier, DIF as a source of measurement error in surveys is a long-standing concern. Under IRT, DIF assessment focuses on the relationship of an item to the trait assessed. DIF can be due to a difference in item discrimination (denoted in the literature as the *a* parameter) and/or item location (denoted in the literature as the *b* parameter or parameters). Item discrimination DIF reflects differences in the strength of the relationship between the item and the trait, with the item having a stronger relationship with the trait in one group than the other. Item location DIF, on the other hand, suggests that the item is "easier," or more likely to be endorsed at a lower level of the trait, for one group than the other. To identify DIF, discrimination and location parameters are estimated using an IRT model; differences in these parameters between two groups are tested statistically and examined in terms of the magnitude and nature of the difference.

IRT model.—We used the 2-parameter logistic model implemented with the computer program Multilog (Thissen, Chen, & Bock, 2002) to estimate one discrimination (a) and one location (b) parameter for each ADL/IADL item. The *a* parameter reflects the ability of an item to discriminate between levels of functioning, with higher *a* values indicating better discrimination. The *b* parameter refers to the location on the underlying trait or dimension (in this case disability) where the probability of indicating functional difficulty relative to no difficulty is 50%. Using HRS as the reference group, we estimated separate *a* and *b* parameters for each survey (ELSA and SHARE), yielding four parameters for each item in freely estimated models for paired analyses.

IRT evaluation of item-level effect.—Likelihood ratio difference tests were used to test whether item parameters were significantly different by survey. We used an iterative process implemented using the computer program IRTLRDIF (Thissen, 2001) to identify anchor items that did not show DIF for each pairwise survey comparison. Once a set of anchor items was determined, we evaluated DIF for each nonanchor item. We first tested for a difference in the slope parameter. The value of $-2 \times$ log likelihood for the model that constrained the *a* parameter to be equal in both surveys was compared with the corresponding $-2 \times$ log likelihood for the model that specified a separate *a* parameter for each group (the *b* parameters were unconstrained in both models). If no difference by survey was observed in the *a* parameter, we continued testing for a difference in the location parameters by comparing the model that constrained both the *a* and the *b* parameters to be equal across surveys with the model that only constrained the *a* parameter. If the *a* parameter differed significantly across surveys, however, no test for a difference in the location parameters was performed, as the interpretation of tests for location differences is unclear in this situation (Thissen, Steinberg, & Wainer, 1993). Differences in $-2 \times$ log likelihood was evaluated using the chi-square distribution, with $p < .05$ indicating significant difference. The Benjamin–Hochberg method was used to adjust for multiple comparisons (Thissen, Steinberg, & Kuang, 2002).

To examine the nature of DIF more closely, we compared the item characteristic curves (ICC) for each item using parameters estimated from the models for the paired survey

analyses. These curves plot the probability of endorsing the item over the range of underlying disability. Differences in these curves for the two groups reveal the magnitude and direction of the DIF at the item level. Nonoverlapping ICCs by group indicate DIF; coincident curves reflect lack of DIF.

IRT evaluation of scale-level mode effect.—The aggregate effect of observed item DIF at the scale level is evaluated by comparing test characteristic curves and estimating the difference in scores for the two groups. For each group of respondents, the curve plots the expected score over the range of disability. The curves were compared across groups to illustrate the magnitude and direction of the overall DIF effect at the scale level. Given the number of countries participating in SHARE, we also investigated whether findings for SHARE at the survey level remained robust for major geographic regions within SHARE: Scandinavia (Denmark, Sweden), Mediterranean (Spain, Italy, Greece, Israel), and Central Europe (Austria, France, Germany, Switzerland, Belgium, The Netherlands). Each region was compared with HRS.

Evaluation of DIF-adjusted score distribution and implications of DIF for cross-survey comparisons.—The effect of DIF on disability scores was evaluated in two ways. First, to determine the effect on the distribution of scores, we plotted the percentage of respondents for (a) the original summed score, (b) an IRT score estimated using item parameters not adjusted for DIF, and (c) an IRT score estimated using item parameters adjusted for DIF. Both sets of IRT scores were rescaled via a linear transformation to a range of 0–11 to facilitate comparison with the original summed scores. Specifically, the IRT score corresponding to "0" ADL/IADL was subtracted from each person's IRT score and then multiplied by a factor to recalibrate the IRT value to the 0–11 score range (i.e., 11 divided by the IRT score range, the absolute value of difference in IRT scores that correspond to the original 0 and 11 scores). Respondents reporting no difficulty with any ADL or IADL (73.5%) and those reporting difficulty in all 11 ADL and IADL (0.9%) had the same score regardless of approach.

To plot the percentage of respondents at each score (from 0 to 11) for the two IRT scores, which are continuous in nature, a score window of −0.5 and +0.5 was used. For example, scores of between 0.5 and 1.5 after rescaling were set to a score of 1. We also plotted the distribution of rescaled DIF-adjusted and unadjusted scores for several sample original scores to illustrate the impact of IRT modeling on scores. For these graphs, score windows of −0.1 and +0.1 were used.

To assess the implications of DIF for cross-survey comparisons of disability, we examined mean disability scores by basic demographic characteristics and compared differences (using *t* tests) based on the original summed score with those based on a DIF-adjusted IRT score (rescaled via linear transformation). The objective was to assess whether conclusions regarding differences in mean disability levels would change when scores have been adjusted to achieve measurement equivalence.

Results

Sample Characteristics

The pooled sample (n = 29,750) spans a broad age range, with 32% between 65 and 69 years old and 23% who were aged 80 years or older (Table 1). Three quarters completed a secondary/high school education and 44% were men. Among ADL, difficulty in dressing and bathing had the highest prevalence and eating the lowest, overall and in each survey. For the IADL, difficulty in shopping had the highest prevalence overall and across surveys. Due to the large sample sizes, with only a few exceptions, significant differences in demographic characteristics and in ADL and IADL difficulty were observed across the three surveys ($p < .001$).

Item-Level DIF

Item-level DIF was evaluated in terms of statistical significance and, more importantly, whether it qualifies as meaningful. Statistically significant DIF at the item level was found for 8 of 11 items between HRS and SHARE. However, using a .1 difference in probability as the criterion for determining meaningful DIF (Perkins et al., 2006), only two items, walking and bathing, demonstrated DIF between HRS and SHARE. The maximum difference for the walking item occurred at about 1.5 *SD*s above the HRS group mean, with HRS respondents at this location about 20% more likely to endorse difficulty with walking than SHARE respondents. Although the maximum difference for bathing was also found near 1.5 *SD*s above the HRS group mean, the direction of the bias was in the opposite direction. For this item, the probability that HRS respondents would report difficulty was slightly less than 20% compared with SHARE respondents. The remaining comparisons between HRS and SHARE appeared negligible (Figure 1).

Between the HRS and the ELSA, of the eight items that demonstrated statistically significant DIF, six appeared meaningful using the .1 difference rule (Figure 2). Three ADL items, dressing, bathing, and transferring, and one IADL item, shopping, showed differences in item location. For these items, ELSA respondents at the same estimated level of functioning as HRS respondents were more likely to report difficulty with performing these tasks. The strongest effect for each item appeared to be within 2 *SD*s of the HRS group mean. Two IADL items, making phone calls and managing money, showed discrimination and location DIF. For these two items, HRS respondents at the same estimated level of functioning were more likely to report difficulty with these tasks than ELSA respondents. The DIF

Table 1. *Sample Characteristics (unweighted)*

Characteristic	Total	HRS	ELSA	SHARE	HRS vs. ELSA *p* Value	HRS vs. SHARE *p* Value	ELSA vs. SHARE *p* Value	Overall *p* value
Sample size	29,750	10,905	5,437	13,408				
Age group (%)								
65–69	31.6	30.3	31.2	32.9	.243	<.001	.021	<.001
70–74	25.7	23.7	26.8	27.0	<.001	<.001	.855	
75–79	19.4	18.6	19.8	20.0	.052	.004	.775	
80–85	13.9	15.1	14.4	12.7	.219	<.001	.003	
85+	9.3	12.5	7.8	7.4	<.001	<.001	.343	
Gender (%)								
Men	44.1	42.4	44.6	45.3	.008	<.001	.362	<.001
Education (%)								
Secondary/high school or less	76.7	66.0	74.9	86.1	<.001	<.001	<.001	<.001
Beyond secondary/high school	21.1	34.0	15.5	13.0	-			
ADL difficulty items (%)								
Dressing	12.7	13.4	16.9	10.4	<.001	<.001	<.001	<.001
Walking	6.6	10.9	4.7	3.8	<.001	<.001	.006	<.001
Bathing	11.9	11.8	17.9	9.6	<.001	<.001	<.001	<.001
Eating	3.9	5.7	2.4	3.0	<.001	<.001	.011	<.001
Getting into/out of bed	6.7	8.5	7.5	4.9	.022	<.001	<.001	<.001
Toileting	5.6	8.5	4.5	3.7	<.001	<.001	.008	<.001
IADL difficulty items (%)								
Using telephone	5.2	8.3	2.7	3.7	<.001	<.001	.001	<.001
Taking medication	4.2	6.2	1.9	3.6	<.001	<.001	<.001	<.001
Handling money	7.9	10.5	3.6	7.4	<.001	<.001	<.001	<.001
Shopping	12.3	14.0	13.2	10.5	.137	<.001	<.001	<.001
Preparing meals	8.2	10.9	5.9	6.8	<.001	<.001	.020	<.001

Notes. ADL = activities of daily living; ELSA = English Longitudinal Study of Aging; HRS = Health and Retirement Survey; IADL = instrumental activities of daily living; SHARE = Survey of Health, Ageing and Retirement in Europe.

effect for these two items was strong, with maximum difference in probability of 40% for difficulty making phone calls and more than 30% for difficulty managing money; the effect was also broad, affecting responses in the range up to approximately 5 *SD*s above the HRS group mean.

Scale-Level DIF

For the summary scale, the difference in expected scores between HRS and SHARE that is attributable to DIF is small. This result was expected given the small item-level differences, and for the two items with meaningful DIF, the fact that the direction of the DIF for each was offsetting. As shown in Figure 3, the maximum score difference observed between HRS and SHARE was 0.36 (for the range from 0 to 11). Otherwise, differences were largely negligible (<0.1). The only differences of note—0.36 and 0.32—were observed between theta scores 1.0 and 2.0, which is equivalent to 1 and 2 *SD*s of the HRS group mean. Similarly, we found negligible differences between HRS and each of the three SHARE regions at the item and scale level. Details on these regional findings are available upon request from the authors.

The difference in expected scores due to DIF between HRS and ELSA is substantially larger and affects a wider spectrum of scores. Differences in expected scores range in size from −1.55 to 0.53 (for the 0–11 score; Figure 3). HRS scores were lower than those for ELSA between theta scores of −1.0 and 2.0, with the largest differences observed between 0 and 1.5. HRS scores were moderately higher than ELSA at theta scores above 2.0, particularly between 2.0 and 3.0.

Figure 4 illustrates the differences between the original summed score and the IRT-based rescaled scores, unadjusted and adjusted for DIF in the pooled sample. The original score distribution suggests a less disabled population than either of the IRT score distributions. Based on the original scores, nearly 15% of the respondents had one or two ADL/IADL limitations, with declining percentages reporting four or more limitations. By contrast, both sets of IRT scores are shifted in the direction of increased numbers of limitations. However, the shift is greater at the lower end of the scale than the upper end. These findings suggest that the original scoring method categorizes more individuals at the lower end of the scale (less disability) than does an IRT scoring method that takes into account characteristics such as item location and discrimination.

Figure 5 offers a more detailed illustration of the effects of IRT modeling and DIF adjustment using original scores of 1, 4, 7, and 10 ADL/IADL limitations. The IRT scores within each reveal variation that is missed with the original score. As already noted, at original scores of 1 or 4 (at the lower end of the scale), IRT scores generally indicate greater disability than do the original scores. DIF adjustment changes the distribution, however, by extending the range, especially at the lower end for these scores. For original scores of 7 and 10, DIF adjustment also extends the range somewhat

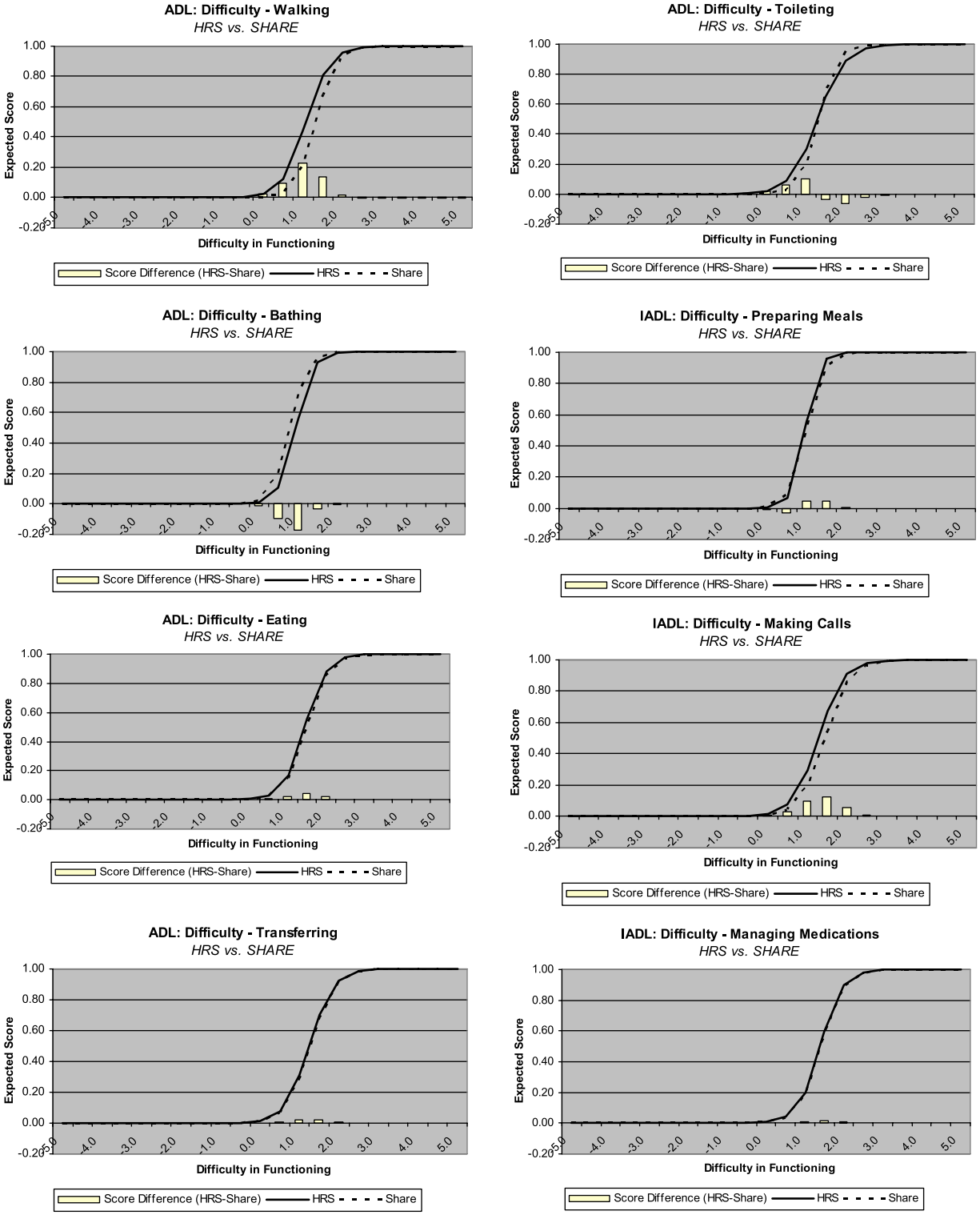


Figure 1. Item characteristic curves for Health and Retirement Survey (HRS) and Survey of Health, Ageing and Retirement in Europe (SHARE; statistically significant items).

in the direction of higher scores. These effects in the pooled sample are in the direction that would be expected after accounting for DIF between HRS and ELSA and likely reflect the score adjustments for these study populations.

Implications of DIF Adjustment for Cross-Survey Comparisons of Mean Disability Scores

Table 2 shows some simple demographic comparisons among the HRS, ELSA, and SHARE to demonstrate how

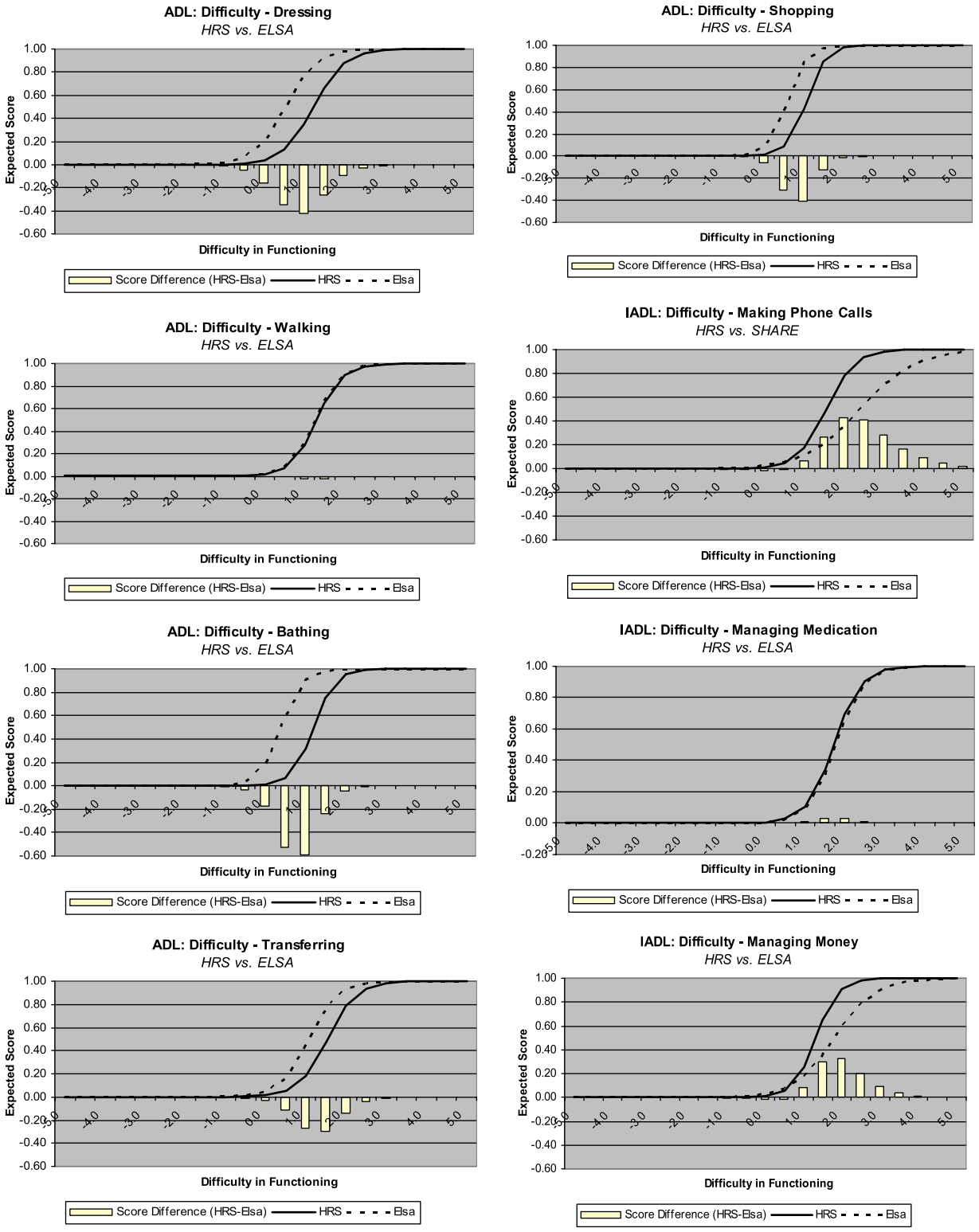


Figure 2. Item characteristic curves for Health and Retirement Survey (HRS) and English Longitudinal Study of Aging (ELSA; statistically significant items).

results may differ if DIF in reported difficulty with routine activities is taken into account. Overall, mean disability levels (as measured by reported difficulty) are significantly different between the HRS and the ELSA (higher in HRS), the HRS and the SHARE (higher in HRS), and between ELSA and SHARE (higher in ELSA), regardless of whether results are based on original scores or DIF-adjusted IRT scores. (For ease of comparison, the linear transformed

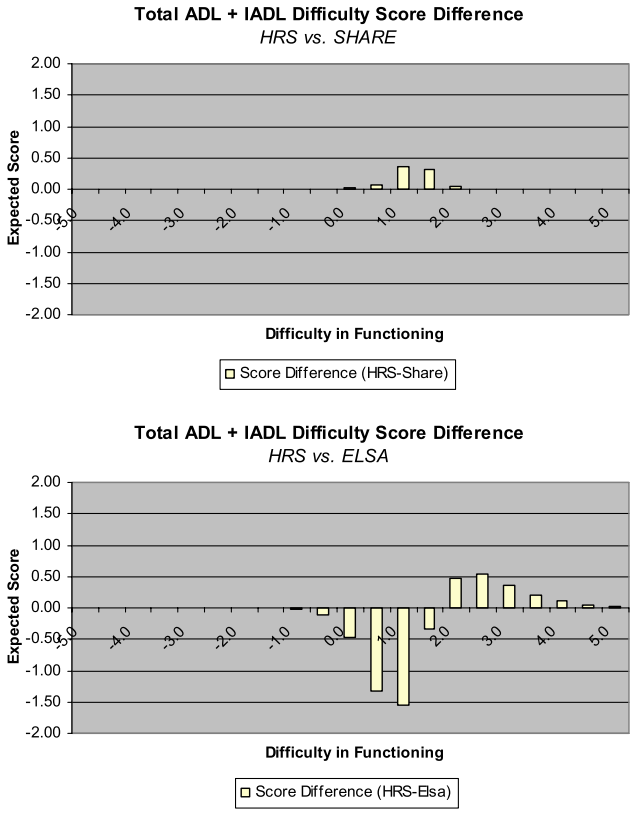
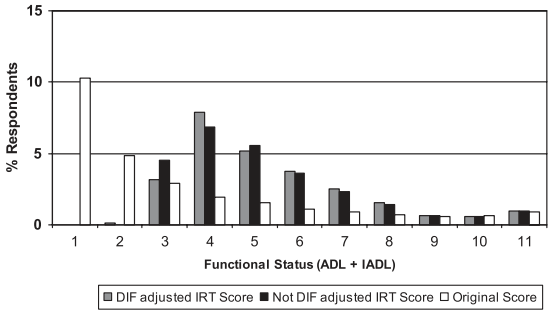


Figure 3. Overall impact of differential item functioning (DIF): Survey of Health, Ageing and Retirement in Europe (SHARE) and English Longitudinal Study of Aging (ELSA) compared with Health and Retirement Survey (HRS).



Note: Original scores are based on a count of ADL and IADL items with reported difficulty (0-11). Respondents with a score of 0 comprise 73.5% of the sample and their scores are unchanged across the three scoring approaches. They are excluded from the graph.

Figure 4. Activities of daily living (ADL) and instrumental activities of daily living (IADL) summary scores 1–11, differential item functioning (DIF)–adjusted item response theory (IRT), IRT not DIF adjusted, and original.

DIF-adjusted IRT scores are used; continuous with range 0–11.) The overall higher IRT scores reflect the shifts in distribution of scores discussed earlier.

Many other comparisons also hold regardless of whether comparisons are based on original scores or DIF-adjusted IRT scores. For some subgroup comparisons, however, results differ. Without DIF adjustment, mean differences between HRS and ELSA among 65- to 74-year olds and between ELSA and SHARE among 75- to 84-year olds are not detected. On the other hand, comparisons of mean original scores indicate a difference between men in the HRS and ELSA, whereas a comparison of DIF-adjusted mean scores shows no difference.

Looking at age within gender, DIF adjustment changes results for 75- to 84-year olds and does so among both men and women. Among men, DIF adjustment indicates a difference between ELSA and SHARE that is otherwise not detected. Among women, DIF adjustment indicates that there is no difference in this age group between women in HRS and women in ELSA.

DIF adjustment results in no changes in interpretation of disability differences within education groups (high school and some college) among surveys. All the changes that result from DIF adjustment involve comparisons with ELSA, consistent with the finding that no DIF was detected between HRS and SHARE.

Disability Differences Across Surveys Using DIF-Adjusted Results

The focus of this paper is on assessing measurement equivalence of disability measures across surveys and the implications of adjusting for DIF. Nonetheless, a few observations concerning international differences in mean disability levels by demographic characteristics drawing on the DIF-adjusted disability scores are useful.

Among 65- to 74-year olds, disability is highest in the ELSA followed by the HRS. At older ages, however, mean disability levels are highest in the HRS. Although disability levels in the ELSA exceed those in SHARE below age 85, mean disability is higher in SHARE among those aged 85 years and older.

Among men, disability levels do not differ between HRS and ELSA, but both have higher disability compared with men in SHARE. Disability among women differs across all survey comparisons, those in the HRS have the highest disability level, followed by ELSA, and then SHARE.

For both men and women below age 85, disability levels are no different between the HRS and ELSA (for women 65–74, the difference approaches significance, $p = .06$). For men younger than 85 years, there are significant differences between those in HRS and SHARE and those in ELSA and SHARE (SHARE has lower disability than either of the others). For women 65–74, as for men, those in HRS and ELSA have higher disability than women in SHARE. Among women 75–84, however, only women in HRS and SHARE have different disability levels (higher in HRS).

Among women aged 85 years or older, there are differences among all surveys with the highest disability in the HRS and the lowest in ELSA. Among men aged 85 years

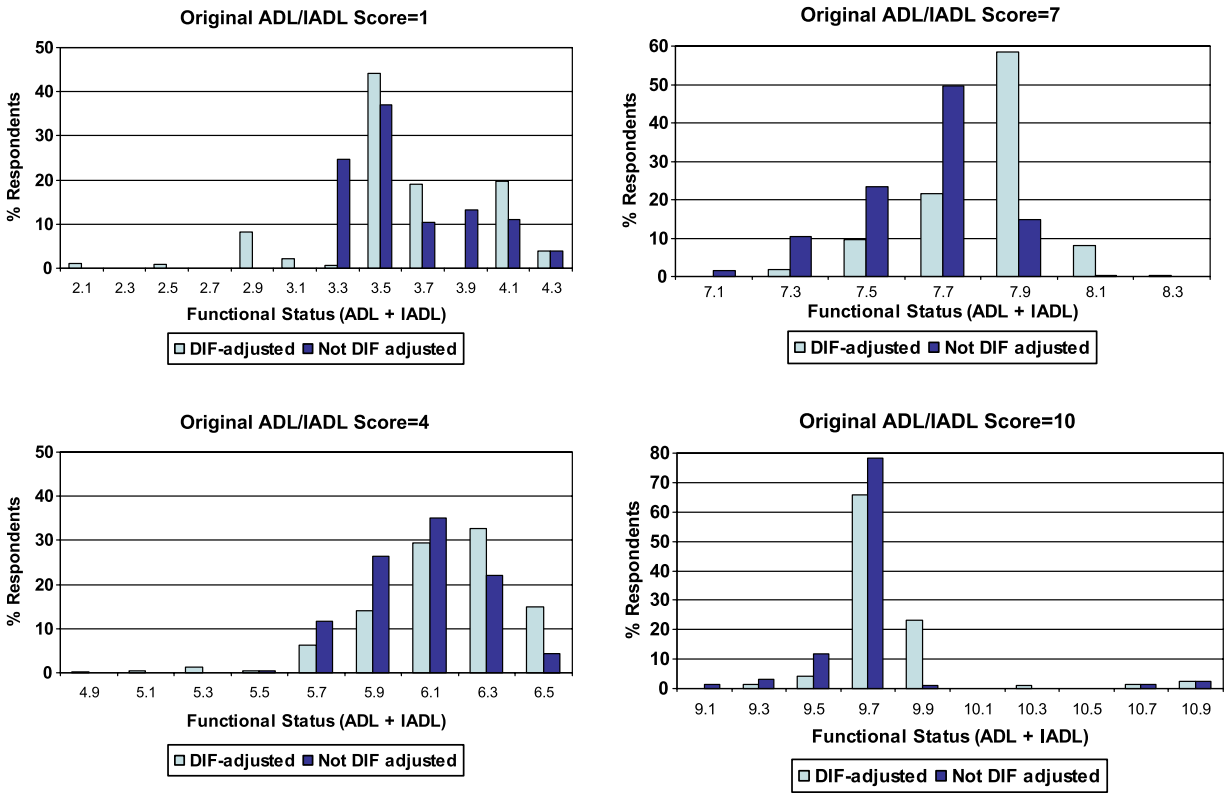


Figure 5. Distribution of rescaled item response theory (IRT) scores, both adjusted and unadjusted for four original summed activities of daily living (ADL)/instrumental activities of daily living (IADL) scores.

and older, the only difference is between men in the HRS and in ELSA and disability is higher in the HRS.

Discussion

Measurement equivalence is often assumed when questions are identically phrased, but considerable evidence suggests the importance of testing for it, particularly when cross-cultural comparisons are involved. Our study undertook examination of DIF for a measure of disability, defined as reported difficulty with ADL and IADL, across major cross-national surveys of aging. The spread of international surveys of aging presents new and promising opportunities for cross-national comparisons of the aging process and the effect of differing national and policy contexts. Despite the use of standardized question wording, investigation of cross-survey measurement equivalence for key outcomes remains a critical step in optimizing the value of these surveys to support international comparative studies.

Our study found equivalence for 11 ADL and IADL difficulty items (individually and as a scale) between the HRS and SHARE. This finding suggests that it would be appropriate to make comparisons regarding disability prevalence and levels between these surveys. We also found six items with substantial DIF between HRS and ELSA, however. For four items—dressing, bathing, transferring, and shopping—ELSA respondents at the same level of disability as HRS respondents were more likely to endorse difficulty with these tasks. Two other items—making phone calls and managing money—that demonstrated significant DIF were more discriminating for HRS than ELSA respondents, and in addition, ELSA respondents at the same level of disability as HRS respondents were less likely to endorse difficulty with these tasks.

Using all items to form a scale, the overall effect of measurement nonequivalence between HRS and ELSA is mitigated, in part because when all items are used the direction of DIF is offsetting (four items with greater endorsement by HRS and two items with greater endorsement by ELSA). The difference between DIF-adjusted and unadjusted scores (IRT) was negligible over much of the observed score range. Observed DIF predominated at the upper end of the score range. ELSA overestimated difficulty (for four items) relative to the HRS within 2 *SD*s above the group mean (HRS) and underestimated difficulty (for two items) in the upper ranges (beyond 2 *SD*s).

Two previous studies suggest DIF may be more of a concern for ADL measures and for scales based on these measures. One early study of cross-national DIF in measures of functioning (Teresi et al., 1989) compared probability samples of elderly people living in long-term care institutions in

Table 2. *Comparisons Between Surveys of Original and DIF-Adjusted IRT Scores Reflecting Mean Disability*

	Original			DIF adjusted			*N*		
	HRS	ELSA	SHARE	HRS	ELSA	SHARE	HRS	ELSA	SHARE
Total	1.09[a,b]	.81[a,c]	.68[b,c]	1.70[a,b]	1.46[a,c]	1.17[b,c]	10,905	5,437	13,408
Age									
65–74	0.55[b]	0.59[c]	0.31[b,c]	0.99[a,b]	1.10[a,c]	0.63[b,c]	5,879	3,152	8,022
75–84	1.17[a,b]	0.96[a]	0.92[b]	1.90[a,b]	1.75[a,c]	1.59[b,c]	3,668	1,859	4,390
85+	3.17[a,b]	1.80[a,c]	2.55[b,c]	4.24[a,b]	2.96[a,c]	3.66[b,c]	1,358	426	996
Gender									
Men	0.85[a,b]	0.72[a,c]	0.53[b,c]	1.39[b]	1.30[c]	0.93[b,c]	4,620	2,423	6,070
Women	1.26[a,b]	0.89[a,c]	0.80[b,c]	1.93[a,b]	1.60[a,c]	1.36[b,c]	6,285	30,140	7,338
Men									
65–74	0.52[b]	0.54[c]	0.28[b,c]	0.93[b]	1.01[c]	0.55[b,c]	2,678	1,462	3,805
75–84	1.00[b]	0.85	0.75[b]	1.65[b]	1.53[c]	1.33[b,c]	1,503	792	1,929
85+	2.37[a]	1.67[a]	2.07	3.29[a]	2.67[a]	2.90	439	171	336
Women									
65–74	0.57[b]	0.64[c]	0.34[b,c]	1.05[b]	1.17[c]	0.70[b,c]	3,201	1,692	4,217
75–84	1.29[a,b]	1.04[a]	1.06[b]	2.07[b]	1.91	1.79[b]	2,165	1,067	2,461
85+	3.55[a,b]	1.89[a,c]	2.80[b,c]	4.69[a,b]	3.16[a,c]	3.99[b,c]	919	255	660
Education[d]									
Secondary/high school or less	1.26[a,b]	0.90[a,c]	0.72[b,c]	1.95[a,b]	1.59[a,c]	1.24[b,c]	7,197	4,070	11,548
Beyond secondary/high school	0.74[a,b]	0.49[a,c]	0.38[b,c]	1.22[a,b]	0.96[a,c]	0.73[b,c]	3,705	843	1,744

Notes. ADL = activities of daily living; DIF = differential item functioning; ELSA = English Longitudinal Study of Aging; HRS = Health and Retirement Survey; IADL = instrumental activities of daily living; IRT = item response theory; SHARE = Survey of Health, Ageing and Retirement in Europe; Original scores = sum of 11 ADL and IADL items where difficulty was reported; DIF-adjusted IRT scores = linear transformation of DIF-adjusted IRT scores so that range = 0–11.

[a]HRS and ELSA different at $p \leq .05$.

[b]HRS and SHARE different at $p \leq .05$.

[c]ELSA and SHARE different at $p \leq .05$.

[d]Missing cases in Education due to Don't Know responses or inability to classify; 3 in HRS; 524 in ELSA (mostly in a "foreign education category" that could not be classified); 116 in SHARE.

New York City and London, England, and found DIF for items on bathing, eating, orientation, and stair climbing. LaPlante (2010) found age-related DIF for 8 of 14 items on receipt of help for ADL and IADL. Significant impact at the scale level was observed for a scale based on ADL items. When ADL and IADL items were combined, the age-related measurement bias was substantially reduced. In our study, DIF for three of six ADL items was found between the HRS and the ELSA. These results, and earlier research, suggest a comparison between persons in HRS and ELSA on ADL items alone or, as a scale, would likely be subject to meaningful measurement bias.

A comparison of the original 0–11 scale scores with scores generated by IRT methods suggests that intervals between scores in the original scale do not represent equivalent unit changes in disability. IRT methods spread out the score distribution in all three study populations—generating higher mean scores because the distribution shifted in the direction of greater disability particularly at the lower end of the scale (and somewhat more so with DIF adjustment). This result suggests that the threshold for scores of 1 or 2—reporting difficulty with one or two items—appears higher than that for reporting difficulty on additional items once several have been endorsed (e.g., reporting difficulty in five vs. six items). This is consistent with Torrence, Zhang, Feeny, Furlong, and Barr (1992, p. 38) who suggested that the "additional disutility added by a particular deficit is greater if it is the first and only deficit and less if it is the last of two or more deficits."

Results from comparisons of mean disability by basic demographic characteristics across surveys showed numerous differences in disability levels (regardless of score methodology) among the populations in the HRS, ELSA, and SHARE. As expected, total population comparisons between HRS and SHARE were unchanged whether based on original or DIF-adjusted IRT scores because no DIF was detected between HRS and SHARE. A few differences emerged in comparisons with ELSA: some differences between HRS and ELSA were no longer significant (men, women 75–84); some differences between ELSA and SHARE (age 75–84, men 75–84) and between ELSA and HRS (65–75) reached significance.

The greater disability we observed for participants in the HRS compared with those in ELSA are consistent with findings reported in other cross-country studies that found greater disease burden, with higher incidence, prevalence, and worse outcomes based on biological markers, in the United States compared with the United Kingdom (Banks, Marmot, Oldfield, & Smith, 2006; Banks, Muriel, & Smith, 2010). Furthermore, we found, similar to the health differences reported by Banks and colleagues (2006), that the disability differences between these two countries were evident at both higher and lower socioeconomic status levels (as measured by education).

We examined only a few demographic differences in mean disability for purposes of illustrating the potential impact of DIF in cross-national comparisons. The differences observed reflect both the score distribution effects of IRT methods (as noted earlier) and the DIF adjustment. Using a score generated by IRT methods has a stronger effect on scores at the lower end of the scale, thus potentially affecting more people than DIF adjustment, which in this case is focused at the upper end of the range where there are fewer people.

Our study also highlights a key challenge to the emerging body of research in this field. Similar to prior studies of DIF, we considered magnitude of DIF, in addition to statistical significance, when determining how meaningful observed differences were. However, there is little consensus at present regarding the size of the difference that constitutes meaningful DIF. We used the criterion of a 0.1 difference along a 0–1 probability scale that Perkins and colleagues (2006) used to define a meaningful difference in their study of DIF for the SF-36. McHorney and Fleishman (2006) have noted, however, that this rule, adapted from studies in educational testing, may be less applicable to patient-reported outcomes. Furthermore, they noted the ambiguity around how to determine the impact of the difference along the spectrum of the underlying scale. One strategy may be to examine the score distribution as well, as we have done, to gauge the specific impact of observed DIF in a population or sample of interest.

We focused our examination of DIF at the survey level given the practical value that generating comparable disability scores and pooling data across surveys could have for future research. Therefore, we did not perform a more extensive analysis of the potential contributors to the DIF we observed, such as variation in the ethnic composition of survey populations. We believe that this would be an important area for further investigation, however.

In summary, our results indicate measurement equivalence between HRS and SHARE on measures of ADL and IADL difficulty. Using DIF-adjusted scores for ELSA respondents would improve measurement equivalence. Furthermore, IRT methods provide a scoring methodology that better reflects the distribution of the population along the underlying trait of disability as measured by the 11 items examined here. The goal of this paper was to explore the extent of DIF in disability measures administered in these three major national surveys and the potential for DIF to affect cross-survey comparisons. Fielding a common set of items does not ensure measurement equivalence. Future efforts that involve pooling data for common measures across surveys should first examine these measures for the presence of DIF.

Funding

This research was supported by the National Institute on Aging (grant AG032502).

Acknowledgments

K. S. Chan helped to plan the study, conducted the statistical analyses, and wrote the paper; J. Kasper helped to plan the study and wrote the paper; J. Brandt helped to plan the study and contributed to revising the paper; L. Pezzin helped to plan the study and contributed to revising the paper.

Correspondence

Correspondence should be addressed to Kitty S. Chan, PhD, Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, 624 North Broadway, Room #633, Baltimore, MD 21205-1901. E-mail: kchan@jhsph.edu.

References

Banks, J., Marmot, M., Oldfield, Z., & Smith, J. P. (2006). Disease and disadvantage in the United States and in England. *Journal of the American Medical Association*, *295*, 2037–2045. doi:10.1001/jama.295.17.2037

Banks, J., Muriel, A., & Smith, J. P. (2010). Disease prevalence, disease incidence, and mortality in the United States and in England. *Demography*, *47*(Suppl.),S211–S231. doi:10.1353/dem.2010.0008

Black, S., Roman, G. C., Geldmacher, D. S., Salloway, S., Hecker, J., Burns, A., . . . Pratt, R. (2003). Efficacy and tolerability of donepezil in vascular dementia: Positive results of a 24-week, multicenter, international, randomized, placebo-controlled clinical trial. *Stroke; A Journal of Cerebral Circulation*, *34*, 2323–2330. doi:10.1161/01.STR.0000091396.95360.E1

Carpenter, I., Gambassi, G., Topinkova, E., Schroll, M., Finne-Soveri, H., Henrard, J. C., . . . Ljunggren, G. (2004). Community care in Europe. The aged in home care project (AdHOC). *Aging Clinical and Experimental Research*, *16*, 259–269.

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., & Sherbourne, C. D. (2004). Interview mode effect on the Center for Epidemiological Studies Depression (CES-D) scale: An item response theory analysis. *Medical Care*, *42*, 281–289. doi:10.1097/01.mlr.0000115632.78486.1f

Epstein, A. M., Hall, J. A., Tognetti, J., Son, L. H., & Conant, L. (1989). Using proxies to evaluate quality of life: Can they provide valid information about patient's health status and satisfaction with medical care? *Medical Care*, *27*, S91–S98. doi:10.1097/00005650-198903001-00008

Ferrucci, L., Guralnik, J. M., Pahor, M., Corti, M. C., & Havlik, R. J. (1997). Hospital diagnoses, medicare charges, and nursing home admissions in the year when older persons become severely disabled. *Journal of the American Medical Association*, *277*, 728–734. doi:10.1001/jama.277.9.728

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *57*, 275–284. doi:10.1093/geronb/57.5.S275

Freedman, V. A., Martin, L. G., Cornman, J., Agree, E. M., & Schoeni, R. F. (2009). Trends in assistance with daily activities: Racial/ethnic and socioeconomic disparities persist in the US older population. In D. A. Cutler, & D. A. Wise (Eds.), *Health in older ages: The causes and consequences of declining disability among the elderly* (pp. 411–438). Chicago, IL: University of Chicago Press.

Freedman, V. A., Martin, L. G., & Schoeni, R. F. (2002). Recent trends in disability and functioning among older adults in the United States: A systematic review. *Journal of the American Medical Association*, *288*, 3137–3146. doi:10.1001/jama.288.24.3137

George, L. K. (2010). Still happy after all these years: Research frontiers on subjective well-being in later life. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *65*, 331–339. doi:10.1093/geronb/gbq006

Guralnik, J. M., LaCroix, A. Z., Branch, L. G., Kasl, S. V., & Wallace, R. B. (1991). Morbidity and disability in older persons in the years prior to death. *American Journal of Public Health*, *81*, 443–447. doi:10.2105/AJPH.81.4.443

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Johnson, T. P., & van de Vijver, F. J. R. (2003). Social desirability in cross-cultural research. In J. A. Harkness, van de Vijver, F. J. R., & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 195–204). New York, NY: Wiley.

Jürges, H. (2007). True health vs. response styles: Exploring cross-country differences in self-reported health. *Health Economics*, *16*, 163–178. doi:10.1002/hec.1134

Kapteyn, A., Smith, J. P., & Van Soest, A. (2007). Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review*, *97*, 461–473. doi:10.1257/aer.97.1.461

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of survey research. *American Political Science Review*, *98*, 191–207. doi:10.1017/S0003055404001082X

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, *15*, 46–66. doi:10.1093/pan/mpl011

Kinsella, K., & He, W. (2009). *An aging world: 2008*. U.S. Census Bureau, International Population Reports, P95/09–1. Washington, DC: Government Printing Office.

LaPlante, M. P. (2010). The classic measure of disability in activities of daily living is biased by age but an expanded IADL/ADL measure is not. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *65*, 720–732. doi:10.1093/geronb/gbp129

McHorney, C. A., & Fleishman, J. A. (2006). Assessing and understanding measurement equivalence in health outcome measures issues for further quantitative and qualitative inquiry. *Medical Care*, *44*(Suppl. 3), S205–S210. doi:10.1097/01.mlr.0000245451.67862.57

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361–388. doi:10.1177/1094428104268027

National Institute on Aging (2007). *Why population aging matters: A global perspective*. NIA Publication No. 07–6134. Bethesda, MD: National Institute on Aging.

Østbye, T., Tyas, S., McDowell, I., & Koval, J. (1997). Reported activities of daily living: Agreement between elderly subjects with and without dementia and their caregivers. *Age and Ageing*, *26*, 99–106. doi:10.1093/ageing/26.2.99

Perkins, A. J., Stump, T. E., Monahan, P. O., & McHorney, C. A. (2006). Assessment of differential item functioning for demographic comparisons in the MOS SF-36 health survey. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *15*, 331–348.

Ploubidis, G. B., & Grundy, E. (2009). Later-life mental health in Europe: A country-level comparison. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *64*, 666–676. doi:10.1093/geronb/gbp026

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552–566. doi:10.1037/0033-2909.114.3.552

Salomon, J. A., Tandon, A., & Murray, C. J. L. (2004). Comparability of self rated health: Cross-sectional multi-country survey study using anchoring vignettes. *British Medical Journal*, *328*, 258–261. doi:10.1136/bmj.37963.691632.44

Spector, W. D., & Fleishman, J. A. (1998). Combining activities of daily living with instrumental activities of daily living to measure functional disability. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *53*, 46–57. doi:10.1093/geronb/53B.1.S46

Teresi, J. A., Cross, P. S., & Golden, R. R. (1989). Some applications of latent trait analysis to the measurement of ADL. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *44*, 196–204.

Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the computation of the statistic involved in item response theory likelihood-ratio tests for differential item functioning*. University of North Carolina at Chapel Hill.

Thissen, D., Chen, W., & Bock, D. (2002). *Multilog (computer program). version 7.0 for windows*. Chicago, IL: Scientific Software.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamin-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*, 77–83. doi:10.3102/10769986027001077

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 67–114). Hillsdale, NJ: Lawrence Erlbaum Associates.

Torrence, G. W., Zhang, Y., Feeny, D., Furlong, W., & Barr, R. (1992). *Multi-attribute preference functions for a comprehensive health status classifications system*. Ontario, Canada: Centre for Health Economics and Policy Analysis, McMaster University.

Weir, D., Faul, J., & Langa, K. (2011). Proxy interviews and bias in the distribution of cognitive abilities due to non-response in longitudinal studies: a comparison of HRS and ELSA. *Longitudinal and Life Course Studies*, *2*, 170–184.