THE GERONTOLOGICAL SOCIETY OF AMERICA®

Journals of Gerontology: Social Sciences
cite as: *J Gerontol B Psychol Sci Soc Sci*, 2017, Vol. 00, No. 00, 1–9
doi:10.1093/geronb/gbw209
Advance Access publication February 1, 2017

OXFORD

Original Article

Gendered Expectations Distort Male–Female Differences in Instrumental Activities of Daily Living in Later Adulthood

Connor M. Sheehan[1] and Elliot M. Tucker-Drob[2]

[1]Population Research Center and Department of Sociology and [2]Population Research Center and Department of Psychology, University of Texas at Austin.

Correspondence should be addressed to Connor M. Sheehan, Population Research Center, 305 E. 23rd Street, Stop G1800, CLA 2.602, Austin, TX 78712-1699; E-mail: connor.sheehan@utexas.edu

Received June 27, 2016; Editorial Decision Date December 8, 2016

Decision Editor: J. Scott Brown, PhD

Abstract

Objectives: The ability of older adults to live independently is often assessed with a battery of questions known as Instrumental Activities of Daily Living (IADLs). Many of these questions query the difficulty conducting household activities that have been predominantly conducted by women (e.g., the ability to prepare a meal), especially for cohorts now in old age. Although previous research has documented gender differences in IADL limitations, it has not been documented whether IADLs equivalently measure the same latent construct for men and women.

Methods: We apply psychometric tests of measurement invariance to data from the 1998 Health and Retirement Study. We then estimate corrected models that account for violations of measurement invariance across genders.

Results: We find that IADLs do not equivalently measure same latent construct for men and women. We find that men are more likely not to do the IADL activities for reasons unrelated to health limitations, which may reflect gendered expectations regarding household activities. Accounting for this we still find that women report greater health-related IADL limitations than men.

Discussion: Researchers should be cautious making gender comparisons for IADLs without attending to the gender-specific measurement properties of many of the items of which the IADL is comprised.

Keywords: Disability—Gender—Independent living—Measurement

The United States is experiencing population aging, as the population is increasingly comprised of adults entering or living into old age. Recent projections indicate that by 2030 about one in five Americans will be older than 65 (Vincent & Velkoff, 2010). This population aging has profound implications. Aging-related declines in physical health and cognitive functioning may restrict the ability to live independently (Thomeer, Mudrazija, & Angel, 2016; Tucker-Drob, 2011). Long-term care alternatives to independent living are costly, associated with lower subjective assessments of quality of life, and can place stress on the family (Kane, 2001; Whitlatch, Schur, Noelker, Ejaz, & Looman, 2001). In response to this growing public health and economic concern, there is a large body of scientific research regarding the social, biological, psychological, and demographic factors associated with the ability to live independently, generally measured and conceptualized as Instrumental Activities of Daily Living (IADLs; Bell-McGinty, Podell, Franzen, Baird, & Williams, 2002; Crimmins & Saito, 2000; Katz, 1983; Manton & Gu, 2001; Melvin, Hummer, Elo, & Mehta, 2014; Nagi, 1991; Pudaric, Sundquist, & Johansson, 2003; Verbrugge & Jette, 1994). The quality of the inferences drawn in such research, however, depends on the employment of valid instruments for assessing independent living that can be implemented across the entire range of the population, particularly with respect to key

subgroups that are compared. Differences in the measurement properties of these instruments across subgroups have the potential to not only introduce imprecision, but also systematically and directionally bias estimates.

The current study tests the measurement properties of a commonly employed battery of questions that has been used for decades to evaluate independent living: the IADL questionnaire (Katz, 1983). We apply latent variable modeling methods to data from the Health and Retirement Study (HRS), to test whether the IADL questionnaire equivalently measures the same latent construct for men and women. We then examine how correcting for gender differences in the measurement properties of the IADL questionnaire and gender differences in the reports of not doing the activity for non-health reasons influences the substantive conclusions of health related gender differences in IADL limitations.

IADLs are a battery of questions that evaluate if cognitive or physical health problems impede the ability to: use a map, manage money, prepare meals, shop, use a phone, or take medicine. IADL limitations are thought to occur relatively late in the disability process and are associated with admission into long-term care (Thomeer, Mudrazija, & Angel, 2016; Verbrugge & Jette, 1994). Although the ability to conduct these tasks is generally critical for a household to function, the expectation of who fulfills each of these household tasks has historically been gendered, especially for cohorts that are now in old age (Becker, 1985; Press & Townsley, 1998; Szinovacz & Harpster, 1994). Given the gendered division of household labor, we hypothesize that some specific activities queried by the IADL battery may be poorer indices of overall everyday living for men or women who are less likely to regularly preform the queried activities.

We anticipate greater expectation for males to use a map and manage money and greater female expectations to prepare meals and shop (we show this visually in Supplementary Figure 1). We anticipate less gendered expectations for using a phone and taking medication. Despite our hypothesized differences in expectations for activity fulfillment, the battery of IADL questions are almost always combined to implicitly or explicitly measure the latent ability to function independently. However, combining the activities does not account for the differential "exposure to risk" of conducting the activities, which we argue, varies by gender. Research that documents gender differences in IADLs without explicit analysis of and correction for unequal measurement of IADLs risks having systematically biased estimates of gender differences in IADLs. Thus we test for measurement invariance by gender, or more simply if IADLs measure the same construct for men and women. Importantly, HRS respondents can also indicate that they do not do the activity for reasons aside from health or cognitive problems, allowing us to explore how reports of not doing the activity influences estimates of health related gender differences in IADLs.

To understand how this can be problematic, consider a hypothetical scenario in which a researcher creates a sum-score of dichotomous responses (1 = unable to perform activity due to health issues/ 0 = able to perform activity and those who do not do the activity for non-health reasons coded as missing) to the IADL questions and compares these scores between men and women. If otherwise highly independent men, but not women, have a systematic tendency to indicate that they do not prepare meals and thus are excluded from the analysis, then the remaining men will have higher levels of limitations than the actual population. Latent variable models can formally evaluate whether patterns such as this exist and correct for them.

Literature Review

Previous research has compared men and women in their levels or risk of impairments in IADLs. When men are compared to women, women report more health related impairments in IADLs and are also at higher risk of reporting new health related IADL impairments in the United States (Melvin, Hummer, Elo, & Mehta, 2014; Murtagh & Hubert, 2004; Peek & Coward, 1999), and in countries of Europe (Crimmins et al., 2010). We build on this research by comparing men and women in IADL limitations while also estimating how reports of not doing the activity for reasons aside from health influence gender differences in health related IADL limitations.

Other work has indeed analyzed how gender roles may influence conclusions regarding IADLs. For instance, previous researchers found that among 629 advanced cancer patients 80% of males and 30% of females attributed help with IADLs to the traditional division of labor, rather than heath-related impairments. After adjusting for the male patients need for help with female-associated tasks, males level of need was reduced (Allen, et al. 1993). Other researchers analyzed differential item functioning with respect to item difficulty in a combined measure of Activities of Daily Living (ADLs) and IADLs using the National Health Interview Survey. They concluded there was "especially large" differential item functioning for two items: managing money and shopping (Fleishman, Spector, & Altman, 2002), two items that we also hypothesize may exhibit measurement non-invariance across gender. Once differential item functioning (DIF) was accounted for they found that the within age group gender differences were reduced. Although not focused explicitly on gender differences other research combined ADLs and IADLs and compared DIF in the HRS to European surveys (Chan, Kasper, Brandt, & Pezzin, 2012). We build on these studies by analyzing multiple specifications of IADLs, by using tests for measurement invariance in not just difficulty but also discrimination parameters, and by illustrating how not accounting for certain responses may lead to biased estimates of the gender differences. We did not include ADLs in this analysis as impairments in ADLs are generally indicative of more severe health problems than are impairments in IADLs, other work analyzes ADLs and IADLs separately

(Crimmins et al., 2010; Melvin et al., 2014; Montez & Hayward, 2014), and, we had no a prioi reason to assume that activities such as being able to use the toilet or getting dressed, differed due to gender roles in a manner similar to that hypothesized for IADLs.

Data & Methods

Data

We used the HRS to accomplish our objective of testing if IADLs equivalently measure the same latent construct for men and women, and adjusting the measure accordingly to reestimate the gender differences. The HRS (Health and Retirement Study, 2016) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. It contains detailed information regarding older Americans health measured longitudinally, is nationally representative of persons aged ≥51, and follows respondents even if they enter long-term care. The IADL measures from the HRS have also been used in previous research (Berger & Porell, 2008; Bowen, 2012). We used the fourth or 1998 wave of the HRS. In this wave new cohorts entered, making the HRS generalizable to all Americans aged ≥51. This wave also is less negatively influenced by mortality selection and attrition compared to later waves (Zajacova & Burgard, 2013). To be consistent with previous research that has compared men and women prevalence rates of IADLs we use the HRS as cross-sectional data. This did, however, preclude us from separating age and cohort effects. We used the Rand "n" file, which is cleaned and tested for consistency by Rand Corporation (Rand 2014). Our sample of 20,218 is limited to those who were included in Wave 4, aged ≥51 at interview at Wave 4, and those who provided valid information regarding their gender and did not have missing responses for *all* IADL questions.

There was generally little missing data. However, approximately 6% of the respondents indicated that they did not take medication. An argument could be made for coding these respondents as not limited based on a hypothetical follow up question. However, we maintained the Rand coding scheme, but when we coded individuals replying "no" to this hypothetical question as "no difficulty," model parameter estimates were exceedingly similar to those reported here (Supplementary Table 1). Aside from that there were very little missing data. For example, only 28 respondents were dropped for missing all the questions. For participants with partial IADL data (<1%), missing responses were handled with Full Information Maximum Likelihood Estimation (Muthén & Muthén 2010). For a more detailed and technical overview of our sampling criteria, methods, missing data, and results please see the Supplementary Material.

Measures

We used all six IADL of the questions asked in the HRS. As discussed above respondents were asked if they had difficulty: using a map, using a telephone, managing money, taking medication, shopping for groceries, and preparing hot meals. For all the questions the respondents were able to select the following responses (0) if they had no difficulty doing the activity, (1) if they had difficulty doing the activity due to health reasons, (2) if they could not do the activity due to health reasons, (3) if they did not do the activity, and then were asked a follow up question to determine if (3) if they did not do the activity for reasons aside from health impairment ("don't do"). Consistent with Rand coding, we combined (1) and (2), into a single health limited category.

We primarily coded IADLs in two ways and estimated measurement invariance tests for each specification. First, to examine gender differences in not doing the activities we created a latent construct that measured not doing the activity for reasons aside from health. For this coding scheme, we coded do not do the activity for reasons aside from health ("don't do," (3) from above) "1" and all the other responses (except missing) as "0." We refer to this as the *Gender Role Specification*. Rather than representing IADL capability or impairment, this factor represents a normative tendency for an individual not to engage in IADLs, for social, lifestyle, or patriarchal, rather than health-related reasons. This specification is critical because if men have a significantly higher average on this latent factor it indicates that they are *systematically* less likely to do the queried activities than women. Second, we employed a coding scheme that mirrored previous research interested in health limitations. This scheme combined the some difficulty (1) and can't do (2) responses into a single category coded "1," the have no difficulty responses into another category coded "0" and coded the don't do (3) and other missing responses as missing. We refer to this as the *Traditional Health Specification*. To gauge the bounds of potential bias for the *Traditional Health Specification,* we also conducted sensitivy analysis where we assumed all the "don't do" (3) responses to be unimpaired and then assumed they were impaired.

Methods

We began with descriptive tables that depict the distribution of the responses of each of the IADL questions by gender. Next, we fit common factor models to the IADL items and employed tests of measurement invariance based on techniques developed by previous researchers (Milsap & Yun-Tein, 2004; van de Schoot, Lugtig, & Hox, 2012). Testing measurement invariance entailed the estimation of a series of nested models where sets of parameters representing item "difficulty" item "discrimination" are increasingly set to be equal between males and females. If setting the parameters to be equal reduces fit significantly, this provides evidence that males and females are not being measured equally on the latent construct. More simply, if parameters are set to be equal and the model fit becomes significantly worse it indicates that setting the parameters between males and females to be equal does not fit the data

as well as allowing the parameters to be different by gender. Conversely, if the model fit is not significantly worse when the parameters are constrained between genders it suggests that men and women are measured equivalently enough that the same set of parameters can be used in both groups.

We employed techniques developed for dichotomous responses (Milsap & Yun-Tein, 2004). The first model is known as the configural model. To begin with a mathematically identified model, we set the residual variance of the items to be equal for men and women. The thresholds of the latent factors were freely estimated for males and females. The factor mean was fixed at 0 and factor variance was fixed to 1 for females, but was freely estimated for males, such that gender differences in factor means and variances are represented by the factor mean and variance estimates for mean in standardized-units relative to females. In the second model, we tested weak invariance by setting all the loadings to be equal for males and females. In the third model, we tested for strong invariance by additionally setting all the thresholds to be equal for males and females.

To evaluate the model fit we used the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error Approximation (RMSEA), for an overview please see (Kline, 2015). For the nested models, we compared the models using chi-squared tests from the difftest function of Mplus, where a significant value indicates statistically significantly *worse* fit relative to the previous model where more parameters were freely estimated. After the freely estimated models were compared to the models where men and women were fully set to be equal, we went onto free the most problematic individual parameters from the fully constrained model, based on modification indices. Modification indices indicate the chi-square unit change in model fit improvement that would be achieved by freely estimating a previously fixed parameter, thus modification indices provide insight at the gender differences in the items. This procedure allowed us to document what the mean and variance are for men and women are if we assumed that IADLs equivalently measure the same construct for men and women, and then reestimate these values after correcting for violations of measurement invariance. For all the specifications and models, we used Mplus's WLSMV estimation with theta parameterization.

Results

Table 1 presents descriptive statistics of the sample. The average age was relatively similar between men (66.5) and women (67.4). Appreciable gender differences can be seen in the response distributions for a number of the IADL items in directions consistent with those we hypothesized earlier. Almost 18.7% of women indicated that they do not use a map compared to only 6.7% of males. Conversely, 2.2% of females reported that they do not shop whereas the comparable percentage of men is more than twice as high (5.3%). While 10.5% of males indicated they do not prepare meals only 1.5% of women reported that they do not prepare meals. Some of the IADLs had similar response distributions for males and females. Specifically, using a telephone and taking medication are activities we hypothesized to have little differences in gendered expectations. The distribution of these activities were similar between males and females, particularly those who reported not doing those items. For health-related differences in IADLs, more women (13.8%) reported difficulty using a map than men (7.5%) and more women reported difficulty shopping (13.0%) than men (7.1%). Overall there was little missing data, except among taking medication (for a sensitivity analysis on those missing taking medication see Supplementary Table 1).

Table 1. Descriptive Statistics, Respodents Aged ≥ 50 Health and Retirement Study (n = 20,218), 1998

	Males	Females
Age (average)	66.5	67.4
Instrumental activities of daily living		
Self-reported difficulty using a map[a]		
No	85.7%	67.4%
Yes/can't do	7.5%	13.8%
Don't do	6.7%	18.7%
Missing	0.0%	0. 0%
Self-reported difficulty using a telephone[a]		
No	93.1%	94.6%
Yes/can't do	6.0%	5.1%
Don't do	0.9%	0.3%
Missing	0.0%	0.0%
Self-reported difficulty managing money[a]		
No	90.4%	89.9%
Yes/can't do	5.7%	7.8%
Don't do	3.9%	2.3%
Missing	0.0%	0.0%
Self-reported difficulty taking medicine[a]		
No	89.0%	90.7%
Yes/can't do	3.2%	4.5%
Don't do	0.2%	0.2%
Missing	7.7%	4.6%
Self-reported difficulty shopping[a]		
No	87.6%	84.9%
Yes/can't do	7.1%	13.0%
Don't do	5.3%	2.2%
Missing	0.0%	0.0%
Self-reported difficulty preparing meals[a]		
No	84.2%	89.9%
Yes/can't do	5.3%	8.6%
Don't do	10.5%	1.5%
Missing	0.0%	0.0%
N	8,727	11,491

Note: Source—Health and Retirement, 1998.
[a]Statistically significant chi-square test for over-dispersion by gender.

Specification 1: Gender Role Specification

Table 2 documents results from the series of nested models on different specifications of IADLs to explicitly examine

Table 2. Model Fit for Measurement Invariance Tests on Different Specifications of Instrumental Activities of Daily Living (IADLs), Respondents aged ≥ 51 Health and Retirement Study (n = 20,218), 1998

	Model	X^2	df	CFI	TLI	RMSEA	X^{2a}	X^2 p	Mean sex difference[b]	p	Male factor variance[c]	p
IADL gender role specification												
Stratified models												
Males		41	9	0.972	0.953	0.020	—	—				
Females		23	9	0.990	0.983	0.012	—	—				
Multi-group models (full results Supplementary Table 2)												
Model 1	Configural	64	18	0.982	0.969	0.016	—	—	0.25	.66	0.70	.37
Model 2	Weak	58	23	0.986	0.982	0.012	6.3	.28	0.32	.19	0.54	0
Model 3	Strong	898	27	0.653	0.614	0.056	857.0	.00	−0.27	.00	1.76	0
Model 4	Corrected	91	26	0.974	0.970	0.016			0.85	.00	0.48	0
IADL traditional health specification												
Stratified models												
Males			9	0.998	0.997	0.029	—	—				
Females			9	0.995	0.991	0.058	—	—				
Multi-group models (full results Supplementary Table 3)												
Model 1	Configural	446	18	0.996	0.993	0.048	—	—	−0.31	.05	1.18	0
Model 2	Weak	349	23	0.997	0.996	0.037	30.2	.00	−0.23	.00	1.09	0
Model 3	Strong	644	28	0.994	0.993	0.047	340.1	.00	−0.41	.00	1.26	0
Model 4	Corrected	393	26	0.996	0.996	0.037			−0.47	.00	1.38	0

Note: CFI = Comparative Fit Index; RMSEA = Root Mean Square Error Approximation; TLI = Tucker-Lewis Index.
[a]Relative to previous model. [b]Male mean relative to females, in female *SD* units. [c]Relative to female factor variance (ratio).

measurement invariance. In the *Gender Role Specification* of IADLs, we analyzed a factor of not doing IADLs for reasons aside from health limitations (for full results see Supplementary Table 2). The overall model fit was slightly better for females than males, but both fit well. The multi-group the configural model also had excellent fit (RMSEA = 0.016, CFI = 0.982, TLI = 0.969). In the configural model, there were large differences between the factor loading of shopping and preparing meals, and larger differences among the thresholds of using a map, using a telephone, and preparing a meal. In this model, the difference in mean was not significant between men and women. In the next model, the factor loadings were set to be equal and the model fit improved slightly, however the differences between models was not statistically significant (X^2 = 6.3, df = 5), suggesting the loadings are comparable. When we constrained the thresholds, the model fit significantly worse than the previous model where they were freely estimated (X^2 = 858, df = 4, p < .01). These results suggest that the factor measuring not doing the activity does not equivalently measure the same latent construct for men and women. In the final model males had a *lower* mean (–0.27, p < .01) and higher variance (1.76, p < .01) than females.

We corrected for violations the measurement invariance using the modification indices. Modification indices were high, as not doing the IADL activities varied substantially by gender; we set the modification index cutoff at 50. Given our large sample size, we chose this cuff off on an a priori basis in order to ensure that the differences identified were not simply statistically significant (a modification index > 3.84 is statistically significant), but also practically significant in terms of size of effect. Based on modification indices we sequentially freed the threshold for not using a map, the residual variance for using a map, and the threshold for preparing meals. These modification indices are all in activities that we hypothesized to have gender role differences. When these were freed the model fit improved substantially, as the corrected model fit the best of all the models for the gender role specification (RMSEA = 0.016, CFI = 0.985, and TLI = 0.981). In line with our discussion of the gendered distribution of household labor, men had a significantly higher mean (0.85, p < .05) and substantially lower variance (0.48, p < .05) of the factor, substantively meaning that men do less of the queried activities measured by IADLs than women. The notion that men do significantly fewer of the IADLs for reasons aside from health suggests that estimates of gender differences in IADLs are likely biased by the lack of exposure to risk of doing these activities by men.

Specification 2: Traditional Health Specification

The second specification of IADLs analyzed the health limitation based specification of IADLs. For this specification, the gender stratified models fit slightly better for males (RMSEA = 0.029) than for females (RMSEA = 0.058), but both models fit reasonably well. The configural model fit well (RMSEA = 0.048, CFI = 0.996, TLI = 0.993; for full results please see Supplementary Table 3). When we set the loadings to be equal to test weak invariance, the constraint

did not lead to worse model fit suggesting that the loadings are invariant. In the final model, we additionally constrained the thresholds to be equal. The model fit became significantly worse (X^2 = 340.1, df = 5, p < .01). In this model, men still had a lower mean on the factor (–0.41, p < .01) and a higher variance (1.26, p < .01).

We then corrected for measurement invariance in the traditional health based specification. We used the same cutoff (50) as the first specification. Based on the modification indices we sequentially freed the unique variance of using a phone and threshold for using a map. After freeing these parameters to differ by gender, there were no other modification estimates >50. The corrected model had excellent fit (RMSEA = 0.037, CFI = 0.996, TLI = 0.996). Consistent with previous research, in this model males had lower levels of the latent construct of IADL limitations than females, (–0.47, p < .01) and had greater variation (1.38, p < .01). In other words, females had higher levels of IADL health based limitations than men. The differences in means is greater for the corrected model than the fully constrained model, suggesting that not correcting for measurement invariance may lead to *underestimates* of gender differences in health related IADLs limitations.

Sensitivity Analyses: Traditional Health Specification

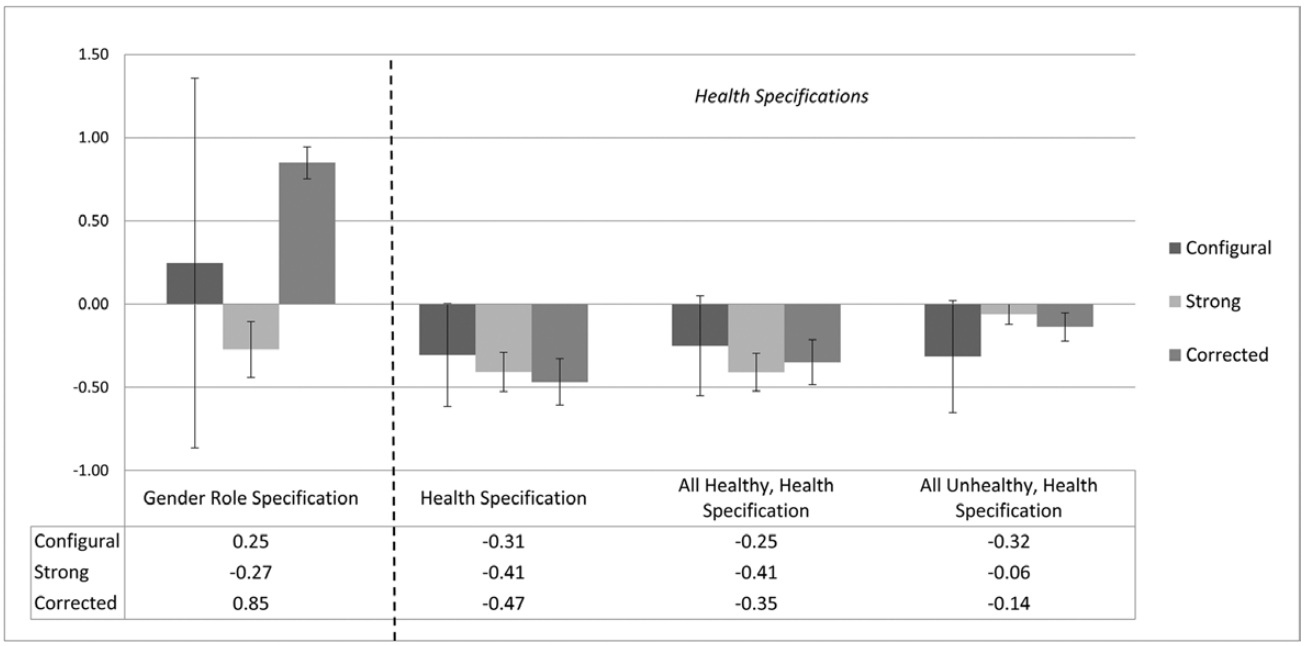
We further tested two other counter-factual specifications of the health limitation measure to understand the bounds of influence that reporting not doing the activity may have on estimates of gender differences in IADLs (For full results see Supplementary Table 4). First, we assumed all of those who reported not doing the activity for non-health issues as if they had no health problems doing the activity. In this specification the fully constrained model had good fit (RMSEA = 0.047, CFI = 0.991, TLI = 0.991) and males had a significantly lower mean level of the IADL factor (–0.40, p < .01), and higher variance than women (1.26, p < .01). However, some of the parameters had modification indices above our cutoff. When we fit models where we progressively freed the residual variance for using a phone, then the residual variance for using a map, no more parameters were >50. The corrected model had excellent fit (RMSEA = 0.038, CFI = 0.995, TLI = 0.995) and males still had a lower factor mean (–0.35, p < .01) and higher variance (1.17, p < .01). The smaller gender difference than the corrected *Traditional Health Specification* is likely due to the substantial proportion of women who reported not using a map having been coded as healthy.

We next assumed all of the "do not do" responses as if they were limited doing the activity due to health problems. The constrained model had poor fit (RMSEA = 0.09, CFI = 0.965, TLI = 0.962), and males had a slightly lower factor mean (–0.053, p < .01) and slightly lower variance (0.919, p < .01). Based on the modification indices we progressively freed the threshold for not using a map, the residual variance for preparing a meal, then the residual variance for using a phone. The corrected model had improved fit (RMSEA = 0.043, CFI = 0.993, TLI = 0.992), and males had a slightly lower mean (–0.14, p < .01), and higher variance (1.12, p < .01) than females. Likely as a results of men disproportionately reporting not shopping and preparing meals, coding all the do not do reports as unhealthy equalizes the difference in IADLs between men and women compared to the other health specifications. We show all the gender differences in means and variances in each specification in Figures 1 and 2. Overall, these results depict the potential biases that arise from assumptions on how to handle the reports of not doing the activity, but are consistent with previous research suggesting women have greater levels of limitations.

Discussion

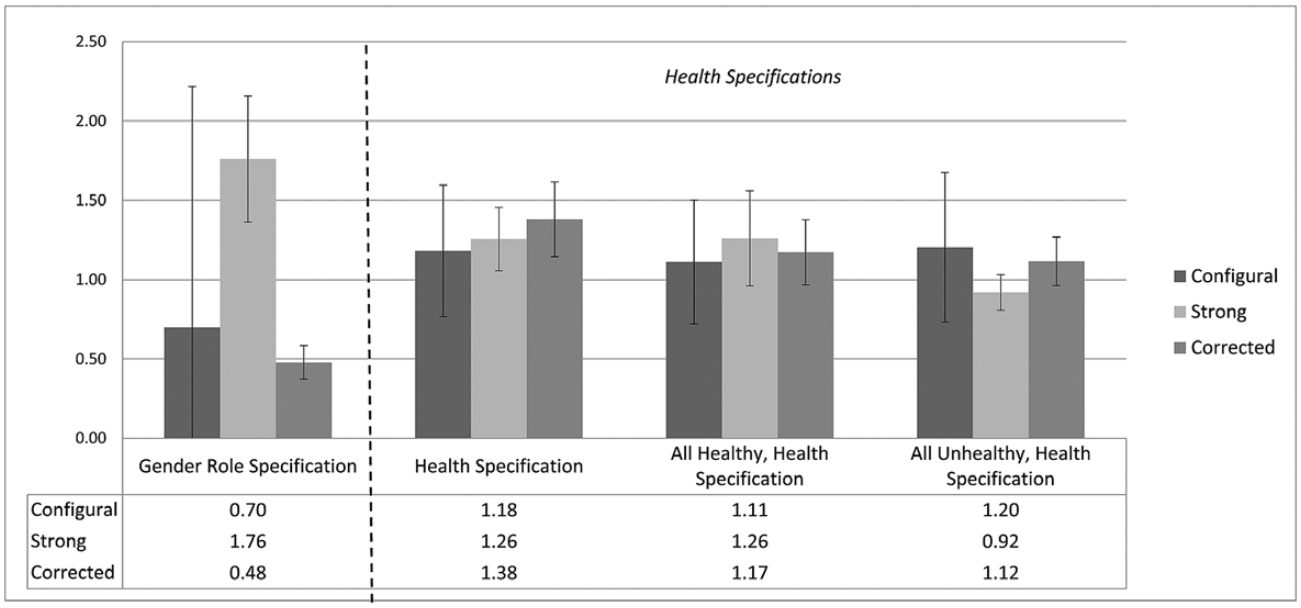
Our aim was to analyze if the battery of questions comprising the IADL questionnaire equivalently measured independent living for men and women and, if not, correct for any misfit resulting from differences in gender based measurement properties. Since some of the activities that comprise IADLs have expectations of fulfillment that differ by gender, we anticipated that IADLs would not equivalently measure the same factor for men and women. Indeed, we found strong evidence supporting the notion that men and women are not equivalently measured on IADLs. Most simply, when men and women had the factor loadings, thresholds, and residuals set to be equal the model fit became significantly worse, substantively meaning that constraining the items to be equal by gender did not closely approximate the actual data. As predicted, the poor fit was largely due to the items that have more traditional gendered expectations of fulfillment: using a map, preparing food, and going shopping. When the measurement models were corrected to allow for these measurement differences across gender, the model fit improved and the inferences regarding gender differences in mean IADL performance and variation in IADL performance changed (see Figures 1 and 2).

Our two specifications of IADLs provide insight in gender differences in health related IADL limitations and what IADLs can measure. The first *Gender Role Specification* analyzed not doing the activity for reasons aside from health. After correcting for problematic parameters, we found men had significantly higher means of the gender role factor, suggesting that men were significantly more likely to *not* do the IADLs for reasons aside from their health. This suggests that there are gender differences in who does and does not do IADLs, and is consistent with the notion that men conduct less housework than women. The excellent model fit of the corrected model indicates that the IADL questions can be used to construct a latent construct of household labor. Unfortunately, this also suggests that not accounting for the systematically lower levels



	Gender Role Specification	Health Specification	All Healthy, Health Specification	All Unhealthy, Health Specification
Configural	0.25	-0.31	-0.25	-0.32
Strong	-0.27	-0.41	-0.41	-0.06
Corrected	0.85	-0.47	-0.35	-0.14

Figure 1. Differences in mean, men relative to women in instrumental activities of daily living (IADLs) latent factors, Health and Retirement Study. Error bars represent 95% confidence Interval. Estimates indicate difference in the Mean of the latent variable in standard deviation units for males relative to females.



	Gender Role Specification	Health Specification	All Healthy, Health Specification	All Unhealthy, Health Specification
Configural	0.70	1.18	1.11	1.20
Strong	1.76	1.26	1.26	0.92
Corrected	0.48	1.38	1.17	1.12

Figure 2. Differences in variance, men relative to women in instrumental activities of daily living (IADLs) latent factors, Health and Retirement Study. Error bars represent 95% confidence Interval. Estimates indicate difference in the male factor variance relative to female factor variance (ratio).

of housework conducted by men likely biases gender differences in IADLs health based limitations.

Our second specification was the *Traditional Health Specification* of IADLs. The fully constrained model reflects the inferences that would be made under a naïve approach, which relies on a strong assumption of no gender differences in measurement properties of the individual items. Consistent with previous research, men had lower mean levels of health related IADL difficulty than women. However, when we freed the most problematic parameters, we found that men had *lower* levels of IADL impairment relative to women. In other words, previous research, which has implicitly or explicitly assumed IADLs measure the same latent construct for men and women, and found men have less IADL limitations than women may have *underestimated* or been conservative to the actual difference.

Sensitivity analyses on the health-based specification further highlight the perils of assuming not doing the

activity to be diagnostic of either health-related impairment, or lack thereof. When we assumed that all those who reported they do not do the activity were healthy, the mean gender differences were *less* than in the corrected model, largely due to the substantial proportion of females who report not using a map. Assuming all those who reported not doing the activity were impaired also underestimates the difference between men and women because of the large proportion of men who report not shopping or preparing meals. Although our results suggest that researchers should be cautious when making gender comparisons for IADLs, we did find similar substantive results to previous research (e.g., Crimmins et al., 2010) in our health specification and sensitivity analysis—females have greater levels of health-related limitations of IADLs than males even after accounting for measurement invariance and gender differences in reports of not doing the activities. The greater level of IADL health-based limitations among women is consistent with the higher number of chronic conditions faced by elderly women (Case, 2005).

Our results suggest a number of possible avenues for avoiding bias associated with gendered expectations in IADLs. First, one could exclusively use measures such as ability to take medication and use a telephone, while excluding the more gendered items. This of course has the disadvantage of losing potentially important information, so researchers could compare results from the full specification to the abbreviated specification. Researchers could also employ all available data in latent models and follow our protocol to freely estimate the most problematic items individually for men and women.

It is useful to discuss the limitations of this work. Our results are only generalizable to the population sampled (U.S. Residents aged ≥51 in 1998) and IADL questions employed by the HRS. This limits the generalizability of our results if household labor becomes more egalitarian in younger cohorts. Future research should examine how shifting household labor influences gender differences in IADLs, or at minimum be wary of cohort shifts in gender differences in IADLs. By examining the entire HRS sample our results are likely more conservative than if we had just limited our sample to married heterosexuals. Our selected modification index cutoff was arbitrary, however the other modification indices in the corrected models were well below the cutoff. Finally, setting data from participants who do not report doing the activity to missing in the second specification does rely on the assumption that these individuals' missing scores on whether or not their health would have caused them limitations in performing the tasks (should they have been inspired to perform them) are missing at random, conditional on their scores on the other items. If individuals who engage in more traditional gendered divisions of household labor are at heightened risk or resilience to IADL impairment beyond what we would expect on the basis of their degree of impairment on the non-gendered IADL items, then our results would be biased. That said, we have no reason to suspect this to be the case. We are also reassured that our sensitivy analysis, which assumed all missing due to not doing the activity were either limited and healthy, produced similar substantive results.

Overall, we suggest that researchers should carefully evaluate and consider the strengths and weaknesses of the measures they select to conceptualize health. Specific batteries may have substantially different meanings and/or levels of exposure to risk that can vary systematically by population subgroup. With a focus on reliable and valid measures, scientists will be better equipped to understand changes in population health.

Supplementary Material

Supplementary data are available at *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences* online.

Funding

This work was supported by the University of Texas Population Research Center who provided administrative and computing support; the National Institute of Child Health and Human Development Ruth L. Kirschstein National Research Service Award for training support (grants R24 HD42849, T32 HD007081-35); the University of Michigan and Rand Corporation for making the data available to the public.

Acknowledgments

We thank the members of the Population Health and Structural Equation Modeling lab for their helpful suggestions. We also thank Frank Mann, Mark Hayward, and Mijke Rhemtulla for their help. C. Sheehan conducted most of the data analysis and writing with supervision and input from E. M. Tucker-Drob. The contents of this manuscript are solely the responsibility of the authors and do not represent the official views of NICHD, Rand Corporation, or the University of Texas at Austin.

Conflict of Interest

Despite these grants we have no financial conflicts of interest.

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*, **48**, S204–S211. doi:http://dx.doi.org/10.1093/geronj/48.4.S204.

Becker, G. S. (1985). Human capital, effort, and the sexual division of labor. *Journal of Labor Economics*, **3**, S33–S58. doi:http://dx.doi.org/10.1086/298075

Bell-McGinty, S., Podell, K., Franzen, M., Baird, A. D., & Williams, M. J. (2002). Standard measures of executive function in predicting instrumental activities of daily living in older adults.

International Journal of Geriatric Psychiatry, **17**, 828–834. doi:http://dx.doi.org/10.1002/gps.646

Berger, S., & Porell, F. (2008). The association between low vision and function. *Journal of Aging and Health*, **20**, 504–525. doi:http://dx.doi.org/10.1177/0898264308317534

Bowen, M. E. (2012). The relationship between body weight, frailty, and the disablement process. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, **67**, 618–626. doi:http://dx.doi.org/10.1093/geronb/gbs067

Case, A., & Paxson, C. (2005). Sex differences in morbidity and mortality. *Demography*, **42**, 189–214. doi:http://dx.doi.org/10.1353/dem.2005.0011

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

Crimmins, E. M., Kim, J. K., & Solé-Auró, A. (2010). Gender differences in health: Results from SHARE, ELSA and HRS. *The European Journal of Public Health*, **21**, 81–91. doi:http://dx.doi.org/10.1093/eurpub/ckq022

Crimmins, E., & Saito, Y. (2000). Change in the prevalence of diseases among older Americans: 1984–1994. *Demographic Research*, **3**, 1–20. doi:http://dx.doi.org/10.4054/DemRes.2000.3.9

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, **57**, S275–S284. doi:http://dx.doi.org/10.1093/geronb/57.5.S275

Health and Retirement Study. (2016). *Rand N File. Public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740)*. Ann Arbor, MI: University of Michigan.

Kane, R. A. (2001). Long-term care and a good quality of life bringing them closer together. *The Gerontologist*, **41**, 293–304. doi:http://dx.doi.org/10.1093/geront/41.3.293

Katz, S. (1983). Assessing self-maintenance: Activities of daily living, mobility, and instrumental activities of daily living. *Journal of the American Geriatrics Society*, **31**, 721–727. doi:http://dx.doi.org/10.1111/j.1532–5415.1983.tb03391.x

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.

Manton, K. G., & Gu, X. (2001). Changes in the prevalence of chronic disability in the United States black and nonblack population above age 65 from 1982 to 1999. *Proceedings of the National Academy of Sciences*, **98**, 6354–6359. doi:http://dx.doi.org/10.1073/pnas.111152298

Melvin, J., Hummer, R., Elo, I., & Mehta, N. (2014). Age patterns of racial/ethnic/nativity differences in disability and physical functioning in the United States. *Demographic Research*, **31**, 497–510. doi:http://dx.doi.org/10.4054/DemRes.2014.31.17

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, **39**, 479–515.

Montez, J. K., & Hayward, M. D. (2014). Cumulative childhood adversity, educational attainment, and active life expectancy among US adults. *Demography*, **51**, 413–435. doi:http://dx.doi.org/10.1007/s13524-013-0261-x

Murtagh, K. N., & Hubert, H. B. (2004). Gender differences in physical disability among an elderly cohort. *American Journal of Public Health*, **94**, 1406–1411. doi:http://dx.doi.org/10.2105/AJPH.94.8.1406

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide: Statistical analysis with latent variables: User's guide*. Los Angeles, CA: Muthén & Muthén.

Nagi, S. Z. (1991). Disability concepts revisited: Implications for prevention. In A. M. Pope & A. R. Tarlov (Eds). *Disability in America: Toward a National Agenda for Prevention* (pp. 309–327). Washington, DC: Division of Health Promotion and Disease Prevention, Institute of Medicine. National Academy Press.

Peek, M. K., & Coward, R. T. (1999). Gender differences in the risk of developing disability among older adults with arthritis. *Journal of Aging and Health*, **11**, 131–150. doi:http://dx.doi.org/10.1177/089826439901100201

Press, J., & Townsley, E. (1998). Wive's and husband's housework reporting: Gender, class, and social desireability. *Gender & Society*, **12**, 188–218. doi:http://dx.doi.org/10.1177/089124398012002005

Pudaric, S., Sundquist, J., & Johansson, S.-E. (2003). Country of birth, instrumental activities of daily living, self-rated health and mortality: A Swedish population-based survey of people aged 55–74. *Social Science & Medicine*, **56**, 2493–2503. doi:http://dx.doi.org/10.1016/S0277-9536(02)00284-8

Szinovacz, M., & Harpster, P. (1994). Couples' employment/retirement status and the division of household tasks. *Journal of Gerontology*, **49**, S125–S136.

Thomeer, M. B., Mudrazija, S., & Angel, J. L. (2016). Relationship status and long-term care facility use in later life. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, **71**, 711–723. doi:http://dx.doi.org/10.1093/geronb/gbv106

Tucker-Drob, E. M. (2011). Neurocognitive functions and everyday functions change together in old age. *Neuropsychology*, **25**, 368. doi:http://dx.doi.org/10.1037/a0022348

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, **9**, 486–492.

Verbrugge, L. M., & Jette, A. M. (1994). The disablement process. *Social Science & Medicine*, **38**(1), 1–14. doi:http://dx.doi.org/10.1016/0277-9536(94)90294-1

Vincent, G. K., & Velkoff, V. A. (2010). *The next four decades: The older population in the United States: 2010 to 2050*. Washington, DC: US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

Whitlatch, C. J., Schur, D., Noelker, L. S., Ejaz, F. K., & Looman, W. J. (2001). The stress process of family caregiving in institutional settings. *The Gerontologist*, **41**, 462–473. doi:http://dx.doi.org/10.1093/geront/41.4.462

Zajacova, A., & Burgard, S. A. (2013). Healthier, wealthier, and wiser: A demonstration of compositional changes in aging cohorts due to selective mortality. *Population Research and Policy Review*, **32**, 311–324. doi:http://dx.doi.org/10.1007/s11113-013-9273-x