Journal of Gerontology: SOCIAL SCIENCES
2002, Vol. 57B, No. 5, S275–S284

Impact of Differential Item Functioning on Age and Gender Differences in Functional Disability

John A. Fleishman,[1] William D. Spector,[1] and Barbara M. Altman[2]

[1]Agency for Healthcare Research and Quality, Rockville, Maryland.
[2]National Center for Health Statistics, Hyattsville, Maryland.

Objectives. Estimates of group differences in functional disability may be biased if items exhibit differential item functioning (DIF). For a given item, DIF exists if persons in different groups do not have the same probability of responding, given their level of disability. This study examines the extent to which DIF affects estimates of age and gender group differences in disability severity among adults with some functional disability.

Methods. Data came from the 1994/1995 National Health Interview Survey Disability Supplement. Analyses focused on 5,750 adult respondents who received help or supervision with at least one of 11 activities of daily living/instrumental activities of daily living tasks. We estimated gender and age group (18–39, 40–69, and 70+) differences in disability, using multiple-indicator/multiple-cause models, which treat functional disability as a latent trait.

Results. Nine items manifested significant DIF by age or gender; DIF was especially large for "shopping" and "money management." Without adjusting for DIF, middle-aged persons were less disabled than elderly men, and women were less disabled than men among nonelderly persons. After adjusting for DIF, middle-aged persons did not differ from elderly persons, and gender differences within age groups were not significant.

Discussion. Comparisons of disability across sociodemographic groups need to take DIF into account. Future research should examine the causes of DIF and develop alternative question wordings that reduce DIF effects.

FUNCTIONAL disability refers to limitations in the performance of basic daily activities necessary to maintain personal hygiene or reside independently in the community. Functional disability includes limitations in activities of daily living (ADLs)—such as bathing, dressing, and eating—and instrumental activities of daily living (IADLs)—such as shopping, meal preparation, and managing finances (Katz, Ford, Moskowitz, Jackson, & Jaffe, 1963; Lawton & Brody, 1969). Accurate measurement of functional disability is important at both the population and individual levels. Knowledge of the prevalence and incidence of functional disability in the population is essential for anticipating demand for services and for program planning. At the individual level, functional disability is assessed to determine eligibility for participation in long-term care programs, and to assist in discharge and care planning.

In addition to information on the presence or absence of any functional disability, assessment of the *severity* of disability among the disabled serves important purposes. Measurement of the severity of functional disability is especially important in monitoring changes in health status for persons with chronic illnesses, whose progress may be measured in partial recovery of function instead of complete absence of limitations. Aggregate severity information can also be used for program planning and group comparisons.

Although functional disability is often measured by asking about difficulty performing ADL and IADL tasks, a tradition in this literature, beginning with Katz and colleagues (1963), distinguishes between dependence and independence. Operationally, in this tradition, disability is assessed by asking whether a person receives human or mechanical help with these tasks (Spector & Fleishman, 1998). The present study falls in this tradition and measures the severity of functional disability in terms of receipt of human help or supervision with ADL and IADL tasks.

ADL and IADL assessment instruments were originally designed for elderly or chronically ill adults (Katz et al., 1963). Some national surveys, however, ask the same questions about functional disability of both elderly and nonelderly adults. Findings from the 1994 National Health Interview Survey–Disability Supplement (NHIS-D) indicate that functional disability rates increase in older age groups. The proportion of people receiving help with ADLs or IADLs is about 1% for persons aged 18–29, 10% for persons aged 70–74, and 80% for those aged 95 and over (Spector, Fleishman, Pezzin, & Spillman, 2000). Data from this study also show that adult women of all ages are more likely than adult men to receive help with ADLs or IADLs: 12.8% versus 2.4% for persons under age 65 and 20% versus 12% for those 65 and older.

Age and gender comparisons of the *severity* of disability show a slightly different pattern. Spector and Fleishman (1998) assessed severity of functional disability by measuring the number of IADL and ADL items for which a person received human help. Among an elderly population with some disability, they found that people aged 80 and older reported more severe disability than those who were younger. Elderly men also reported more severe disability than elderly women, even though elderly women were more likely to be disabled.

Despite the widespread use of ADL and IADL questions across the age spectrum, the measurement properties of

these scales have not been established. Valid comparison of disability severity across age or gender groups requires that the measure be comparable in these groups. The validity of any comparison of functional disability across age or gender groups (or groups differing in other characteristics) rests on the assumption that the measure is *invariant* across these different groups. If extraneous factors influence people to respond differently in one group than in another, then the resulting lack of invariance confounds group comparisons. In other words, for valid group comparisons, measures should not be affected by differential item functioning (DIF; Camilli & Shepherd, 1994; Millsap & Everson, 1993; Thissen, Steinberg, & Wainer, 1993).

In terms of measurement theory, responses to survey questions are considered to be observed indicators of an unobserved latent variable. In the present context, measured responses to questions about receipt of help with ADLs and IADLs are viewed as reflecting a latent factor of functional disability. This factor is not observed directly, but indirectly through its effect on observed responses to ADL or IADL questions. If individuals at the same level of underlying disability differ in their responses to a specific item—depending on their age, gender, or other characteristics—then the item exhibits DIF. (The term "item bias" is often used, as well. DIF is a more neutral term referring to the existence of differential response patterns in different groups, whereas item bias specifically implies invidious group comparisons resulting from DIF [Camilli & Shepherd, 1994; Marshall, Mungas, Weldon, Reed, & Haan, 1997].) The mere existence of a group difference in observed disability is not sufficient to demonstrate DIF, which reflects group differences in responses to an item *after* respondents' status on the latent factor is controlled (Millsap & Everson, 1993). If DIF is present, then observed group differences will reflect something other than the latent factor. In the present case, if DIF is present, observed differences between age groups or genders in responses to ADL and IADL items will be a combination of real differences in latent disability and group-specific effects. It is, therefore, important to be able to separate true differences from measurement differences when comparing groups.

This study examines the extent to which comparisons of the severity of functional disability across age and gender groups are affected by DIF. Prior literature raises the possibility of DIF by gender and age in IADL and ADL items. Concerns about the comparability of IADL questions for men and women have existed since the 1960s, when functional disability measures were first being developed (Lawton & Brody, 1969). Tasks like laundry or meal preparation, especially in cohorts of elderly persons, may represent gender-typed activities that men may not normally do, and thus men may be more likely to receive help when they attempt them (Allen, Mor, Raveis, & Houts, 1993). Although evidence of gender DIF for functional disability items has been found for elderly persons (Spector & Fleishman, 1998), gender DIF has not been studied among nonelderly persons. In an attempt to reduce potential gender bias when assessing functional disability, surveys often ask whether nonperformance or help is caused by a "health problem or a disability or physical, mental, or emotional problems" (Spector et al., 2000). It is not known, however, to what extent this strategy is successful in eliminating DIF.

In addition to gender DIF, age-related DIF may occur. Nonelderly persons may not respond in the same way to certain ADL/IADL questions as elderly persons. Elderly and nonelderly respondents may have different perceptions of disability, different levels of support, varying role expectations, or varying coping styles that may result in a reluctance either to report receiving assistance or to seek assistance in the first place (Groot, 2000). If these tendencies are particularly strong for a subset of items, DIF may result. Physical or mental impairments that are more prevalent in certain age groups and affect only a few items may also produce DIF.

In this study, we compare three age groups—18–49, 50–69, and 70 and over—and two gender groups to assess DIF in commonly used ADL and IADL items. To gauge the potential impact of DIF, we estimate age and gender differences in underlying disability levels controlling for DIF and not controlling for DIF. We identify items that have particularly large DIF effects and compare results from models with and without these items to test the stability of our results.

Methods

Data

Data come from the 1994 and 1995 NHIS-D. The NHIS is a nationally representative household survey of the civilian, noninstitutionalized U.S. population. The 1994 and 1995 surveys included a special module of questions pertaining to disability and program participation; these questions (Phase 1) were used to screen respondents for a second phase of interviews. The analyses in this study use data only from 1994 and 1995 Disability Supplement Phase 1 interviews; they do not include data from the Phase 2 interviews. We limited the sample to adults because IADL questions are not applicable to children, and IADL questions in the NHIS-D were not asked of children under age 18.

Functional Disability Measures

Respondents were asked a series of questions about ADLs and IADLs. ADLs included bathing, dressing, eating, using the toilet, getting in or out of bed or chairs, and getting around inside the home; IADLs included preparing meals, shopping, managing money, using the telephone, doing heavy housework, and doing light housework. Respondents were asked whether they get help from another person for each ADL because of a physical, mental, or emotional problem; a separate question asked whether the person needs to be reminded or needs to have someone close by when performing each task. For each of the six ADLs, we created a dichotomous variable, which had a value of one if the person answered yes to either of these questions. For each of the six IADLs, respondents aged 18 or older were asked whether they get help or supervision from another person because of a physical, mental, or emotional problem; we created dichotomous variables that had values of one if the person received help or supervision performing each task. Use of equipment to perform any task, in the absence of human help or supervision, was not counted as being disabled.

Multiple-Indicator/Multiple-Cause (MIMIC) Models

Initial analyses were conducted on the 10,371 adult NHIS-D respondents who received help or supervision with at least one of the 12 ADL/IADL tasks. The remaining 134,440 (93%) respondents, persons who answered "no" to all items, have no variation in their responses and consequently contribute no information regarding item performance. More important, because the focus of the analyses is on measuring the severity of disability among persons with any disability, including the large number of respondents with no disabilities in the analyses would obscure systematic variation among the disabled.

Standard DIF assessment procedures assume that all items measure a single underlying latent trait. Recommended practice is to demonstrate the existence of a single dominant dimension (Hambleton, Swaminathan, & Rogers, 1991). One indication of a single dominant dimension is if the magnitude of the first eigenvalue of the item correlation matrix is large, relative to the second eigenvalue (Lord, 1980). We examined eigenvalues of the interitem correlation matrix based on the 10,371 adult NHIS-D respondents who received help or supervision with at least one of the ADL/IADL tasks. Because the items were dichotomous, tetrachoric correlations were computed.

There are several approaches to measuring DIF. We used a MIMIC latent variable model. MIMIC models have been used previously to investigate DIF in depression screening scales (Gallo, Anthony, & Muthen, 1994; Grayson, Mackinnon, Jorm, Creasey, & Broe, 2000). The MIMIC model postulates that a latent factor gives rise to associations among several observed indicators. In the present case, IADL and ADL items are observed indicators assumed to measure an unobserved functional disability factor. In addition, the MIMIC model extends the standard factor analysis model by including observed exogenous variables that affect the latent factor. In the present instance, latent disability is regressed on age and gender. Finally, the model includes direct effects from the exogenous age and gender variables to the indicators. Thus, the MIMIC model distinguishes two ways in which group differences in ADL/IADL may manifest themselves. First, age or gender groups may be more or less disabled, which in turn affects responses to the observed indicators. Second, age or gender groups may differ in their responses to particular items, over and above any differences in disability. Such direct effects correspond to DIF; they represent systematic differences in item responses controlling for the latent factor.

Many authors have used Item Response Theory (IRT) models to investigate DIF (e.g., Morales, Reise, & Hays, 2000; Teresi, Kleinman, & Ocepek-Welikson, 2000a). We decided to estimate MIMIC models instead for several reasons. First, it has been shown that a dichotomous factor analysis model (without exogenous covariates) is equivalent to and a reparameterization of the standard two-parameter IRT model (McDonald, 1999; Muthen & Lehman, 1985; Takane & de Leeuw, 1987). Second, procedures for testing DIF in an IRT framework become cumbersome when there are more than two groups. The MIMIC model has the advantage that multiple exogenous variables can be included simultaneously.

In the measurement part of the MIMIC model, the dichotomous ADL/IADL items were specified as indicators of a single latent disability factor. To identify the model, the loading of the toileting item on the latent factor was fixed to equal 1.0. (Preliminary analyses using IRT modeling showed no significant DIF for this item as a function of age or gender.) In the structural part of the model, the latent factor was regressed on five age-gender indicators. We classified respondents into three age groups: 18–39, 40–69, and 70 or older to differentiate young, middle-aged, and older adults. We included persons aged 65–69 in the middle-aged group because we wanted the elderly group to exclude the relatively healthy younger-old. Because we anticipated that the impact of gender-based DIF might vary in different age groups, we included interaction effects in the model. To examine each combination of age group and gender, analyses included five dummy variables (young men, young women, middle-aged men, middle-aged women, and elderly women). The reference category was men aged 70 or older. These two parts constituted the *no-DIF model*.

Next, each of the IADL and ADL items was examined individually for DIF. For each item, DIF was captured by a set of five direct effects, one from each dummy age/gender indicator to that item. A set of "forward inclusion" models was estimated, each adding the five age/gender DIF effects for one item to the no-DIF model. Items that did not manifest DIF were identified by a nonsignificant difference between the no-DIF model and the model containing DIF for that item. In view of the large sample size, this was a stringent test of the absence of DIF. These analyses validated the choice of toileting as an anchor item and identified other potential anchor items.

Next, we estimated a *DIF model*, which contained direct age-gender effects to all items, other than the no-DIF items identified previously. A series of "backwards elimination" models was then estimated, removing the five DIF effects from the DIF model one item at a time. We compared the chi-square for each of these models to that for the DIF model to identify items with especially severe DIF. Finally, to examine the extent to which adjusting for DIF affected estimates of age and gender differences in latent functional disability, we compared the magnitude of the direct effects of the age-gender indicators on the disability factor in the no-DIF and DIF models. We report unstandardized regression coefficients; because the estimated variance of the latent factor was .98, values of standardized coefficients were virtually identical to unstandardized ones.

The DIF and no-DIF MIMIC models were evaluated using standard criteria. A goodness of fit statistic, reflecting the discrepancy between the observed data (item means and covariances) and the model's predictions, can be compared with a chi-square distribution. However, because statistical power increases with sample size, chi-square goodness of fit tests should be viewed with caution because trivial differences often appear statistically significant. Consequently, we also examined other indicators of goodness of fit. The Comparative Fit Index (CFI) and the Tucker–Lewis Index (TLI) compare the substantive model to a baseline null model of independence among the observed variables; values of 0.95 or higher suggest acceptable fit (Hu & Bentler, 1999). The root mean square error of approximation (RMSEA)

assesses misfit per degree of freedom; values less than 0.08 suggest an acceptable fit, whereas values less than 0.05 suggest very good fit (Browne & Cudeck, 1993).

Among the subset of persons who received help with at least one task, 92% had complete data. To deal with missing data, we replicated MIMIC analyses twice. In one set of analyses, we removed any case with missing data (listwise deletion). In the second set of analyses, we assigned missing data a value of zero (i.e., did not receive help or supervision); this procedure reflects the fact that this response was most likely, occurring more than 92% of the time for each item among those aged 18 or older. Both sets of analyses led to the same conclusions, raising confidence that biases from missing data were minimal. We report analyses in which missing values were recoded to zero.

All analyses were conducted using Mplus software, version 2.01 (Muthen & Muthen, 1998). Because the observed indicator variables were dichotomous, we used weighted least squares estimation, which is appropriate for models containing categorical variables. All analyses incorporated the NHIS-D sampling weight, normalized so that the sum of the weights equaled the unweighted sample size.

Results

Dimensionality

For the 10,371 adults with a disability, the first eigenvalue for the matrix of tetrachoric correlations among the 12 ADL/IADL items was 7.968; the second and third eigenvalues were 1.570 and 1.318, respectively, with all remaining eigenvalues less than 1.0 in magnitude. The large discrepancy between the first and remaining eigenvalues is consistent with a single latent dimension for these items. In a confirmatory factor analysis specifying a single factor (results not shown), all items but one had significant ($p < .001$) factor loadings, ranging in magnitude from .865 (light housework) to .981 (getting around inside). "Heavy housework" was the exception, with a nonsignificant factor loading (.007). Receipt of help for heavy housework was virtually uncorrelated (<.04) with all other ADLs/IADLs, with the exception of light housework (tetrachoric correlation = .55). We concluded that receipt of help with heavy housework does not reflect the same dimension as the remaining ADL/IADL items. Consequently, we removed this item from subsequent analyses, which were conducted using the eleven remaining ADL/IADL items.

Removing heavy housework from the analysis meant that persons who received help with *only* heavy housework were no longer considered disabled. Among the 10,371 persons who received help with at least one ADL/IADL task, 45% (n =4,621) indicated that they received help or supervision *only* with heavy housework. These persons had "no" responses to the remaining 11 items and thus provide little information regarding item performance. Consistent with our focus on measuring severity of disability among persons with disabilities, analyses were conducted on the remaining 5,750 respondents.

The first two eigenvalues of the tetrachoric correlation matrix of the remaining 11 items, calculated on the 5,750 disabled respondents, were 6.28 and 1.89, with the remaining eigenvalues less than 1.0. This is consistent with a single major dimension.

Demographic Characteristics

Among the 5,750 adults with a functional disability on at least 1 of the 11 ADL/IADL tasks, 17% were between 18 and 39 years old; 38% were between 40 and 69 years old, and 45% were aged 70 or older. Almost two-thirds were women (63% vs. 37%). The proportion of women was greater in older age groups: 53% among the young, 59% among the middle aged, and 70% among the old.

Table 1 shows the mean number of tasks for which people received help or supervision, by age-gender groups. Overall, adults with functional disabilities on average received help with 3.34 of 11 ADL/IADL tasks. To examine age and gender differences, we conducted multiple regression analyses using SUDAAN to incorporate the complex sampling design. Men had significantly more disabilities than women (3.47 vs. 3.26, $p < .01$). Elderly respondents had more limitations than the younger groups (3.63, 3.15, and 2.97 for elderly, middle-aged, and young persons, respectively, $p < .0001$), but differences between middle-aged and young persons were not significant.

Table 2 reports percentages of adults with disabilities who received help or supervision with each task. Items are ordered by prevalence of disability. IADLs were more prevalent than ADLs, but there was not a strict IADL-ADL hierarchy. Two items, bathing and telephoning, did not follow a strict IADL-ADL ordering. The proportion receiving help with telephoning (15%) was similar to the proportion receiving help using the toilet, whereas the proportion receiving help with bathing (38.4%) was similar to the proportions helped with preparing meals and managing money.

A confirmatory factor analysis of the 11 items, using tetrachoric correlations and weighted least squares estimation, resulted in a CFI of .96 and a TLI of .95, which indicated acceptable fit. However, the goodness of fit chi-square was significant ($\chi^2 = 2171.17$, $df = 44$, $p < .000$), and the RMSEA was .092. In view of the somewhat high RMSEA, we examined derivatives among residual covariances to ascertain which ones might be contributing to lack of fit. We incorporated four covariances among residuals into the model: money management with telephoning, meal preparation with shopping and with light housework, and bathing with dressing. These residual covariances were between complementary tasks (e.g., bathing and dressing or meal preparation

Table 1. Mean Number of ADL/IADL Tasks With Which Help Was Received, by Age and Gender

Age Group	Men	Women	Total
18–39	3.10 (.15)	2.85 (.11)	2.97 (.10)
40–69	3.32 (.10)	3.03 (.08)	3.15 (.06)
70 and over	3.84 (.12)	3.54 (.07)	3.63 (.07)
Total	3.47 (.07)	3.26 (.05)	3.34 (.04)

Notes: Standard errors are in parentheses. Standard errors are adjusted for complex sampling. Sample is 5,750 adults receiving help with at least one of 11 ADL/IADL tasks. ADL = activity of daily living; IADL = instrumental activity of daily living.

Table 2. Proportion Receiving Help or Supervision With Each IADL or ADL Task

IADL/ADL Task	Proportion
Shopping	64.2
Doing light housework	50.0
Preparing own meals	38.5
Bathing	38.4
Managing money	36.2
Dressing	29.0
Getting in and out of bed or chair	22.1
Getting around inside	16.7
Using the telephone	15.1
Using the toilet	15.0
Eating	9.0

Notes: *Data Source*: Disability Supplement to the National Health Interview Survey Phase I, 1994 and 1995 combined sample, adults receiving help with at least one of 11 ADL/IADL tasks (n = 5,750). IADL = instrumental activity of daily living; ADL = activity of daily living.

and shopping) or between tasks that share a strong cognitive component (i.e., money management and telephoning). Residual correlations ranged from .12 for bathing and dressing to .27 for money management with telephoning. The goodness of fit chi-square for the revised model was 1,218.89 (df = 40, $p < .000$), and CFI and TLI were .98 and .97, respectively. Inclusion of the residual covariance parameters reduced the RMSEA to .072.

Table 3 presents the factor loadings and thresholds derived from the confirmatory factor analysis, including these four covariance parameters. Factor loadings were statistically significant ($p < .001$). The loadings were very high for all ADLs, ranging from .85 to .97; they were somewhat lower for IADLs, with managing money, light housework and shopping having the lowest loadings, ranging from .47 to .58.

The item threshold reflects the point along the latent disability continuum at which the item provides the strongest discrimination between those with more versus less disability. The results suggest that shopping provides the most information for those with the least severe disability, whereas eating provides information at the most severe levels of disability. The order of items in terms of their thresholds was generally consistent with the order in terms of proportion receiving help (as shown in Table 2). Managing money and using the telephone, however, had different positions in the overall order of items by threshold than they did when ordering the items by proportion receiving help.

Table 3. Item Parameters (Loadings and Thresholds) Estimated in Confirmatory Factor Analysis Model

IADL/ADL Tasks	Loadings	Thresholds
Shopping	.466	−.440
Doing light housework	.549	−.064
Preparing own meals	.810	.229
Bathing	.851	.312
Using the telephone	.757	.345
Dressing	.919	.541
Getting in/out of bed/chair	.974	.773
Getting around inside	.968	.886
Using the toilet	1.000	.924
Managing money	.584	.975
Eating	.917	1.189

Notes: Item parameters obtained from one-factor confirmatory factor analysis. Model includes four error correlations (N = 5,750). All loadings are significant ($p < .001$). IADL = instrumental activity of daily living; ADL = activity of daily living.

Identifying Items With DIF

To identify items with DIF, we estimated several MIMIC models; all incorporated four residual covariances, specified previously. The first model (no-DIF) contained no direct (DIF) effects from the five age-gender groups to individual items. The goodness of fit chi-square for the no-DIF model was 2,398.08 (df = 90, CFI = .96, TLI = .96, RMSEA = .067). We then estimated 11 models, each adding five DIF effects for one item. Models with DIF effects for toileting (χ^2 = 2,386.71, df = 85) and for getting around inside (χ^2 = 2,393.08, df = 85) did not differ significantly from the no-DIF model ($p < .01$), suggesting that DIF was not present for these items. Subsequent models did not include age-gender DIF for these two items.

The chi-square for a model with DIF for the remaining nine items (the *DIF* model) was 1,384.32 (df = 45). The difference in chi-squares between this model and the no-DIF model (1,013.84, with 45 degrees of freedom) was statistically significant. Thus, including differential effects for age and gender significantly improved the fit of the model. Other fit indices indicated acceptable overall fit (RMSEA = .072, TLI = .95, CFI = .98).

To combine the five DIF effects for each item into a summary statistic, we compared the chi-square of the full-DIF model (i.e., nine items with DIF) with the chi-square from a model that eliminated DIF effects for that specific item. The last column of Table 4 reports the chi-square difference for each item, with five degrees of freedom. Each of the nine items had a significant chi-square ($p < .001$). Three items had relatively small chi-square difference values—eating, dressing, and meal preparation. Two items, shopping and managing money, had chi-square values that were notably higher than the rest. Other items with relatively large DIF were doing light housework, bathing, and using the telephone.

Table 4 shows the direct effects of age-gender groups on each IADL/ADL item, with elderly men as the reference group. Because they are estimated controlling for latent disability, these effects represent estimates of DIF. For young and middle-aged persons, most of the significant effects were negative. (Significance was assessed at the .001 level to adjust for the large number of comparisons that are being made in this table.) A negative coefficient indicates that help with a task was less likely to be received, compared with elderly men, than would be expected given latent disability. For example, shopping had significant and large DIF effects for all age-gender groups except middle-aged women. Compared with elderly men, young people and middle-aged men were less likely to receive help with shopping at the same level of latent disability. In contrast, managing money had a high positive DIF effect among young men, who were much *more* likely than elderly men to report they received help than would be expected. The DIF effect for young women

Table 4. Estimates of DIF Effects for Each Item (Elderly Men Is Reference Group)

IADL/ADL Tasks	Young Men	Young Women	Middle-Aged Men	Middle-Aged Women	Elderly Women	Chi-Square Difference[b]
Shopping	−.293***	−.316***	−.354***	−.145	.384***	275.91
Doing light housework	−.449***	−.117	.019	.160	−.021	92.10
Preparing meals	−.049	−.161	−.197	−.168	−.039	18.02
Bathing	−.105	−.430***	−.109	−.283***	.013	88.29
Managing money	.723***	.207	−.149	−.324***	.034	218.47
Dressing	−.041	−.157	.039	−.127	−.108	20.57
Getting in/out of bed/chair	.034	.092	.143	.138	−.048	41.88
Getting around inside	a	a	a	a	a	a
Using the telephone	−.039	−.372***	−.357***	−.523***	−.248***	74.46
Using the toilet	a	a	a	a	a	a
Eating	.010	.073	−.119	−.204	−.122	19.71

Notes: DIF = differential item functioning; IADL = instrumental activity of daily living; ADL = activity of daily living.

[a]Constrained to zero, based on nonsignificant χ^2 difference from no-DIF model.

[b]Values are differences in χ^2 between model with DIF for 9 items and model excluding DIF for specific item in each row (5 *df* for χ^2 difference test).

***$p < .001$.

was also positive, although not significant at the .001 level ($p = .006$). In addition, middle-aged men and women were more likely to receive help transferring from a bed or chair, compared with elderly men at the same level of latent disability. Most coefficients for elderly women were nonsignificant. Compared with elderly men, elderly women were more likely to receive help with shopping and less likely to receive help with using the telephone.

To gauge the overall impact of adjusting for DIF for all nine items, we compared age-gender group effects on latent disability, with and without DIF adjustment. These effects can be interpreted as estimated differences in latent functional disability between each age-gender group and elderly men. As shown in the top section of Table 5, without adjusting for DIF, young women and both middle-aged groups appeared to be significantly less disabled than elderly men, but (surprisingly) young men were not. Controlling for DIF altered the estimates of the effects of age and gender on latent disability, and changed the conclusions about the relative disability of these groups. In the DIF model, the effects of

Table 5. Impact of DIF on Estimated Age and Gender Effects, Comparing Models With and Without "Shopping" and "Money" Items

Age-Gender Group	No-DIF Model	DIF Model	DIF Adjustment
Model with money and shopping			
Young women	−.267 (.050)***	−.211 (.073)	.056 (21.0%)
Young men	−.084 (.049)	−.306 (.072)***	−.222 (264.3%)
Middle-aged women	−.244 (.044)*	−.100 (.060)	.144 (59.0%)
Middle-aged men	−.142 (.046)***	−.028 (.061)	.114 (80.3%)
Elderly women	−.064 (.042)	−.094 (.055)	−.030 (46.9%)
Model excluding shopping			
Young women	−.255 (.050)***	−.249 (.074)*	.006 (2.4%)
Young men	−.071 (.050)	−.275 (.072)*	−.204 (287.3%)
Middle-aged women	−.242 (.045)***	−.108 (.060)	.134 (55.4%)
Middle-aged men	−.134 (.046)*	−.040 (.061)	.094 (70.1%)
Elderly women	−.144 (.043)*	−.089 (.056)	.055 (38.2%)
Model excluding money			
Young women	−.385 (.055)***	−.213 (.073)*	.172 (44.7%)
Young men	−.292 (.056)***	−.184 (.072)*	.108 (37.0%)
Middle-aged women	−.234 (.045)***	−.136 (.060)*	.098 (41.9%)
Middle-aged men	−.128 (.047)*	−.036 (.061)	.092 (71.9%)
Elderly women	−.088 (.043)*	−.105 (.055)	−.017 (19.3%)
Model excluding money and shopping			
Young women	−.380 (.055)***	−.240 (.074)*	.140 (36.8%)
Young men	−.254 (.057)***	−.165 (.073)*	.089 (35.0%)
Middle-aged women	−.242 (.046)***	−.148 (.061)*	.094 (38.8%)
Middle-aged men	−.140 (.047)*	−.057 (.061)	.083 (59.3%)
Elderly women	−.165 (.044)***	−.102 (.056)	.063 (38.1%)

Notes: Young is ages 18–39, middle-aged is ages 40–69, and elderly is 70 or older. Reference group is elderly men. Standard errors appear in parentheses in columns 2 and 3. In column 4, numbers in parentheses are the percentage changes in the corresponding coefficient from the no-DIF to the DIF models. DIF = differential item functioning.

*$p < .05$; ***$p < .001$.

being in the middle-aged group diminished for both men and women, resulting in a nonsignificant difference from elderly men. In contrast, the effect for young men became more negative, resulting in both young men and young women estimated to be significantly less disabled than elderly men.

The difference between elderly women and elderly men in latent functional disability was not significant in both the no-DIF and the DIF models. In the no-DIF model, young women were significantly less disabled than young men ($-.18$, $t = -3.68$), and middle-aged women were significantly less disabled than middle-aged men ($-.10$, $t = -2.56$). After controlling for DIF, however, gender differences among the young ($.10$, $t = 1.18$) and among the middle-aged ($-.07$, $t = .06$) were not significant. Gender DIF effects appear to be stronger among the young and middle-aged groups, compared with elderly persons. Observed gender differences in severity of functional disability diminish after DIF adjustment.

To provide a sense of the relative impact of adjusting for DIF, the last column in Table 5 shows the percentage change in each coefficient effected by controlling for DIF. The magnitude of the DIF adjustment was large, except for young women. It was especially large for young men (decreasing the disability estimate by 264%) and middle-aged men (increasing the disability estimate by 80%).

Effects of Deleting Items

If an item shows large DIF, one option is to drop the item rather than try to adjust for it in a statistical model. We assessed whether deleting items with high chi-squares for DIF effects would affect the magnitude of the DIF adjustment. Managing money and shopping had chi-square values that were much larger than the other items (Table 4). Consequently, we re-estimated the no-DIF and the DIF models first deleting only shopping, then removing only money, and finally removing both items (sections 2–4 of Table 5).

When the "managing money" item was removed from the analysis, the no-DIF model estimates changed dramatically for young men, who appeared much less disabled, compared with old men. Removing shopping reduced the estimate of disability for elderly women. Nonetheless, even when these two most problematic items were removed, DIF adjustment remained important. Controlling for DIF produced, at a minimum, a 35% change in parameter estimates. Estimated differences between each age-gender group and elderly men were reduced by controlling for DIF. Significant effects for middle-aged men and elderly women became nonsignificant when DIF effects were included in the model. In general, without DIF adjustment, we would conclude disability differences across age-gender groups were greater than they actually were.

Discussion

Summary of DIF Findings

Using the large and nationally representative samples in the 1994 and 1995 Disability Supplements of the NHIS, we examined the extent of DIF by age and gender in a standard set of items measuring functional disability. We identified substantial DIF across age and gender groups using MIMIC models. Only two items showed no significant DIF—toileting and getting around inside. Group comparisons that did not adjust for DIF (e.g., Table 1 and the no-DIF model in Table 5) suggested that elderly respondents had more severe functional disability than both younger and middle-aged respondents. After DIF adjustment, however, the differences between elderly and middle-aged groups in the severity of disability were not significant, whereas the difference between the young and other age groups became more pronounced.

After controlling for DIF, the gender effects on the latent factor show that women were slightly (but not significantly) less disabled than men in the middle-aged and elderly age groups. In contrast, the common finding in studies of functional disability is that women are more likely to have disabilities than men (Dawson, Hendershot, & Fulton, 1987; U.S. Bureau of the Census, 1990). The higher prevalence of any disability among women was replicated in this study, because nearly 9% of women in the overall NHIS-D sample had a functional disability, compared with 5% of men. However, when analyses were restricted to the subset of persons with some functional disability, the level of disability tended to be more severe among men in the two older groups. Spector and Fleishman (1998) found similar results among elderly adults. These results highlight the importance of distinguishing factors that affect the prevalence of *any* disability in the general population from factors that affect the *severity* of disability among persons with disabilities.

IADL items, especially managing money and shopping, tended to have larger DIF effects than ADL items. Relative to men aged 70 and older, at the same disability level, young men were more likely to respond that they received help with managing money, and both men and women in the two youngest age groups were less likely to respond they were receiving help with shopping. DIF was less substantial for ADLs, except bathing.

Strategies for Dealing With DIF

Studies that include participants from across the age spectrum or examine gender differences and do not adjust for DIF may produce biased estimates of age or gender differences in functional disability. The potential existence of DIF needs to be addressed in the design and analysis phases of such studies. Statistical adjustment, by using latent variable models, is one approach to reducing the impact of DIF on group comparisons. Another approach is to reword questions that exhibit DIF. The NHIS-D questions themselves are very general, leaving substantial room for interpretation; thus, improvements may be possible.

As noted, a third strategy to reduce the impact of DIF is to delete problematic items, such as managing money. To assess whether removing problematic items was sufficient to reduce the impact of DIF, we compared models that excluded managing money and/or shopping. Removing these items did not eliminate DIF, suggesting the importance of statistically adjusting for DIF in making group comparisons of functional disability. When considering whether to delete an item from a scale, one must also consider the impact on the content validity of the instrument and the potential loss of information for measuring certain levels of the latent trait. If

an item's threshold differs from those of other items, the item is measuring a point on the latent dimension that is not well represented by the other items in the scale, and removing the item may have a negative impact on precision or content validity.

Psychometric Properties of IADL and ADL Items

The present results are consistent with Spector and Fleishman's (1998) findings, among elderly people with disabilities, that ADL and IADL items do not form a clear hierarchy. In the present study, the ADL task of bathing had a threshold parameter similar in magnitude to some IADL items, whereas the IADL task of telephoning had a threshold parameter similar to some ADL items. As in the earlier analyses, the threshold parameters display a gap between shopping and the item with the next lowest threshold, indicating an area on the latent continuum of functional disability in which the items do not provide a great deal of information. In an analysis of IADL and ADL items in the National Long Term Care Survey, doing laundry followed shopping in terms of thresholds and had a high loading (Spector & Fleishman, 1998). The addition of this item, which is not included in the NHIS-D, would likely improve discrimination among those with mild functional disabilities.

Heavy housework ("Doing heavy work around the house like scrubbing floors, washing windows, and doing heavy yard work") should not routinely be included in measures of functional disability. This item had virtually negligible correlations with all but one other ADL/IADL item and thus did not appear to indicate the same construct. Excluding heavy housework had a large impact on the size of the sample defined as disabled. If heavy housework was included among the tasks that define functional disability, 6.9% of adults were disabled. However, if receipt of help with heavy housework was excluded as an indicator of disability, then the estimated proportion of adults with disabilities dropped to 3.8%.

Latent trait analyses assume that the items all reflect a single underlying dimension. We used the criterion (Lord, 1980) that a large first eigenvalue relative to the second was sufficient evidence to suggest unidimensionality. For 11 ADL/IADL items, the first two eigenvalues were 6.3 and 1.9. Using this criterion, the present study provides evidence that NHIS-D IADL and ADL items can be combined in a unidimensional scale. Others (e.g., Teresi et al., 2000a) have used similar eigenvalue patterns as evidence of unidimensionality. For example, Teresi and colleagues (2000b), in analyses of a cognitive screening measure, reported first eigenvalues ranging from 5.1 to 5.7 in different subgroups and second eigenvalues ranging from 1.4 to 1.5 as evidence of unidimensionality.

Although we have met a statistical standard commonly used for determining unidimensionality, the existence of age and gender DIF, and the inclusion of correlated errors in the model, suggest that the scale is not perfectly unidimensional. The strong DIF exhibited by managing money, and the correlated error between money and telephoning, suggest the presence of a secondary cognitive factor. Of the 11 ADL/IADL items, managing money and using the telephone appear to have the strongest cognitive component. Only two items, however, may not be sufficient to demonstrate the existence of a cognitive factor. IADL/ADL scales have been criticized for being insensitive to functional losses that result from cognitive deficits (Spector, 1997; Tappen, 1994). Future research could include items intended to tap functional loss associated with mild cognitive deficits, such as items in the Pfeffer Functional Activity Scale (Pfeffer, Kurosaki, Harrah, Chance, & Filos, 1982). Combined with standard ADL/IADL items, such an expanded item pool may provide clearer evidence for a separate cognitive dimension of functional disability.

Limitations

Limitations of the analyses need to be acknowledged. The MIMIC model allows the loadings to vary across items. However, the MIMIC model inherently imposes the restriction that the loading for each item does not vary as a function of age or gender. Future research on the psychometric properties of ADL/IADL items should examine the comparability and equivalence of factor structures and loadings across age and gender groups.

Another limitation pertains to the sampling design. Although we incorporated the sampling weights into the analyses, software limitations precluded us from adjusting for other aspects of the complex sample, such as stratification and clustering. In part, this motivated our selection of a conservative .001 level for significance tests.

The chi-square goodness of fit tests for both the DIF and the no-DIF MIMIC models were statistically significant, indicating that the models' predictions did not perfectly match the observed means and correlations. However, a large sample size—5,750 in this study—inflates the value of the chi-square statistic. More troubling were the values of the RMSEA, which were still somewhat high, despite the expedient of estimating four disturbance covariances. To ascertain the degree to which imposing a single disability factor might contribute to lack of fit, we estimated a MIMIC model with two factors, corresponding to ADL and IADL items, respectively. The two-factor DIF model, with no correlated disturbances, had a chi-square of 1,844.59 ($df = 48$) and an RMSEA of .081. These values were worse than those for the single-factor DIF model with correlated disturbances; a two-factor specification may not dramatically improve the model's fit. (The ADL and IADL factors correlated .67.)

The nature of the functional disability criterion—receiving human help or supervision—may shape the generalizability of the results. In many studies, rather than indicate receipt of help, respondents indicate how much difficulty they have performing ADL or IADL tasks. Using a criterion of difficulty may give rise to different pattern of results. In particular, the criterion of receipt of help may result in higher inter-item correlations than the difficulty criterion; once a network of helpers is activated, it may provide assistance with multiple tasks. Receiving help with ADL/IADL tasks may also reflect cultural or social factors, such as expectations concerning when it is appropriate to request or offer help with these tasks. Incorporating use of mechanical aides into the definition of disability may also alter the pattern of results. Future research should examine more closely the effect of the disability criterion on the psychometric properties of functional disability measures.

This study has focused on ADLs and IADLs as measures of functional disability. Other measures of functional disability include items assessing mobility or cognition (Kempen, Miedema, Ormel, & Molenaar, 1996; Mahoney & Barthel, 1965). The psychometric properties of several functional disability scales have been reported elsewhere (Cohen & Marino, 2000; Spector, 1996). In addition, measures of disability, more broadly construed, ascertain performance of social roles and participation in socially valued activities. Although our analyses do not address these other dimensions, our results suggest that researchers be attentive to the possibility of DIF in these measures.

In conclusion, evidence of substantial age and gender DIF implies that adjustments for DIF are necessary when making comparisons of disability levels across age and gender groups. Concern with DIF first arose in educational testing. Correctly answering certain test items could potentially require extraneous information that was differentially available to members of certain sociodemographic groups. Decisions for individual students based on such biased items could be inappropriately disadvantageous. A similar situation may exist with measures of functional disability. ADL/IADL items are often used to determine eligibility for services or to allocate program resources to individual clients. Some states have initiated efforts to consolidate in one agency long-term care programs for elderly persons and for people with disabilities (e.g., Oregon, Texas, Wisconsin). This could increase the likelihood that clients of various ages would be compared. Based on this study's findings, if age-based DIF is ignored, one consequence could be that middle-aged persons may appear to be less disabled, compared with elderly persons, and may receive fewer program resources than their underlying severity of disability would merit. In addition, programs for nonelderly persons that use ADLs/IADLs as criteria for allocating resources may make inappropriate decisions due to gender DIF.

Acknowledgments

The views expressed here are those of the authors. No official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred.

Address correspondence to Dr. John A. Fleishman, Center for Cost and Financing Studies, Agency for Healthcare Research and Quality, 2101 East Jefferson Street, Rockville, MD 20852. E-mail: jfleishm@ahrq.gov

References

Allen, S. M., Mor, V., Raveis, V., & Houts, P. (1993). Measurement of need for assistance with daily activities: Quantifying the influence of gender roles. *Journal of Gerontology: Social Sciences, 48*, S204–S211.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Thousand Oaks, CA: Sage.

Camilli, G., & Shepherd, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Cohen, M. E., & Marino, R. J. (2000). The tools of disability outcomes research: Functional status measures. *Archives of Physical Medicine and Rehabilitation, 81*(12 Suppl. 2), S21–S29.

Dawson, D., Hendershot, G., & Fulton, J. (1987). *Aging in the eighties, functional limitations of individuals age 65 years and over* (No. 133, DHHS Publication No. PHS 87-1250). In *Advanced data from vital and health statistics*. Hyattsville, MD: Public Health Service.

Gallo, J. J., Anthony, J. C., & Muthen, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology: Psychological Sciences, 49*, P251–P264.

Grayson, D. A., Mackinnon, A., Jorm, A. F., Creasey, H., & Broe, G. A. (2000). Item bias in the Center for Epidemiologic Studies Depression Scale: Effects of physical disorders and disability in an elderly community sample. *Journal of Gerontology: Psychological Sciences, 55B*, P273–P282.

Groot, W. (2000). Adaptation and scale of reference bias in self-assessments of quality of life. *Journal of Health Economics, 19*(3), 403–420.

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

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Katz, S., Ford, A. B., Moskowitz, R. W., Jackson, B. A., & Jaffe, M. W. (1963). Studies of illness in the aged. The index of ADL: A standardized measure of biological and psychosocial function. *Journal of the American Medical Association, 185*, 914–919.

Kempen, G. I., Miedema, I., Ormel, J., & Molenaar, W. (1996). The assessment of disability with the Groningen Activity Restriction Scale. *Social Science and Medicine, 43*(11), 1601–1610.

Lawton, M. P., & Brody, E. M. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. *The Gerontologist, 9*, 179–186.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mahoney, F. I., & Barthel, D. W. (1965). Functional evaluation: The Barthel Index. *Maryland State Medical Journal, 14*, 61–65.

Marshall, S. C., Mungas, D., Weldon, M., Reed, B., & Haan, M. (1997). Differential item functioning in the Mini-Mental State Examination in English- and Spanish-speaking older adults. *Psychology and Aging, 12*(4), 718–725.

McDonald, R. (1999). *Test theory*. Mahwah, NJ: Erlbaum.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297–334.

Morales, L. S., Reise, S. P., & Hays, R. D. (2000). Evaluating the equivalence of health care ratings by whites and Hispanics. *Medical Care, 38*(5), 517–527.

Muthen, B. O., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*(2), 133–142.

Muthen, L. K., & Muthen, B. O. (1998). *Mplus user's guide*. Los Angeles: Author.

Pfeffer, R. I., Kurosaki, T. T., Harrah, C. H., Jr., Chance, J. M., & Filos, S. (1982). Measurement of functional activities in older adults in the community. *Journal of Gerontology, 37*, 323–329.

Spector, W. D. (1996). Functional disability scales. In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trials* (2nd ed., pp. 133–143). Philadelphia: Lippincott-Raven.

Spector, W. D. (1997). Measuring functioning in daily activities for persons with dementia. *Alzheimer Disease and Associated Disorders, 11*(Suppl. 6), 81–90.

Spector, W. D., & Fleishman, J. A. (1998). Combining activities of daily living with instrumental activities of daily living to measure functional disability. *Journal of Gerontology: Social Sciences, 53B*, S46–S57.

Spector, W. D., Fleishman, J. A., Pezzin, L. E., & Spillman, B. C. (2000). *The characteristics of long-term care users* (AHRQ Publication No. 00-0049). Rockville, MD: Agency for Healthcare Research and Quality.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408.

Tappen, R. M. (1994). Development of the refined ADL Assessment Scale for patients with Alzheimer's and related disorders. *Journal of Gerontological Nursing, 20*(6), 36–42.

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000a). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19*(11-12), 1651–1683.

Teresi, J. A., Kleinman, M., Ocepek-Welikson, K., Ramirez, M., Gurland, B., Lantigua, R., et al. (2000b). Applications of item response theory to the examination of the psychometric properties and differential item

functioning of the Comprehensive Assessment and Referral Evaluation Dementia Diagnostic Scale among samples of Latino, African American and White non-Latino elderly. *Research on Aging, 22*(6), 738–773.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland (Ed.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

U.S. Bureau of the Census. (1990). The need for personal assistance with everyday activities: Recipients and caregivers (Series P-70, No. 19). In *Current Population Reports*. Washington, DC: U.S. Government Printing Office.

Received November 12, 2001
Accepted February 20, 2002
Decision Editor: Fredric D. Wolinsky, PhD