Combining Activities of Daily Living With Instrumental Activities of Daily Living to Measure Functional Disability

William D. Spector\(^1\) and John A. Fleishman\(^2\)

\(^1\)Center for Organization and Delivery Studies, Agency for Health Care Policy and Research.
\(^2\)Center for Cost and Financing Studies, Agency for Health Care Policy and Research.

Measures of functional disability typically contain items that reflect limitations in performing activities of daily living (ADLs) or instrumental activities of daily living (IADLs). Combining IADL and ADL items together in the same scale would provide enhanced range and sensitivity of measurement. This article presents psychometric justification for a combined ADL/IADL scale. Data come from 2,977 disabled respondents in the 1989 National Long-Term Care Survey. Respondents indicated whether they received human help on 7 ADL items; they also indicated whether they were unable to perform each of 9 IADL items due to health reasons. Factor analyses using tetrachoric correlations demonstrated that 15 of the 16 items reflected one major dimension. Item response theory (IRT) methods were used to calibrate the items; a one-parameter IRT model fit the data. Item calibrations showed that ADL and IADL items were not hierarchically related. Analyses showed that a simple sum of item responses could be used to derive a measure of functional disability. Implications of using a 15-item ADL/IADL scale for eligibility determination and for comparing groups are discussed.
extends the range of disability measurement and avoids a narrow focus on only severe functional disability. But it ignores information about IADLs among those with ADL disability. Both approaches avoid combining ADL and IADL information for each respondent, regardless of ADL status. A third approach, which is considered in this article, is to develop a functional disability score that synthesizes all information about ADL and IADL disabilities for each respondent in a psychometrically appropriate manner.

Combining ADL and IADL items into one measure requires (a) determining if the measure is unidimensional, and (b) determining how to combine responses to several questions into an aggregate score. With regard to the first issue, if separate questions measure different constructs, combining responses into an overall score is misleading and inappropriate. With regard to the second issue, unjustified protocols for combining responses to multiple items may result in incorrect inferences about the relative disability of individuals or groups. This article examines the legitimacy of and the proper procedure for combining ADLs and IADLs, using factor analysis and item response theory (Muthen, 1978; Baker, 1992).

**Dimensionality.** — Controversy exists over whether functional disability is a unidimensional or multidimensional construct. If it is unidimensional, then ADLs and IADLs can be combined into one scale and a single score can describe a unique level of disability. If, on the other hand, it is multidimensional, a set of measures is necessary to provide a functional disability profile for an individual. Lawton and Brody (1969) suggested that IADLs require "a greater complexity of neuropsychological organization" than ADLs. However, they did not explicitly consider IADLs as a dimension distinct from ADLs. They observed that, although IADL tasks were typically more complex than ADLs, some overlap in degree of complexity between ADLs and IADLs was possible.

Subsequent research addressing the issue of dimensionality is diverse and results are inconsistent. Some studies suggest that functional disability is multidimensional and have found three or four separate constructs (Clark, Stump, & Wolinsky, 1997; Fitzgerald et al., 1993; Wolinsky & Johnson, 1991). Other studies argue that ADLs and IADLs comprise only one dimension (Kempen & Suurmeijer, 1989). Spector et al., 1987; Suurmeijer et al., 1994). These different conclusions may result from differences in the specific items used, and from variation in the methods used to determine dimensionality.

The set of items analyzed in prior studies often includes a mixture of items measuring impairments, functional limitations, and functional disabilities. Analyzing items reflecting multiple constructs may affect the number of dimensions ascertainment. For example, some studies include items concerning bowel and bladder incontinence (Jagger, Clarke, & Davies, 1986; Linacre et al., 1994); these items often comprise a separate dimension in factor analyses. Incontinence is often measured in terms of bowel or bladder control, reflecting an abnormality in a particular physical system; phrased in this manner, such items measure impairments, not a functional disability. Other studies combine functional disabilities with functional limitations, such as walking up stairs or upper or lower body limitations (e.g., bending and reaching; Clark et al., 1997; Haley, McHorney, & Ware, 1994; Jette, 1980; Linacre et al., 1994; Wolinsky & Johnson, 1991). Few studies have focused on a core set of ADL and IADL activities when assessing dimensionality (Clark et al., 1997; Fitzgerald et al., 1993).

Prior studies have typically used factor analysis or principal components analysis to determine dimensionality, although the decision rules and methods have varied (Fitzgerald et al., 1993; Wolinsky & Johnson, 1991). One reason for conflicting results may be that some studies factor analyze dichotomous variables, indicating presence or absence of a limitation for a particular task. Standard factor analysis methods using Pearson correlation coefficients, however, are not appropriate when variables are discrete or highly skewed (Rummel, 1970; Muthen, 1988). Artifactual "difficulty" factors may appear as a result of items with similar degrees of skewness loading together. Results may therefore be sensitive to the distribution of responses to the items. In addition, the magnitude of Pearson correlations between dichotomous items may have an upper limit less than 1, depending on the relative skewness of both variables.

When factor analyzing dichotomous variables, a recommended approach is to use tetrachoric correlations. Tetrachoric correlations assume that observed dichotomous variables represent unobserved continuous variables with a bivariate normal distribution; the tetrachoric correlation estimates the relationship between these latent continuous variables. The range of tetrachoric correlations is not constrained by disparate and skewed marginal distributions of observed variables. In this study, we perform factor analyses of dichotomous variables measuring functional disabilities (i.e., ADLs and IADLs) using tetrachoric correlations and the statistically appropriate generalized least squares estimation procedure (Muthen, 1978, 1988).

**Scaling procedures: IRT models.** — After the issue of dimensionality has been resolved, the researcher must consider the psychometrically appropriate procedure for combining items into multi-item scales. Item response theory (IRT) provides a sophisticated framework for developing measurement scales and calibrating individual items within a scale. For many years, IRT has been used to develop educational and psychological tests (Lord, 1980). Only recently has IRT been applied to assess health status measures (Haley, McHorney, & Ware, 1994; Granger et al., 1993; Linacre et al., 1994; Silverstein et al., 1992; Teresi, Cross, & Golden, 1989). Previous research applying IRT methods to assess disability has analyzed heterogeneous sets of items that combine ADLs, IADLs, impairments, and functional limitations; the present research applies IRT methods to a standard set of IADL and ADL items.

As applied to functional disability, IRT assumes that there exists an unobserved (latent) continuous dimension of disability, symbolized by θ, and that each individual can be placed along this dimension at a point that reflects the extent of his or her functional disability. IRT estimates item characteristic curves (ICCs) for each item. The ICC shows the probability of a positive response on a specific dichoto-
mous item as a function of the latent disability level \( \theta \). In the present context, the ICC shows how the probability of reporting a specific ADL or IADL disability varies, depending on one's underlying level of disability. (Figure 1 shows four different ICCs.)

IRT offers several advantages over classical psychometric approaches. First, classical item statistics, such as the proportion answering an item correctly or the item-total correlation, depend on the composition of the sample and the specific combination of items used. Classical item statistics are typically based on correlations, which are affected by the variances of the variables. In contrast, IRT item parameter estimates are invariant (within a linear transformation) to the population distribution of the trait being measured, analogous to regression coefficients (Hambleton, Swaminathan, & Rogers, 1991). Second, the IRT model is nonlinear, with an S-shaped ICC. In the analysis of dichotomous variables, nonlinear models have several advantages that make them more attractive than the linear model of classical test theory (Waller et al., 1996). Third, IRT provides an estimate of reliability for each item. Unlike classical test theory, which assumes that the scale's reliability is constant for all respondents regardless of \( \theta \), in IRT an item's reliability is not constant but varies, depending on \( \theta \). IRT thereby provides a mechanism for selecting items to measure a specific level of \( \theta \) most accurately. Fourth, IRT provides a framework for assessing item bias.

The shape of the ICC is determined by two item parameters, \( \beta \) and \( \alpha \), where "i" refers to different items. Items may vary in their parameters, and their ICCs will have different shapes. \( \beta \), the location parameter, represents the point on the latent dimension that corresponds to a .5 probability of a positive response to the item (in a two-parameter IRT model). A high value of \( \beta \) implies that a high probability of making a positive response to the item will be predominantly limited to individuals with high values of \( \theta \). \( \alpha \) indexes the discrimination ability of the item (the slope of the curve). The greater the value of this parameter, the more sharply the item discriminates among individuals with different values of \( \theta \) (at the point of inflection of the ICC). This parameter is roughly analogous to a factor loading in factor analysis; the greater the magnitude, the greater the degree to which responses are affected by the latent variable. \( \beta \) reflects the point along the range of \( \theta \) where the item discriminates best. Each item provides information about \( \theta \) over the entire range of disability, but the amount of information varies across the range of \( \theta \), and the maximum information occurs at the value of the location parameter (Lord, 1980). Thus, one item may provide information about relatively mild levels of disability, whereas

![Figure 1. Illustrative item characteristic curves.](http://psychsocgerontology.oxfordjournals.org/Downloaded from)
another item may discriminate best at severe levels of disability. IRT provides estimates of the $\beta$ and $\alpha$ parameters for each item.

Descriptions of the IRT model typically refer to the location parameter ($\beta$) as the item’s “difficulty.” The greater the magnitude of this parameter, the lower the proportion of people who answer positively. In the familiar context of IRT modeling — educational testing — estimates of ability are derived. In the gerontological literature, interest focuses on disability. We have chosen to be consistent with this literature and scale the items to measure disability. To avoid confusing double negatives, we refer to the item’s “location,” instead of its “difficulty.”

One- and two-parameter IRT models must be distinguished. In the one-parameter model, all items have equal slopes ($\alpha_i$), and the ICCs for all items are parallel. With a one-parameter model the location parameter establishes an unequivocal order of the items. In the two-parameter model, $\alpha_i$ varies across items, and the ICCs may cross each other. In Figure 1, ICCs 1, 3, and 4 have the same slope, whereas the slope is different for Item 2. The probability of a positive response to Item 1 is higher than Item 2 at some levels of disability and lower at other levels.

The number of parameters in the IRT model has significant implications for estimating each person’s latent disability level, $\theta$. In a one-parameter model, the estimated disability score for an individual depends only on the number of items that person answers positively. For example, all persons with four positive responses to a set of items receive the same estimated disability score, regardless of the particular items each person answered positively. In contrast, with a two-parameter model the score depends on the particular pattern of responses as well as the number of items answered positively.

The goals of this article are to assess the dimensionality of ADLs and IADLs and to use IRT methods to estimate disability scores for individuals. Several studies applying the IRT methodology have assumed a one-parameter model without testing to see if it is appropriate (Haley, McHorney, & Ware, 1994; Heinemann et al., 1993). We explicitly compare the one- and two-parameter IRT models. We discuss the implications for research and for using functional disability measures to determine program eligibility.

**Methods**

**Data**

Data are from the 1989 National Long-Term Care Survey (NLTCs; Manton, Corder, & Stallard, 1993). This survey was designed to serve two purposes: (a) to provide a nationally representative cross-sectional sample of functionally disabled persons in 1989, and (b) to link with the prior 1982 and 1984 surveys to provide estimates of change in functional disability. The original (1982) sample was taken from a Medicare enrollee frame provided by the Health Care Financing Administration. Persons who were impaired or institutionalized in 1982 and 1984 were automatically included in the 1989 survey and provided data on ADL and IADL problems. In addition, a screening process was used in 1989: Persons defined as unimpaired in 1982 and 1984 surveys, as well as a sample of Medicare enrollees who turned age 65 between 1984 and 1988, were given a screening interview to determine if they had IADL or ADL problems that had lasted or were expected to last three months. A total of 15,255 older adults (unweighted) completed screening interviews (Clark et al., 1997). Those with ADL or IADL problems were eligible for the detailed 1989 survey, in addition to those with prior impairment or institutionalization. A total of 4,463 people received the detailed interview.

Because the health status of a large number of disabled persons actually improves, many of the 1982 and 1984 impaired or institutionalized group were not functionally disabled in 1989. Of the 4,463 persons who were interviewed, 1,496 had no functional disabilities in 1989. To focus on a cross sectional sample of functionally disabled persons, we restricted the study sample to the 2,977 persons who were functionally disabled in at least one of sixteen ADLs or IADLs in 1989. This is a nationally representative sample of elderly functionally disabled persons in the community.

**Measuring ADL and IADL Disability**

The screened-in sample was given a detailed interview about demographic characteristics, health status, impairments, disabilities, number of caregivers and amount of ADL and IADL help received. Measures of functional disability included activities of daily living (Katz et al., 1963) and instrumental activities of daily living (Lawton & Brody, 1969). The survey measured seven ADLs (bathing, dressing, getting around inside, transferring, toileting, incontinence, and feeding) and nine IADLs (preparing meals, doing housework, doing laundry, shopping, managing money, taking medicines, telephoning, going places outside of walking distance, and getting around outside). The reference period was the week prior to the interview. For ADLs, a positive (disabled) response was defined as receipt of human help to perform the task; human help included standby as well as hands-on help. Use of special equipment to perform a task, in the absence of human help, was coded as nondisabled. This decision is consistent with a large body of literature, beginning with Katz et al. (1963), that distinguishes between dependence and independence. For incontinence, a positive response was defined as receiving human help with cleaning up accidents or with managing a device, rather than bowel or bladder control per se; this reflects a focus on ability to maintain hygiene after accidents, rather than on propensity to have episodes of incontinence. (By focusing on maintaining hygiene, this item represents a functional disability rather than an impairment, and thus is conceptually consistent with the other functional disability items.)

For IADLs, the interview asked whether the person usually performed the task, whether the person could perform the task if necessary, and whether any inability was due to health problems. Respondents were coded as disabled on IADLs if they usually did not perform the task, could not perform the task, and their disability resulted from a health problem. In the survey, housework was defined as either heavy (e.g., scrubbing floors) or light (e.g., straightening up). However, persons were screened into the survey based only on the light housework item. We limited consideration...
to light housework because persons disabled only in heavy housework were ineligible for the survey.

Estimation Procedures
The first phase of the analysis examined the dimensionality of the 16 ADL and IADL items. As a cross-validation procedure, the sample was split into halves, and analyses were conducted separately for each half. In the first half, exploratory factor analyses were conducted; in the second half, confirmatory factor analyses were conducted. Both analyses were based on the tetrachoric correlations among the items, using the LISCOMP program (Muthen, 1988).

IRT analyses were performed using the BILOG-MG program (Zimowski et al., 1996). BILOG-MG estimates the item parameters by a marginal maximum likelihood procedure; estimates of each person's latent disability score are derived using a Bayesian (EAP: expectation a posteriori) approach (Bock & Aitkin, 1981).

The analyses were limited to the subset of respondents who had at least one “disabled” response to the initial set of 16 items; those who gave a “nondisabled” response to all 16 items were removed from the analyses. Ancillary analyses were conducted, using the entire sample of 4,463, and the ordering of items in terms of location parameters was identical to that reported below. In fact, some IRT methods cannot provide an estimate of θ for people who answer all items correctly or incorrectly, and these respondents are automatically removed from the analysis (Wright & Stone, 1979). (The EAP method can provide such an estimate, however.) People who give a uniform response to all items provide relatively little information concerning item parameters, and estimates of their θ are relatively imprecise. If a large proportion of people give uniform responses to all items, this points out that the scale is not sensitive to variations outside of a specific range.

RESULTS

Sample
The study sample consists of functionally disabled persons aged 65 and over in 1989. The mean age of the sample was 79 years; 69% were women; 82% were White. Table 1 presents the frequency of positive responses (indicating disability) on each of the 16 IADL or ADL items. The proportion of the sample indicating a positive response ranged from 73% for “going outside of walking distance” to 14% for “help with incontinence.” Generally, IADL items had the highest proportions with a positive response, and ADL items typically had the lowest. However, there were two anomalies: “bathing,” an ADL, had the fourth highest proportion with a positive response, and “telephoning,” an IADL, had the third lowest.

Exploratory and Confirmatory Factor Analyses
Exploratory factor analyses of the tetrachoric correlations for the first half of the sample (n = 1,489) revealed that the first eigenvalue was substantially greater in magnitude than the remaining eigenvalues. The first three eigenvalues were 10.080, 1.475, and 1.213, respectively, with remaining eigenvalues less than 1.0. One method for determining the number of factors, the scree method, looks for a large drop in the eigenvalues and then a trailing off of subsequent values (Rummel, 1970). If the first eigenvalue is large relative to the second, this suggests that the items are approximately unidimensional (Lord, 1980). A second rule for determining the number of factors is the number of eigenvalues greater than 1. The first method suggests a one-factor solution, whereas the second suggests a three-factor solution.

Table 2 presents the estimated factor loadings for a three-
factor model after oblique (Promax) rotation. The ADL items had strong (> .5) loadings on one factor; three items ("managing money," "taking medications," and "using the phone") that typically reflect cognitive confusion or poor memory loaded strongly on the second factor. The IADL items, however, tended to have moderate loadings on both the second and third factors. Removing one item ("go places outside of walking distance") resulted in a clearer factor structure: ADL items, IADL items, and "cognitive IADL" items each loaded strongly on one factor and relatively less strongly on the others. These results, combined with the fact that "go places outside of walking distance" could reflect the availability of transportation rather than disability, led us to discard this item from further analyses. Eigenvalues were recalculated, based on the reduced set of items. The large first eigenvalue pattern was maintained. The first three eigenvalues were 9.915, 1.326, and 1.020.

Although the three-factor model was clear and interpretable, the factors themselves had moderately strong correlations, which ranged between .608 and .652. In a two-factor model, the ADL items loaded strongly (> .760) on the first factor and IADL items loaded strongly (> .714) on the second. However, three items had high loadings on both factors: taking medicine (.396, .477), getting around outside (.355, .383) and help with incontinence (.596, .361). The two factors correlated .696. The two-factor structure thus was less clear than the three-factor structure. When only one factor was extracted, all items had substantial loadings, with the lowest loading being .710.

Using the second half of the sample, a confirmatory factor analysis specified three factors, corresponding to the three-factor exploratory results. Table 3 presents the estimates for this model. The factor loading for each item was strong and significant. The goodness-of-fit chi-square statistic for the model was 311.37 (df = 87). Compared to a null model of complete independence (χ² = 25553.49, df = 105), the fit of this model was excellent, with Bentler’s (1990) fit index = .99. (Values above .90 are conventionally considered to indicate a good fitting model.) The value of the root mean square error of approximation (RMSEA), which estimates the discrepancy between the population covariance matrix and the covariance matrix implied by the best fitting model, was .042; Browne and Cudeck (1993) suggested that values less than .05 indicate an acceptable model.

Estimated correlations between the three factors were .91 (Factors 1 and 2), .85 (Factors 1 and 3) and .86 (Factors 2 and 3). Factor correlations of this large magnitude suggest that a one-factor model describes the data parsimoniously. The model was re-estimated, constraining the factors to be perfectly correlated. This model resulted in a chi-square of 430.28 (df = 90). Although this was a statistically significant difference from the three-factor model, Bentler’s fit index was still .98, compared to a null model of complete independence. The RMSEA for this model was .050. In addition, the eigenvalues were 10.079, 1.213, and 1.099 in this half of the sample.

A two-factor model was also estimated by constraining the second (IADL) and third (cognitive) factors to be perfectly correlated and to have equal correlations with the first factor. This model resulted in a chi-square of 377.92 (df = 89). The estimated correlation between the two factors was .927, again suggesting that a one-factor model is appropriate.

These results are consistent with the hypothesis that a single dimension underlies responses to these items. Hambleton, Swaminathan, and Rogers (1991) noted that the unidimensionality assumption of IRT is met adequately if a dominant component or factor underlies responses to the items. The results thus justify estimating unidimensional item-response models for these data. Consequently, IRT analyses proceeded, based on a one-factor model.

### IRT Analyses

Table 4 presents results of IRT analyses of 15 items for the entire sample of 2,977 respondents. Both one- and two-parameter models were estimated. Table 4 shows location estimates from the one-parameter model (Column 1), and estimates of item locations (Column 2) and slopes (Column 3) from the two-parameter model. The chi-square (−2 times the log likelihood) for the one-parameter model was 39221.33, whereas the chi-square for the two-parameter model was 38524.16. A likelihood-ratio test comparing the two models indicated that the two-parameter model was significantly different from the one-parameter model (χ² = 697, df = 14). However, a large sample size increases the magnitude of this statistic, making statistical significance easy to attain. The difference in chi-square of 697 amounted to only a 2% reduction in the chi-square of the one-parameter model.

Estimates of the slopes from the two-parameter model ranged from .685 for “finances” to 1.694 for “toileting.” However, many of the slope estimates were concentrated between 1.2 and 1.5. In addition, the order of the items in terms of the location parameters was similar for both models. The only major difference in the ordering of item lo-

### Table 3. Confirmatory Factor Analyses in Second Half of Sample

<table>
<thead>
<tr>
<th>Item</th>
<th>Three-Factor Model</th>
<th>One-Factor Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Bathing</td>
<td>.912</td>
<td>.000</td>
</tr>
<tr>
<td>Dressing</td>
<td>.944</td>
<td>.000</td>
</tr>
<tr>
<td>Feeding</td>
<td>.956</td>
<td>.000</td>
</tr>
<tr>
<td>Toileting</td>
<td>.977</td>
<td>.000</td>
</tr>
<tr>
<td>Transferring</td>
<td>.975</td>
<td>.000</td>
</tr>
<tr>
<td>Getting around inside</td>
<td>.966</td>
<td>.000</td>
</tr>
<tr>
<td>Help with incontinence</td>
<td>.831</td>
<td>.000</td>
</tr>
<tr>
<td>Shopping</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Doing laundry</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Doing light housework</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Preparing meals</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Getting around outside</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Managing money</td>
<td>.000</td>
<td>.884</td>
</tr>
<tr>
<td>Taking medications</td>
<td>.000</td>
<td>.882</td>
</tr>
<tr>
<td>Using the phone</td>
<td>.000</td>
<td>.864</td>
</tr>
<tr>
<td>Chi-Square</td>
<td>311.37</td>
<td>430.29</td>
</tr>
<tr>
<td>df</td>
<td>87</td>
<td>90</td>
</tr>
<tr>
<td>Root mean square residual</td>
<td>0.122</td>
<td>0.182</td>
</tr>
<tr>
<td>Root mean square error of approximation</td>
<td>0.042</td>
<td>0.050</td>
</tr>
</tbody>
</table>
cations occurred for “taking medicines” and “managing money,” which ranked 9 and 11, respectively, in the two-parameter model but ranked 6 and 7, respectively, in the one-parameter model. Other items maintained the same relative order in both models.

Several factors suggest that the one-parameter model provides an acceptable fit. First, the fit of the two-parameter model was not markedly better than the one-parameter model, insofar as a 2% reduction in chi-square was substantively trivial. Second, many of the items had slope parameters of similar magnitude. Third, the correlation between the estimated 6s in the one- and the two-parameter models was .994. We therefore decided that the improvement in fit of the two-parameter model did not outweigh its added complexity. Subsequent analyses focused on the one-parameter model. The reliability of the one-parameter scale was .88, based on estimates of true-score and error variance derived from IRT (Zimowski et al., 1996).

**Item location parameters.** — The item location parameters for the one-parameter model ranged from a low of -.826 (“shopping”) to a high of 1.614 (“feeding”). Because this is a one-parameter IRT model, the ordering of item location parameters was consistent with the proportion reflecting the item, using this formula to calculate the probability of answering each item in the affirmative, and summing these probabilities across all 15 items (Hambleton, Swaminathan, & Rogers, 1991).

In general, the intervals between expected scores were roughly equal. Only one interval, between raw scores of 8 and 9, was as much as two times the average. Relatively smaller than average intervals appeared between 0 and 1, 1 and 2, and 5 and 6. The number of items with positive responses (the “raw count” score) correlated .99 with the estimated 6 of the one-parameter model and with the expected score. This result, together with the relatively equal spacing between the expected scores, suggest that a simple count of the IADLs and ADLs may be a reasonable approximation to the 6 estimate from the one-parameter model when measuring disability.
Relationships with Other Variables

We examined the relationship of the 15-item scale with several relevant variables, including a measure of the total number of hours of IADL and ADL help received in the last week. In the 1989 NLTCs, for each ADL and IADL, the respondent named both formal and informal caregivers. For each caregiver (up to six), respondents indicated how many hours of ADL help and how many hours of IADL help were provided in the last week. The correlation between the IADL/ADL scale and total hours of IADL/ADL help was .7. In a regression analysis of total hours of help on 3 separate scales, corresponding to the three factors postulated in Table 3, the $R^2$ was 0.493, compared to an $R^2$ of 0.487 for the IADL/ADL scale alone, indicating little gain in explanatory power from separating functional disability into three variables, rather than one. Table 5 shows that the mean hours of help ranged from 7 hours a week for persons with no disabilities to 99 hours for persons with 15 disabilities. This demonstrates the broad range of disability needs that this scale covers, from needs averaging as little as 1 hour a day to essentially around-the-clock help.

Table 6 presents mean estimates of $\theta$ for groups based on gender, race (White/non-White) and age (less than 80 vs 80 or older). As expected, older disabled respondents had higher mean disability than younger ones. In addition, average disability was higher among disabled non-Whites than among disabled Whites. It is surprising that mean disability was higher among men than among women (5.6 for men vs 5.0 for women based on the raw count; $t = 3.3$). Previous research has shown that women are more likely than men to be functionally disabled (Dawson, Hendershot, & Fulton, 1987; Foley et al., 1986), mainly because women are more likely than men to have nonfatal chronic conditions (Verbrugge, 1987). The present results show that, despite overall lower prevalence rates, among the disabled, men have more severe disability than women.

When interpreting this finding, one concern is that, controlling for their underlying level of disability, men would be more likely to respond positively to IADL items because they may not normally do these activities. In other words, gender bias in types of activities routinely performed might explain this result. To deal with this concern, some researchers reduce the number of IADL items used for men (Lawton & Brody, 1969). Another approach, which was used in the NLTCs, is to ask if inability to do an IADL is due to health problems. This approach should reduce the gender bias problem, but may not be sufficient to remove it totally.

A major advantage of IRT is its capability of systematically assessing differential item functioning (DIF). If an item has substantially different parameters in different subgroups, this suggests that the item may be "biased," reflecting a variable in addition to the one being measured (Thissen, Steinberg, & Wainer, 1992). We tested for gender DIF by performing a multiple-group IRT analysis and examining the largest differences in location parameters for men and women. For women, location parameters were lower for three items ("shopping," "going outside," and "bathing"), and were higher for three other items ("taking medications," "telephoneing," and "light housework"). Removing the latter three items from the scale did not alter the conclusion that men were more disabled (expected counts were 4.97 for men and 4.64 for women, $t = 3.24$). (The exclusion of the former set of items would make men even more disabled relative to women.) Although an extensive DIF analysis is beyond the scope of this article, these results illustrate the potential for IRT to illuminate issues of DIF.

**DISCUSSION**

This article has addressed the question of whether IADL and ADL items can be combined into a single scale to measure functional disability. Results suggest two major conclusions. First, factor analyses indicated that 15 ADL and IADL items can be represented as a one-dimensional construct. This justifies combining these items into a single overall scale. Major advantages of the 15-item IADL/ADL scale are its larger range, compared to scales based solely on IADLs or on ADLs, as well as its parsimony, compared to using separate ADL and IADL scales. Second, IRT analyses explicitly compared the one- and two-parameter IRT models and found that the one-parameter model adequately represented the data. The implication of this finding is that the total number of items with a "disabled" response is sufficient to represent a person's level of functional disabil-
ity. In addition, the analysis identified gaps in the scale where discrimination could be improved by additional well-chosen items.

**Dimensionality**

Prior analyses that obtained a multidimensional structure generally used more heterogeneous sets of items, which included measures of bowel or bladder continence and physical limitations in addition to ADLs and IADLs. Item heterogeneity may have contributed to the emergence of multiple dimensions. One exception is Fitzgerald et al. (1993), who used a standard set of IADL and ADL items; they found four correlated factors including an incontinence factor. Their study sample, hospitalized persons aged 45 and over, was very different from the other studies. In addition, like many studies, they used principal components analysis with dichotomous variables, which may result in additional artifactual factors.

Using 14,415 respondents from the NLTCs, Clark, Stump, and Wolinsky (1997) fit a three-factor model using 10 of the 15 items considered in this article (excluding dressing, meal preparation, getting around outside, help with incontinence, and eating). They used items from the screener, which did not provide as much detail as corresponding items from the community interview. They found that a three-factor model fit adequately for the whole sample, as well as for subgroups of White men, White women, and Black women. However, Clark et al. did not report interfactor correlations, nor did they directly compare one- and three-factor models.

The use of Pearson correlations in factor analyses of dichotomous items, as in some prior studies, may result in a bias toward multiple-factor solutions. For dichotomous (or polytomous) items, the maximum value of Pearson correlation coefficients may be constrained to be less than 1.0 due to differential skewness of the items. This may lead, in turn, to relatively low estimates of interitem or interfactor correlations. Using data from the first half of the sample, a standard principal components analysis, which used Pearson correlations among the 15 items and specified 3 factors, yielded interfactor correlations ranging from 40 to .52. These interfactor correlations were lower in magnitude than those obtained using tetrachoric correlations, and they make a weaker case for selecting a unidimensional representation for these items. The choice of a correlation coefficient may thus have led other investigators to prefer multifactor models.

The use of tetrachoric correlations when analyzing dichotomous items is considered the methodological "gold standard" (Stump, Clark, Johnson, & Wolinsky, 1997). However, tetrachoric correlations rely on the assumption that underlying latent variables are normally distributed. Checks of this assumption in the current data, using the method proposed by Muthen and Hofacker (1988), suggested that this assumption was violated in a number of instances. The effects of violating this assumption on estimates of factor loadings and interfactor correlations are largely unknown, however (Muthen, 1993). More methodological work needs to be conducted to develop procedures for dealing with any such violations in applied research.

Although the present results warrant treating these items as unidimensional, for some purposes, distinguishing three factors might have heuristic value. It is possible that the three factors might have different sets of antecedents or consequences. However, when examining the consequences of functional disability, using a single scale offers analytic advantages; in particular, including multiple subscales, with substantial intercorrelations, into a single regression model can give rise to problems of collinearity and related difficulty in interpreting regression coefficients. In addition, the present results do suggest that, in contrast to one- or three-factor models, a two-factor model (ADLs and IADLs) was not the best representation of these data. Neither the scree test nor the number of eigenvalues greater than 1 pointed to a two-factor model; in addition, solutions with two factors lacked the simple factor structure of a three-factor model and also revealed very high interfactor correlations. The common practice of deriving one scale of ADLs and a separate scale of IADLs may not be justified.

**IRT Analyses**

The IRT analyses demonstrated that a one-parameter model fit the data adequately. Although a likelihood-ratio test showed that the two-parameter model fit significantly better than a one-parameter model, a large sample size can produce significant differences that are substantively trivial. Moreover, the extremely high correlation between the individual disability estimates from the one- and two-parameter models implies that the two-parameter model does not add a great deal of information beyond the one-parameter model. The one-parameter model also has the advantage of assuring that items can be ordered unambiguously, in the sense that their ICC curves do not cross. In sum, the one-parameter model provides a simple and parsimonious representation for these data, whereas the slightly improved fit of the two-parameter model is outweighed by its added complexity. If a two-parameter model had been preferred, the implication would have been that an optimal disability score would have to consider the pattern of responses to each item, not simply the total number of “disabled” responses.

IRT provides other valuable information for assessing and developing measurement scales. First, the IRT estimates of θ can be treated as comprising an interval scale (Wright & Stone, 1979); in contrast, the number of disabled items does not possess interval-scale properties. Whereas the product-moment correlation between the IRT estimate of θ and the number of “disabled” responses will often be high (Nunnally & Bernstein, 1994, p. 434), for some analyses, such as when estimating change in θ over time, having an interval scale would be useful. Second, unlike classical test theory, IRT highlights the fact that a scale’s reliability fluctuates across the latent dimension, depending in part on the combination of item location parameters. Examination of the scale’s measurement precision at each point along the latent dimension aids in further item and scale development. Third, comparisons of item parameters across different subgroups provide a means of testing for differential item functioning, which may affect group comparisons.

IRT provides explicit procedures for selecting items to measure particular ranges of θ with precision. For example,
interest may center on measuring relatively severe levels of disability. ADLs are often used for this purpose. IRT item location parameters could be used to select a subset of items focused on this range. IRT analyses in this article suggest that most, but not all, ADL items do reflect high levels of disability, but “bathing” had a location parameter lower than some IADL items, while “telephoning” had a location parameter higher than many ADLs. Thus, for measuring severe disability, improvements may be possible by adding IADL items, such as “telephoning,” to the standard set of ADL items. Moreover, if the number of items to measure severe functional disability is limited, one might consider removing “bathing” from a set of ADL items. On the other hand, to measure less severe disability, “bathing” might be combined with IADL items.

The IRT analyses excluded a large number of people without any measured disability. For the purposes of calibrating the scale, persons with a score of zero provide minimal information concerning the performance of the items. Analogously, if one administered a mathematics test consisting of simple addition problems to college students, one could conclude only that the test was too easy for the large number of persons with perfect scores. For this group, one would not have a precise measure of their mathematics ability, nor could one differentiate among them in terms of mathematics ability. Nevertheless, a score of zero on the ADL/IADL scale does provide some important information. It is likely that an expanded disability scale, one which was more sensitive to milder forms of disability, would provide further discrimination among this group. Future efforts in scale development could address this issue.

Advantages of IRT over Guttman Scaling

Historically, ADLs and IADLs have frequently been analyzed using Guttman scaling. If items follow a pure Guttman pattern, they have a hierarchical relationship; that is, they have a specific order in terms of increasing difficulty (i.e., higher location parameters). Katz et al. (1963) argued that the ADL items were approximately hierarchically related, with an order (in terms of difficulty): bathing, dressing, toileting, transferring, incontinence, and feeding. These items have very high Guttman coefficients of reproducibility and scalability for home care and nursing home populations (Kane & Kane, 1981; Spector, 1996; Spector & Takada, 1991).

One problem with Guttman methodology is that it assumes a deterministic, not a stochastic, model. In other words, the ICC in the Guttman model specifies that, below a certain difficulty level, a person has a zero probability of a positive response to the item, and, above that level, a positive response is certain. The absence of a random error component is unrealistic. Another problem is that the Guttman model presumes a specific ordering of items. However, several different orderings of items can produce satisfactory Guttman scale coefficients of reproducibility and scalability (Lazaridis et al., 1994). Finally, it is often unclear how to score respondents who have imperfect scale patterns. For example, Katz developed his ADL scale by observing the patterns of items, combining a number of different patterns into a single level of disability, and developing an “other” category for patterns that were not pure scale types. Often researchers use high Guttman scale coefficients to justify ordering people by counting the number of tasks with limitations, despite the fact that a number of response patterns do not conform to the pure Guttman pattern.

The hierarchical relationship between IADLs and ADLs has received less attention than relationships among ADLs or among IADLs, and Guttman scaling has been used to address this issue. Lawton and Brody (1969) suggested that IADLs were more complex than ADLs, implying that these items would consistently have a lower location on a disability scale. Spector et al. (1987), using Guttman scaling, found that two IADLs, “shopping” and “transportation,” represented a lower level of disability relative to ADLs. The present study, which used a much larger number of IADLs than Spector et al. (1987), demonstrates that IADLs and ADLs overlap in terms of item location parameters. For example, “getting around outside” and “preparing meals” were located between “bathing” and “dressing.” Suurmeijer et al. (1994) and Kempen, Myers, and Powell (1995) found similar overlap with IADL and ADL items. Thus, IADL and ADL items do not necessarily have a strict hierarchical relationship.

Implications for Eligibility Determination

States are using functional disability measures to determine who is eligible for long-term care and to determine triggering criteria for the insured event in long-term care policies. Although approaches vary, essentially all states use activities of daily living, but many do not include instrumental activities of daily living criteria (U.S. General Accounting Office, 1994). Using a scale that includes both ADLs and IADLs may enable more precise assessment of the level of functional disability of applicants than using ADLs alone. This would also affect who would be eligible for these programs.

We compared an eligibility criterion based on the IADL/ADL scale to an ADL-only criterion, with the constraint that roughly the same number of persons would be eligible under both criteria. Although the equal number criterion cannot be precisely met, a close comparison is between a cutpoint of 4 or more ADL deficits using an ADL only scale and 9 or more deficits on the ADL/IADL scale. With this comparison, the eligible populations are 735 and 713, respectively. By considering both ADLs and IADLs when scoring disability, the IADL/ADL scale cutpoint of 9 or more deficits would include 86 persons with less than 4 ADLs who have at least 6 IADL deficits. It excludes 108 persons with four or more ADLs who have few IADL deficits. Thus, in about 15% of these cases, the ADL only criterion and the IADL/ADL criterion lead to discrepant conclusions. (Of course, this discrepancy would depend on the specific ADL cutpoint.) The extent of this misallocation can be viewed by looking at the mean hours of ADL and IADL help received. Those eligible using the ADL only criterion, but not eligible with the ADL/IADL criterion, averaged 45 hours of care; those eligible with the ADL/IADL criterion, but not eligible with the ADL only criterion, averaged 51 hours of care. Those eligible under both criteria averaged 78 hours of care; those not eligible averaged 18
hours. As expected, the ADL/IADL scale, rather than the ADL scale only, more effectively targets persons who would receive more hours of ADL and IADL care.

**Future Research**

We have demonstrated the feasibility and validity of combining ADLs and IADLs into one scale for a representative sample of functionally disabled elderly persons living in the community. It is important to remember that the scaling properties are dependent on the exact items and definitions of disability used. An important research question is the degree to which different definitions of disability affect scale properties. In this study the ADL items measured the receipt of human help, including receipt of standby help. Other studies have included use of mechanical help in defining ADL disability; still others present definitions that exclude both standby and mechanical help (Leon & Lair, 1990; Manton, Stallard, & Corder, 1995; Stone & Murchaugh, 1990).

Other surveys and researchers define disability in terms of respondent reports of difficulty in performing tasks, and they may or may not separate out those who receive human help. Using a definition in terms of difficulty may introduce variation due to respondents’ different conceptions of what is “difficult”; on the other hand, such a definition has the advantage of extending the range of measurement of disability by including persons who have difficulty but do not need human help. In contrast, a definition reflecting received help restricts the range of disability measured, and it confounds aspects of the person’s support network with disability. Ideally, a measure of disability based on human help should be defined in terms of need rather than in terms of receipt, so that it avoids a confound with availability of caregivers. However, no consensus exists on the proper way to measure need. Sensitivity analyses are necessary to assess how changes in wording, definitions of disability, and response scales alter IRT item parameters.

IRT analyses can identify points along the range of θ in which additional items would improve the precision of the scale. “Gaps” between item location parameters indicate such points. Future efforts need to improve the scale by developing items to address the gaps identified in the analyses, including, at the low end of the scale (between “shopping” and “laundry”), at the moderately high end (between “toileting” and “telephoning”), and at the high end of the scale (between “toileting” and “telephoning”/“feeding”/“help with incontinence”). Future research should also extend the range of the scale by developing items to measure less severe functional disability.

Future research should examine whether item locations are independent of demographic and impairment characteristics. For example, comprehensive analyses of possible differential item functioning by gender, which cross-validate the DIF analyses in the present article, would be valuable. Future DIF analyses should also examine if item parameters are sufficiently constant across the age spectrum to justify using the scale to compare elderly persons with younger persons. Further, the present research needs to be replicated for other populations, such as the under-65 disabled, to assess whether the scale provides a common measure of disability across all impairment groups. A study of the Functional Independence Measure, a measure commonly used in medical rehabilitation, indicated that there is little item bias for motor items except for burns and back pain groups (Heinemann, 1993). The extent that these results apply to ADL/IADL scales must be determined.

**Acknowledgements**

A version of this paper was presented at the 123rd Annual Meeting of the American Public Health Association, San Diego, November 1995. The views expressed here are those of the authors and no official endorsement by the Department of Health and Human Services or the Agency for Health Care Policy and Research is intended or should be inferred. We thank Liliana Pezzin for her helpful comments on an earlier draft of this article.

Address correspondence to William D. Spector, Center for Organization and Delivery Studies, Suite 500, Agency for Health Care Policy and Research, 2101 E. Jefferson St, Rockville, MD 20852. E-mail: wspector@ahcpr.gov

**References**


