

RESEARCH ARTICLE

Differential Item Functioning in the SF-36 Physical Functioning and Mental Health Sub-Scales: A Population-Based Investigation in the Canadian Multicentre Osteoporosis Study

Lisa M. Lix^{1*}, Xiuyun Wu², Wilma Hopman³, Nancy Mayo⁴, Tolulope T. Sajobi⁵, Juxin Liu⁶, Jerilynn C. Prior⁷, Alexandra Papaioannou⁸, Robert G. Josse⁹, Tanveer E. Towheed¹⁰, K. Shawn Davison¹¹, Richard Sawatzky¹²

1 Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, Canada, **2** School of Health and Human Performance, Dalhousie University, Halifax, NS, Canada, **3** Department of Public Health Sciences, Queen's University, Kingston, ON, Canada, **4** Division of Clinical Epidemiology, McGill University Health Centre, Montréal, QC, Canada, **5** Department of Community Health Sciences & O'Brien Institute for Public Health, University of Calgary, Calgary, AB, Canada, **6** Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, SK, Canada, **7** Medicine and Endocrinology, University of British Columbia, Vancouver, BC, Canada, **8** Department of Medicine, McMaster University, Hamilton, ON, Canada, **9** Department of Medicine, Faculty of Medicine, University of Toronto, Toronto ON, Canada, **10** Department of Medicine, Queen's University, Kingston, ON, Canada, **11** Faculty of Graduate Studies, University of Victoria, Victoria, BC, Canada, **12** School of Nursing, Trinity Western University & Centre for Health Evaluation and Outcomes Sciences, Providence Health Care, Langley, BC, Canada

* lisa.lix@umanitoba.ca



OPEN ACCESS

Citation: Lix LM, Wu X, Hopman W, Mayo N, Sajobi TT, Liu J, et al. (2016) Differential Item Functioning in the SF-36 Physical Functioning and Mental Health Sub-Scales: A Population-Based Investigation in the Canadian Multicentre Osteoporosis Study. PLoS ONE 11(3): e0151519. doi:10.1371/journal.pone.0151519

Editor: Ulrich S Tran, University of Vienna, School of Psychology, AUSTRIA

Received: May 23, 2015

Accepted: February 29, 2016

Published: March 21, 2016

Copyright: © 2016 Lix et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data have been provided under agreement with the Canadian Multicentre Osteoporosis Study (CaMos). The CaMos has developed an Ancillary Project and Data Release Policy that governs ancillary project approval and access to the data. The Design Analysis and Publications (DAP) Committee acts to oversee the selection and implementation of ancillary projects. They review each project application considering feasibility, priority and its impact on CaMos and will, after due consultation with the applicants, make a

Abstract

Background

Self-reported health status measures, like the Short Form 36-item Health Survey (SF-36), can provide rich information about the overall health of a population and its components, such as physical, mental, and social health. However, differential item functioning (DIF), which arises when population sub-groups with the same underlying (i.e., latent) level of health have different measured item response probabilities, may compromise the comparability of these measures. The purpose of this study was to test for DIF on the SF-36 physical functioning (PF) and mental health (MH) sub-scale items in a Canadian population-based sample.

Methods

Study data were from the prospective Canadian Multicentre Osteoporosis Study (CaMos), which collected baseline data in 1996–1997. DIF was tested using a multiple indicators multiple causes (MIMIC) method. Confirmatory factor analysis defined the latent variable measurement model for the item responses and latent variable regression with demographic and health status covariates (i.e., sex, age group, body weight, self-perceived general health) produced estimates of the magnitude of DIF effects.

decision regarding the acceptability of the project. Ancillary projects may be undertaken in any one of CaMos' nine regional centres, or as a collaboration among investigators, at least one of which must be a CaMos Centre Director. A formal proposal must be submitted to the DAP Committee for review. Following approval, the authors of the proposal will be notified of the release of data, and will sign an agreement, stating that they will only use the data for the purpose described, will follow the timeline specified for the analysis, and will destroy the data files by a given date. The CaMos principal investigators are Dr. David Goltzman and Dr. Nancy Krieger. To obtain a copy of the Ancillary Project and Data Release Policy and the requirements for proposal submission, or to obtain further information from the CaMos principal investigators about data access, please send an e-mail to info@camos.org.

Funding: This work was supported by funding from the Canadian Institutes of Health Research (Funding Reference #122110; website <http://www.cihr-irsc.gc.ca/e/193.html>) to LML. LML receives salary support from a Manitoba Research Chair funded by Research Manitoba (<http://researchmanitoba.ca/>). CaMos is currently supported by the Canadian Institutes of Health Research (website <http://www.cihr-irsc.gc.ca/e/193.html>) Amgen Canada Inc. (website www.amgen.ca), Dairy Farmers of Canada (website <http://www.dairyfarmers.ca/>), Merck Canada (website www.merck.ca), Eli Lilly Canada (website www.lilly.ca), and Novartis Canada (website www.novartis.ca). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Results

The CaMos cohort consisted of 9423 respondents; 69.4% were female and 51.7% were less than 65 years. Eight of 10 items on the PF sub-scale and four of five items on the MH sub-scale exhibited DIF. Large DIF effects were observed on PF sub-scale items about vigorous and moderate activities, lifting and carrying groceries, walking one block, and bathing or dressing. On the MH sub-scale items, all DIF effects were small or moderate in size.

Conclusions

SF-36 PF and MH sub-scale scores were not comparable across population sub-groups defined by demographic and health status variables due to the effects of DIF, although the magnitude of this bias was not large for most items. We recommend testing and adjusting for DIF to ensure comparability of the SF-36 in population-based investigations.

Introduction

Self-report health status measures, like the Short Form 36-item Health Survey (SF-36) [1], can provide rich information about the overall health of a population [2,3] and its components, such as physical, mental, and social health. However, in order for comparisons of health status across population sub-groups to be accurate, these self-report measures must be valid and reliable. Construct validity and test-retest reliability are frequently evaluated for a measure's summary score (s), that is, after the item responses have been summed [4]. Self-report measures are less often evaluated for the effects of differential item functioning (DIF), which can also affect construct validity [5]. DIF occurs when individuals with the same underlying (i.e., latent) level of health do not interpret a measure's items in the same way. DIF can result in an unexpected lack of scale comparability and erroneous conclusions about the presence of group differences [5].

The SF-36 has undergone comprehensive psychometric evaluations of its reliability and validity [6–9]. DIF and related topics of differential scale functioning have been investigated for the SF-12 and SF-36 [10–16], but most of these analyses have been conducted in clinical or disease-specific samples. DIF analyses are often conducted on demographic and ethnic characteristics even though other determinants of health, including risk factors for poor health and presence of chronic conditions may be potential sources of DIF [11,13,17].

The physical functioning (PF) and mental health (MH) sub-scales of the SF-36 are the sub-scales most frequently investigated in psychometric evaluations, and are also commonly used sub-scales to compare health status at the population level. Our study objective was to test for DIF on the SF-36 PF and MH sub-scale items in population-based data on demographic and health-related variables.

Methods

Data Source

Study data were from the Canadian Multicentre Osteoporosis Study (CaMos), a prospective cohort study initiated to provide national prevalence and incidence estimates for osteoporosis and osteoporosis-related fractures in the Canadian population [18]. Baseline data, which were the focus of the current study, were collected in 1996–1997, using both personal interview and paper-based questionnaires, from participants in nine Canadian regional urban centres. Respondents were at least 25 years of age and were recruited without regard for disease status.

Details of the methodology to select the CaMos cohort and collect the study data have been described elsewhere [2,18]. To ensure the quality and integrity of the data, interviewers are trained to minimize the amount of missing data (i.e., probe for responses), and respondents are re-contacted if clarification of responses is required to resolve inconsistencies in the data.

Measures

Version 1 of the SF-36 was used in CaMos; it encompasses eight sub-scales: PF, role physical, bodily pain, general health, vitality, social functioning, role emotional, and MH. Item responses are captured using dichotomous or ordinal scales [1]. The PF sub-scale contains 10 items, each having three response options: limited a lot, limited a little, and not limited at all. The MH sub-scale consists of five items, each having six response options: all of the time, most of the time, a good bit of the time, some of the time, a little of the time, none of the time. Responses for “Have you felt calm and peaceful?” and “Have you been a happy person?” are reverse coded so that higher scores represent better MH, in keeping with the other sub-scale items.

DIF analyses were conducted for the following demographic and health-related variables: sex, age group, body weight status, and self-perceived general (i.e., overall) health. Age was classified as 25–49 years, 50–64 years, 65–74 years, and ≥ 75 years. Body weight status was based on BMI, which was calculated from measured height and weight (kg/m^2), and was categorized as under or normal weight (< 25.0), overweight (25.0–29.9) and obese (≥ 30.0) in accordance with published guidelines [19]. General health was based on a single question in the SF-36 and was categorized as excellent/very good, good, and fair/poor.

Analysis

The analyses were conducted for respondents with complete information on all items or explanatory variables. Descriptive analyses were conducted using frequencies and percentages.

Both non-parametric and parametric approaches have been proposed to test for DIF; they can accommodate multiple population characteristics (i.e., covariates) that may be associated with item responses. Parametric approaches include logistic regression analysis [20], the multiple indicators multiple causes (MIMIC) model, and item response theory (IRT) models [5,21,22], with the latter two being popular because they can be applied to binary and ordinal item responses [5], are flexible to incorporate one or more latent constructs, and can be readily implemented using existing software. Woods [23] demonstrated, via simulation, that the MIMIC model will produce more accurate results than the IRT model, in a simple two-group scenario, for small group sizes.

We adopted the MIMIC model and used the following strategy to test for DIF. First, the assumption of unidimensionality, that is, that all sub-scale items measure a single construct [5], was examined by applying factor analysis with oblique rotation to the polychoric correlations for the sub-scale items [24]. The first eigenvalue should be significantly higher than the second eigenvalue (i.e., ratio $> 4:1$) to support unidimensionality [25]. As well, fit of a unidimensional measurement model was assessed using the comparative fit index (CFI) and the root mean square error of approximation (RMSEA). A RMSEA value ≤ 0.10 , and a CFI value > 0.90 indicate acceptable model fit [23,26,27]. The measurement model defines the relationship between the underlying latent construct (i.e., PF or MH) and the sub-scale items, while a structural model is used to specify the relationships of the observed covariates with the sub-scale items and the latent construct [23,28,29]. A statistically significant direct effect of the observed covariates on the sub-scale item(s) indicates uniform DIF, after controlling for differences in latent health status between comparison groups (Fig 1). In keeping with previous research [28–31], a three-step procedure was adopted: (a) empirically select an anchor item(s)

for the sub-scale, (b) test for DIF on each sub-scale item, and (c) fit a final model that allows for differential functioning on the individual sub-scale items.

At least one anchor item is needed to define the latent construct on which the groups are compared. We adopted the following method to select an anchor item [32]. First, a model was fit to the data that included the effects of the covariates on the latent variable but no direct effects between the covariates and the sub-scale items. Next, a series of models were fit to the data that added paths from the covariates to the items; this was done one sub-scale item at a time. A likelihood ratio (LR) statistic was used to test the difference between the two models for each item [32]. An item was labelled DIF-free if the LR statistic was not statistically significant. The DIF-free item with the smallest LR statistic was selected as the anchor item. If none of the items were DIF-free, then the item with the smallest LR statistic was selected as the anchor item [32].

Next, an unconstrained DIF model was fit to the data that included direct effects from the covariates to each sub-scale item, except for the anchor item. Then we fit a second set of no-DIF models that did not contain direct effects. A LR statistic was used to test the difference between the constrained “no-DIF” model in which the covariate direct effects were set to zero and the unconstrained DIF model with freely-estimated covariate direct effects for each item. A statistically significant LR statistic indicates the presence of uniform DIF on the item.

The third step was to fit a model that included direct effects of the covariates on all the items for which DIF was identified in the previous step, as well as direct effects of the covariates on the latent variable. This final DIF model was used to obtain parameter estimates and predict the factor scores.

Regression coefficients from the final model were exponentiated to produce odds ratios (ORs) to estimate the size of the DIF effects. Cut-points of 0.3, 0.5, 0.7 for the log of the ORs were used to indicate small, moderate and large DIF effect sizes [33]. Accordingly, an OR outside the range of 0.5 to 2.0 (i.e., a large effect size) was used to indicate a clinically meaningful DIF effect [13,34].

The impact of DIF was also investigated by testing the associations of the demographic and health variables with the predicted factor scores in the final DIF and no-DIF models [13].

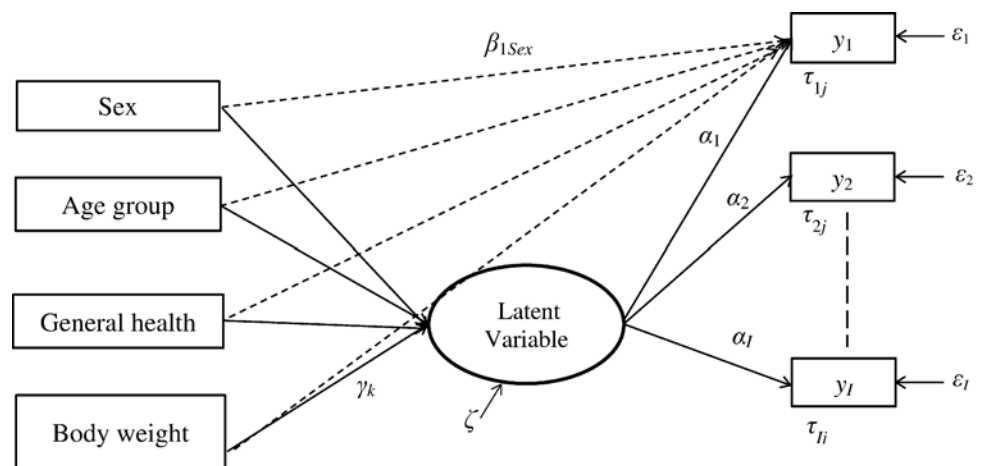


Fig 1. Illustration of the multiple indicators multiple causes model to test for differential item functioning on SF-36 sub-scale items. In this model, y_i is the i th sub-scale item ($i = 1, \dots, l$); the dashed arrow from each covariate to the item represents the DIF (i.e., direct) effect; β_{1Sex} is the regression coefficient for the difference in thresholds on item i for males and females; similar regression coefficients are defined for other model covariates; α_i is the regression coefficient for the latent variable and the i th item; γ_k is the regression coefficient for the latent variable and the k th covariate ($k = 1, \dots, K$); τ_{ij} is the threshold for the ($j - 1$)th response category ($j = 1, \dots, J$) for item i ; ϵ_i is the error term for the i th item; ζ is the residual error for the latent variable.

doi:10.1371/journal.pone.0151519.g001

Table 1. Percentage distribution of item responses on the SF-36 physical functioning and mental health sub-scale items in the Canadian Multicentre Osteoporosis Study.

Item	Response options					
	Limited a lot		Limited a little		Not limited at all	
Physical functioning (N = 9062)						
PF1: Vigorous activities	43.3		34.7		22.0	
PF2: Moderate activities	12.2		24.4		63.4	
PF3: Lifting or carrying groceries	8.5		22.2		69.3	
PF4: Climbing several flights of stairs	14.7		29.8		55.4	
PF5: Climbing one flight of stairs	5.2		15.7		79.1	
PF6: Bending, kneeling or stooping	12.6		33.1		54.4	
PF7: Walking more than a mile	16.0		20.7		63.3	
PF8: Walking several blocks	10.3		14.9		74.8	
PF9: Walking one block	3.6		9.9		86.5	
PF10: Bathing or dressing self	1.7		6.3		92.0	
Mental health (N = 9115)						
	All of the time	Most of the time	Good bit of the time	Some of the time	Little of the time	None of the time
<i>Have you...</i>						
MH1: Been a very nervous person?	1.3	3.1	5.4	17.3	29.9	43.1
MH2: Felt so down in the dumps that nothing could cheer you up?	0.3	1.0	1.7	7.3	17.1	72.5
MH3: Felt calm and peaceful?	10.6	46.4	18.4	15.6	5.9	3.2
MH4: Felt downhearted and blue?	0.3	1.5	3.4	17.6	35.1	42.1
MH5: Been a happy person?	18.4	54.9	13.1	8.6	2.8	2.2

doi:10.1371/journal.pone.0151519.t001

Differences in the predicted scores on the covariates for the two models were computed [13] and subsequently tested for statistical significance using a multivariable linear regression model.

Parameters were estimated using the maximum likelihood method with robust standard errors (MLR) [24]. Details for computation of the differences in log likelihood statistics for nested models can be found on the Mplus website (<https://www.statmodel.com/chidiff.shtml>). Given the increased probability of a Type I error when conducting multiple tests of significance, we adopted the Bonferroni procedure for all inferential analyses [35], adjusting the nominal level of significance (i.e., α) by either the number of items in the PF or MH sub-scales or the number of levels of the covariates, depending on the analysis.

The MIMIC framework was implemented using Mplus version 7.11 [24], while SAS version 9.3 was used to conduct the descriptive analyses [36]. To identify each model in the MIMIC framework, the latent factor mean was constrained to zero and the variance was constrained to one. This research was approved by the University of Manitoba Health Research Ethics Board. CaMos participants provided written informed consent at the time of study entry.

Results

Of the 9423 respondents in the CaMos cohort, 69.4% were women and slightly more than half (51.7%) were under 65 years of age. One third reported being in good health and 11.0% reported being in fair or poor health. Overweight and obese individuals accounted for 40.7% and 22.3% of respondents, respectively. A total of 96.2% ($n = 9062$) were retained in the PF sub-scale analysis and 96.7% ($n = 9115$) were retained in the MH sub-scale analysis after excluding individuals with missing observations.

Table 1 shows the distribution of item responses for the PF and MH sub-scales. Close to half (43.3%) of respondents reported being limited a lot in vigorous activities, while only 16.0%

were limited a lot in walking more than a mile. For the MH sub-scale, few respondents reported being very nervous, feeling so down in the dumps that nothing could cheer them up, or feeling downhearted and blue for some or all of the time. [S1 Table](#) and [S2 Table](#) shows the percentage of respondents for selected categories of the PF and MH sub-scale items on each of the covariates. There was a substantial increase in the percentage of respondents who experienced a lot of limitations in their activities, including walking a single block, with age. While there were few differences in the PF item percentages on “a lot of limitations” between respondents who were normal or underweight and overweight, those who were obese had a number of limitations. Respondents with excellent or very good overall health had substantially fewer functional limitations and were also much more likely to feel calm and peaceful and be happy. There were fewer differences for the MH sub-scale items, except for general health status.

For the PF sub-scale items, the factor analysis produced eigenvalues of 7.61 for the first factor and 0.63 for a second factor; the ratio of eigenvalues of 12.1 suggests a unidimensional factor structure. Further analysis revealed that a one-factor model without error covariances had a reasonable fit to the data (RMSEA = 0.11, 95% CI = 0.10–0.11; CFI = 0.98). However, a single-factor model with error covariances between five pairs of items (PF1 with PF2, PF2 with PF3, PF4 with PF5, PF7 with PF8, PF8 with PF9) resulted in a better fit to the data (RMSEA = 0.057, 95% CI = 0.054–0.067, CFI = 0.99). Factor loadings for the sub-scale items ranged from 0.80 to 0.93.

For the MH sub-scale items, factor analysis revealed that the eigenvalues were 3.13 for the first factor and 0.69 for a second factor. The one-factor model produced CFI = 0.97 and RMSEA = 0.15 (95% CI = 0.14–0.15), and factor loadings ranged in value from 0.61 to 0.82. Both of these analyses support a single dominant factor for both the PF and MH sub-scale items.

In step 1 of the DIF analysis, which focused on identifying anchor items for each of the sub-scales, all of the LR statistics were statistically significant, indicating that no items were DIF-free. However, “walking more than a mile” and “felt so down in the dumps that nothing could cheer you up” had the smallest LR statistics and thus were selected as anchor items. [Table 2](#) displays the LR test results for DIF. All the items except the anchor items showed statistically significant effects. The factor loadings and item thresholds are reported in [S3 Table](#).

As [Table 3](#) reveals, eight of the 10 PF sub-scale items exhibited DIF by age group, sex, body weight status and/or self-perceived health status. Women had a greater odds of reporting limitations on PF2, PF3, and PF4, and a greater odds of having no limitations in bathing or dressing (PF10; OR = 1.66), even after controlling for differences in their underlying functional abilities. Older respondents with the same underlying functional abilities scored lower than younger respondents in vigorous and moderate activities, and scored higher in walking one block and bathing/dressing self. With respect to self-perceived general health status, DIF effects were observed for four PF items, with respondents in poorer health being more likely to endorse limitations in vigorous and moderate activities and also being more likely to report better physical function in walking one block, relative to those in very good or excellent health, after controlling for differences in latent physical functioning. Relative to normal/underweight respondents, overweight and obese people were more likely to report fewer limitations in moderate activities and lifting or carrying groceries, but were more likely to report limitations in bending, kneeling or stooping, after controlling for overall physical health status. The DIF effects were small to large in size across the PF items; the smallest effects were observed for the following PF items: vigorous activities and lifting and carrying groceries. The largest effects were observed for lifting and carrying groceries and walking one block.

Four items in the MH sub-scale exhibited DIF effects. Women tended to endorse having been a very nervous person, feeling less calm and peaceful, and feeling downhearted and blue

Table 2. Tests for differential item functioning on the SF-36 physical functioning and mental health sub-scale items in the Canadian Multicentre Osteoporosis Study.

Item	LL ^a	df ^b	p-value ^c
Physical functioning (N = 9062)			
PF1	200.76	8	<0.0001
PF2	105.79	8	<0.0001
PF3	182.35	8	<0.0001
PF4	45.50	8	<0.0001
PF5	23.24	8	0.0031
PF6	77.18	8	<0.0001
PF7	†	†	†
PF8	27.77	8	0.0005
PF9	50.08	8	<0.0001
PF10	47.40	8	<0.0001
Mental health (N = 9115)			
MH1	75.23	8	<0.0001
MH2	†	†	†
MH3	56.53	8	<0.0001
MH4	24.05	8	0.0023
MH5	41.56	8	<0.0001

^aLL = log of the likelihood function.

^bdf = degrees of freedom. LL values and df were computed as the difference between the unconstrained DIF model (i.e., including DIF effects for all items except the anchor item) and the constrained DIF model (i.e., excluding DIF effects for one item at a time).

^cAll p-values are statistically significant using a Bonferroni-corrected $\alpha = .05/9 = .0056$ for the physical functioning sub-scale items and $\alpha = .05/4 = .0125$ for the mental health sub-scale items.

† = anchor item.

doi:10.1371/journal.pone.0151519.t002

more often than men, after controlling for differences in underlying MH status. Older respondents, relative to younger respondents who had the same latent MH status showed a greater propensity to feel calm and peaceful, and to be happy. Respondents with good and fair/poor health had lower odds of feeling calm and peaceful, and being less happy relative to respondents with better self-perceived general health status, even after controlling for differences in underlying MH status. Respondents who were obese had greater odds of being very nervous. Obese individuals were more likely to endorse being happy people (OR = 1.26) even after controlling for differences in their latent MH status. However, the DIF effects were small to moderate in size for all MH sub-scale items.

Table 4 provides the regression parameter estimates for the effects of the covariates on the latent PF and MH scores in the DIF and no-DIF models, and their absolute differences. For the PF sub-scale, all of the covariates were associated with latent PF scores. However, adjustment for DIF resulted in a reduction in the size of the parameter estimates for all covariates, except for overweight. The magnitude of the change in the coefficients was greatest for the age groups (absolute differences ranged from 0.09 to 0.14) and sex (absolute difference 0.09). For the MH sub-scale, adjustment for DIF changed the direction of the association between body weight and MH. In the no-DIF model, MH scores were significantly higher for obese respondents relative to normal/underweight respondents. However, in the DIF model, MH scores were lower for obese respondents. The largest change in regression coefficients between no-DIF and DIF models was found for fair/poor health (absolute difference 0.12).

Table 3. Odds ratios for differential item functioning (i.e., direct) effects on the SF-36 physical functioning and mental health sub-scale items in the Canadian Multicentre Osteoporosis Study.

Item	Sex ^a	Age Group			General Health		Body Weight	
	Female	50–64	65–74	≥75	Good	Fair/Poor	Overweight	Obese
Physical functioning^b								
PF1	0.86	0.56	0.34	0.24	0.73	0.46	1.04	1.10
PF2	0.71	0.80	0.89	0.73	0.82	0.61	1.43	2.28
PF3	0.39	0.87	1.14	1.13	0.82	0.67	1.41	2.46
PF4	0.65	0.89	1.06	1.05	0.83	0.80	0.79	0.81
PF5	0.88	1.00	1.09	1.62	0.98	1.16	0.85	0.82
PF6	0.93	0.84	0.93	1.18	0.92	0.90	0.72	0.57
PF7	† ^c	†	†	†	†	†	†	†
PF8	1.19	0.99	1.27	1.67	1.22	1.52	1.13	1.14
PF9	1.40	1.59	2.34	4.22	1.31	2.51	1.03	1.10
PF10	1.66	1.51	2.15	2.43	1.12	0.90	0.95	1.12
Mental health^b								
MH1	0.80	0.95	0.93	1.28	0.96	0.96	1.29	1.57
MH2	†	†	†	†	†	†	†	†
MH3	0.70	1.26	1.36	1.49	0.82	0.63	0.97	1.18
MH4	0.80	0.92	0.79	0.84	1.16	0.91	1.08	1.07
MH5	1.03	1.06	1.23	1.48	0.74	0.63	1.07	1.26

^aReference groups for the covariates are male, age 25–49 years, excellent/very good, underweight/normal weight.

^bBold odds ratios denote values that are statistically significant at $\alpha = 0.05/9 = 0.0056$ for the physical functioning sub-scale items and $\alpha = 0.05/5 = 0.0125$ for the mental health sub-scale items.

† = anchor item.

doi:10.1371/journal.pone.0151519.t003

As a sensitivity analysis, we fit separate models for each of the demographic and health status variables to test for DIF when the model was not simultaneously adjusted for all covariates. The ORs and magnitude of change in the coefficient estimates were similar to those reported in Tables 3 and 4.

Discussion

This study revealed that the majority of the SF-36 PF and MH sub-scale items showed evidence of DIF across one or more of the investigated demographic and health status variables after controlling for differences in the latent PF and MH of respondents in this population-based sample. Some of the DIF effects observed for the PF and MH sub-scale items were consistent with those found in previous studies [6,13,14]. Specifically, older people reported more limitations in vigorous and moderate activities, and fewer limitations in bathing or dressing, even after controlling for differences in their underlying PF [13,14]. Older respondents were also more likely to endorse feeling calm and peaceful than younger respondents after controlling for differences in their underlying MH [13,14]. Women tended to identify more problems in lifting or carrying groceries and climbing several flights of stairs, whereas they reported fewer problems in bathing or dressing than men [6,13,14].

This study revealed that body weight status was associated with DIF; this effect was observed for three items on the PF sub-scale and two items on the MH sub-scale. However, the direction of the effect for body weight status was not consistent across all items on the PF sub-scale. Overweight/obese people may not perceive themselves as being limited in activities that occur

Table 4. Regression coefficient estimates for the effects of demographic and health status variables on physical functioning and mental health latent variables in models without differential item functioning (no-DIF) and with differential item functioning (DIF) in the Canadian Multicentre Osteoporosis Study.

Variable ^a	Physical functioning			Mental health		
	No-DIF	DIF	d ^c	No-DIF	DIF	d ^c
	Coeff. (SE) ^b	Coeff. (SE)		Coeff. (SE)	Coeff. (SE)	
Female	-0.34 (0.03)*	-0.25 (0.04)*	0.09*	-0.27 (0.03)*	-0.18 (0.04)*	0.09*
Age Group						
50–64	-0.60 (0.04)*	-0.46 (0.05)*	0.14*	0.21 (0.03)*	0.19 (0.05)*	0.02
65–74	-1.07 (0.04)*	-0.95 (0.05)*	0.12*	0.39 (0.04)*	0.36 (0.05)*	0.03*
≥75	-1.74 (0.05)*	-1.65 (0.06)*	0.09*	0.49 (0.04)*	0.40 (0.06)*	0.09*
General Health						
Good	-0.78 (0.03)*	-0.73 (0.04)*	0.05*	-0.59 (0.03)*	-0.54 (0.04)*	0.05*
Fair/Poor	-1.68 (0.04)*	-1.62 (0.05)*	0.06*	-1.16 (0.05)*	-1.04 (0.06)*	0.12*
Weight Status						
Overweight	-0.10 (0.03)*	-0.10 (0.04)*	0.00	0.03 (0.03)	-0.01 (0.04)	0.04*
Obese	-0.48 (0.03)*	-0.53 (0.04)*	0.05*	0.08 (0.03)*	-0.02 (0.04)	0.10*

^aReference groups for the covariates are male, age 25–49 years, excellent/very good, underweight/normal weight.

^bCoeff. = coefficient; SE = standard error.

^c|d| = absolute value of the difference in coefficient estimates between the no-DIF and DIF models.

*denotes values that are statistically significant at $\alpha = 0.05/8 = 0.00625$.

doi:10.1371/journal.pone.0151519.t004

on a daily basis, like lifting or carrying groceries, but may be limited in activities that are more likely to occur on a daily basis, such as bending, kneeling or stooping. One prior study that evaluated DIF on the SF-36 for BMI showed significant non-uniform DIF in vigorous PF activities by weight category [11]. The inconsistent effect and limited research on the potential for DIF by body weight status suggests an opportunity for further research, including an exploration of potential non-uniform DIF.

Adjustment for DIF did not change the direction of the associations between the covariates and the PF latent variables, but the strength of the associations did change such that they were almost always smaller in size after adjusting for DIF. This finding is consistent with other studies showing that significant sex and age differences in physical ability were not altered after adjusting for DIF [10,13]. The largest change in regression coefficients between the no-DIF and DIF models was observed for age on the PF sub-scale items. However, adjustment for DIF changed the association between body weight and MH. There was a significant difference in the latent MH variable between the obese and underweight/normal weight groups before DIF adjustment, whereas the difference became non-significant after controlling for DIF. Large changes in regression coefficients between the no-DIF and DIF models for MH were observed for fair/poor health and obese groups. The results suggest that the comparison of PF and MH scores across population sub-groups defined by demographic and health status variables may be biased if DIF is ignored. Group comparisons on PF may be most affected by age and sex, while group comparisons on MH may be most affected by body weight and general health status.

The advantages of the MIMIC framework over other methods for DIF detection, such as logistic regression or IRT models, are that it is based on well-established confirmatory factor analysis modeling processes and can be used to test for DIF on multiple observed variables simultaneously. It can also be used to test for group differences in predicted factor scores

simultaneously on multiple observed covariates [13,37]. In addition, it allows for the analysis of DIF effects on a latent construct by comparing group differences in predicted factor scores between DIF-adjusted model and unadjusted models [13,37,38]. The MIMIC model for ordinal indicators is equivalent to the popular IRT graded-response model [22,38]; it is now frequently used to test for DIF and is known to perform well under a wide variety of data-analytic conditions [13,23,28,29]. Previous studies have demonstrated that the MIMIC model is effective in identifying DIF items, and is less sensitive to potential contamination of anchor items than other DIF detection methods [21,23,29]. Simulation studies have demonstrated that the MIMIC framework is more sensitive to identify DIF items and provides better control of Type I errors than conventional DIF test methods [21,31]. The chief disadvantage is that it cannot be used to test for non-uniform DIF (i.e., differential effects on item difficulty).

The study has other strengths. The DIF analysis was applied to a national population-based sample with a large sample size, which allowed for consideration of multiple demographic and health status variables. BMI was based on measured height and weight, and is therefore less susceptible to measurement error than self-reported height and weight. Age and sex are also confirmed by study staff during the data collection process, and are therefore likely to exhibit little, if any, measurement error.

However, this study is not without limitations. The MIMIC approach assumes that items have similar discriminative performance across comparison groups. Non-uniform DIF models, which test interactions between covariates and latent variables on the item responses, were not investigated because previous simulation studies have reported that the MIMIC approach may result in inflated Type I error rates when interaction terms are added to the model [39]. Performance of the MIMIC model for non-uniform DIF needs to be further studied. We only tested for DIF for two SF-36 sub-scales; other sub-scales have been examined for DIF using other statistical methods (i.e., logistic regression) [11,14]. We did not include other sub-scales in the current study; when there are small numbers of items (i.e., less than five items per subscale), the MIMIC framework is not an appropriate choice [13]. The CaMos sample includes a higher proportion of older adults and women than in the general population, thus the findings may under represent younger people (i.e., <50 years of age) and men. As well, for the one-factor model for the MH items, the RMSEA did not suggest a good fit to the data, although this finding is consistent with other studies [40,41]. Additional covariates could be considered in the model, although age and sex are two of the most common demographic variables considered in DIF analyses.

Conclusions

In summary, this study revealed the presence of DIF in population-based SF-36 data. The results indicate that PF and MH sub-scale scores may not be comparable across sub-groups defined by demographic and health status variables without accounting for DIF. DIF can affect the validity of epidemiologic studies that use self-report measures of quality of life. Removing items that exhibit DIF affects the content validity of a measure, which is problematic for well-established and well-known measures such as the SF-36. Removing items can also affect the comparability of sub-scale scores across different studies [6,42]. An alternative strategy is to replace the items with equivalent items that do not exhibit DIF; item banks may be a resource for identifying DIF-free items [43]. However, often the preferred approach is to examine items for DIF prior to conducting other analyses and adjust for any identified DIF effects prior to making comparisons between sub-groups [6,11,42]. Evaluating population-based self-reported outcome measures for DIF, particularly on key demographic and health-related variables, should therefore be a routine component of all comparative analyses.

Supporting Information

S1 Table. Percentages of respondents for the category “Limited a lot” on the PF sub-scale items by demographic and health status variables in the Canadian Multicentre Osteoporosis Study ($n = 9062$).

(DOCX)

S2 Table. Percentages of respondents for the category “All/most/good bit of the time” on the MH sub-scale items by demographic and health status variables in the Canadian Multicentre Osteoporosis Study ($n = 9115$).

(DOCX)

S3 Table. Factor loading and item threshold estimates of the differential item functioning (DIF) model for the SF-36 physical functioning and mental health sub-scale items in the Canadian Multicentre Osteoporosis Study.

(DOCX)

Acknowledgments

We thank all those participants in the Canadian Multicentre Osteoporosis Study (CaMos) whose careful responses and attendance made this analysis possible. We thank the CaMos Research Group for access to the data needed to complete this study. This Group includes:

David Goltzman (co-principal investigator, McGill University), Nancy Kreiger (co-principal investigator, Toronto), Alan Tenenhouse (principal investigator emeritus, Toronto),

CaMos Coordinating Centre, McGill University, Montreal, Quebec: Suzette Poliquin (national coordinator, emeritus), Suzanne Godmaire (research assistant), Silvia Dumont (administrative assistant), Claudie Berger (study statistician), Lisa Langsetmo (Fellow).

Memorial University, St. John’s Newfoundland: Carol Joyce (director), Christopher Kovacs (co-director), Emma Sheppard (coordinator).

Dalhousie University, Halifax, Nova Scotia: Susan Kirkland, Stephanie Kaiser (co-directors), Barbara Stanfield (coordinator).

Laval University, Quebec City, Quebec: Jacques P. Brown (director), Louis Bessette (co-director).

Queen’s University, Kingston, Ontario: Tassos Anastassiades (director), Tanveer Towheed (co-director), Wilma Hopman (research associate), Karen Rees-Milton (coordinator).

University of Toronto, Toronto, Ontario: Bob Josse (director), Sophie Jamal (co-director), Barbara Gardner-Bray (coordinator), Angela Cheung (scientist).

McMaster University, Hamilton, Ontario: Jonathan D. Adachi (director), Alexandra Papaioannou (co-director), Laura Pickard (coordinator).

University of Saskatchewan, Saskatoon, Saskatchewan: Wojciech P. Olszynski (director), K. Shawn Davison (co-director), Jola Thingvold (coordinator).

University of Calgary, Calgary, Alberta: David A. Hanley (director), Steven Boyd (co-director), Jane Allan (coordinator).

University British Columbia, Vancouver, British Columbia: Jerilynn C. Prior (director), Millan Patel (co-director), Brian Lentle (radiologist), Nerkeza Andjelic (coordinator), Yvette Vigna (coordinator, emeritus).

McGill University, Montreal, Quebec: Elham Rahme (biostatistician), Brent Richards (research associate).

University of Alberta, Edmonton, Alberta: Stuart Jackson (medical physicist).

University of Manitoba, Winnipeg, Manitoba: William Leslie (nuclear medicine physician).

Author Contributions

Conceived and designed the experiments: LML WH TTS RS JCP JL. Analyzed the data: LML XW. Wrote the paper: LML XW WH NM TTS JL JCP AP RGJ TET KSD RS. Designed and Implemented Data Collection Protocol: WH JCP AP RGJ TET KSD.

References

1. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Med Care*. 1992; 30: 473–483. PMID: [1593914](#)
2. Hopman WM, Towheed T, Anastassiades T, Tenenhouse A, Poliquin S, Berger C, et al. Canadian normative data for the SF-36 health survey. Canadian Multicentre Osteoporosis Study Research Group. *CMAJ*. 2000; 163: 265–271. PMID: [10951722](#)
3. Watson EK, Firman DW, Baade PD, Ring I. Telephone administration of the SF-36 health survey: validation studies and population norms for adults in Queensland. *Aust N Z J Public Health*. 1996; 20: 359–363. PMID: [8908757](#)
4. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. International consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes: results of the COSMIN study. *J Clin Epidemiol*. 2010; 63: 737–745. doi: [10.1016/j.jclinepi.2010.02.006](#) PMID: [20494804](#)
5. Teresi JA, Fleishman JA. Differential item functioning and health assessment. *Qual Life Res*. 2007; 16 (Suppl 1): 33–42. PMID: [17443420](#)
6. McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care*. 1993; 31: 247–263. PMID: [8450681](#)
7. McHorney CA, Ware JE Jr, Lu JF, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care*. 1994; 32: 40–66. PMID: [8277801](#)
8. McHorney CA, Ware JE Jr. Construction and validation of an alternate form general mental health scale for the Medical Outcomes Study Short-Form 36-Item Health Survey. *Med Care*, 1995; 33: 15–28. PMID: [7823644](#)
9. Sullivan M, Karlsson J, Ware JE Jr. The Swedish SF-36 Health Survey—I. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. *Soc Sci Med*. 1995; 41: 1349–1358. PMID: [8560302](#)
10. Fleishman JA, Lawrence WF. Demographic variation in SF-12 scores: true differences or differential item functioning? *Med Care*. 2003; 41(7 Suppl): III75–III86. PMID: [12865729](#)
11. Pollard B, Johnston M, Dixon D. Exploring differential item functioning in the SF-36 by demographic, clinical, psychological and social factors in an osteoarthritis population. *BMC Musculoskelet Disord*. 2013; 14: 346–247. doi: [10.1186/1471-2474-14-346](#) PMID: [24330385](#)
12. Lix LM, Acan Osman B, Adachi JD, Towheed T, Hopman W, Davison KS, et al. Measurement equivalence of the SF-36 in the Canadian Multicentre Osteoporosis Study. *Health Qual Life Outcomes*. 2012; 10: 29. doi: [10.1186/1477-7525-10-29](#) PMID: [22414200](#)
13. Yu YF, Yu AP, Ahn J. Investigating differential item functioning by chronic diseases in the SF-36 health survey: a latent trait analysis using MIMIC models. *Med Care*. 2007; 45: 851–859. PMID: [17712255](#)
14. Perkins AJ, Stump TE, Monahan PO, McHorney CA. Assessment of differential item functioning for demographic comparisons in the MOS SF-36 health survey. *Qual Life Res*. 2006; 15: 331–348. PMID: [16547771](#)
15. Bjorner JB, Kreiner S, Ware JE, Damsgaard MT, Bech P. Differential item functioning in the Danish translation of the SF-36. *J Clin Epidemiol*. 1998; 51: 1189–1202. PMID: [9817137](#)
16. Sawatzky R, Ratner PA, Kopec JA, Zumbo BD. Latent variable mixture models: a promising approach for the validation of patient reported outcomes. *Qual Life Res*. 2012; 21: 637–650. doi: [10.1007/s11136-011-9976-6](#) PMID: [21814876](#)
17. Hopman WM, Berger C, Joseph L, Barr SI, Gao Y, Prior JC, et al. The association between body mass index and health-related quality of life: data from CaMos, a stratified population study. *Qual Life Res*. 2007; 16: 1595–1603. PMID: [17957495](#)
18. Kreiger N, Tenenhouse A, Joseph L, Mackenzie T, Poliquin S. The Canadian Multicentre Osteoporosis Study (CaMos): Background, rationale, methods. *Can J Aging*. 1999; 18: 376–387.
19. Health Canada. Canadian guidelines for body weight classification in adults. Ottawa, Canada: Health Canada Publications Centre, Publication ID 4645. 2013. Available: http://preventdisease.com/pdf/weight_book-livres_des_poids_e.pdf.

20. Zumbo BD (1999) A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modelling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. [Available: <http://faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf>]. Accessed 19 August 2015.
21. Finch H. The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBT-EST, and the IRT likelihood ratio. *Appl Psychol Meas*. 2005; 29:278–295.
22. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*. 1969; Monograph No. 17. Available: <http://www.psychometrika.org/journal/online/MN17.pdf>.
23. Woods CM. Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Mult Behav Res*. 2009; 44: 1–27.
24. Muthén LK, Muthén BO. *MPlus user's guide*. 7th ed. Los Angeles, CA: Muthén & Muthén; 2012.
25. Lord FM. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: L. Erlbaum Associates; 1980.
26. Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS, editors. *Testing structural equation models*. Newbury Park, CA: Sage; 1993. pp. 136–162.
27. Hu LT, Bentler PM. Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Modeling*. 1999; 6: 1–55.
28. Woods CM, Oltmanns TF, Turkheimer E. Illustration of MIMIC-model DIF testing with the schedule for nonadaptive and adaptive personality. *J Psychopathol Behav Assess*. 2009; 31: 320–330. PMID: [20442793](https://pubmed.ncbi.nlm.nih.gov/20442793/)
29. Wang W, Shih C, Yang C. The MIMIC method with scale purification for detecting differential item functioning. *Ed Psychol Meas*. 2009; 69: 713–731.
30. Teresi JA, Ocepek-Welikson K, Kleinman M, Cook KF, Crane PK, Gibbons LE, et al. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. *Qual Life Res*. 2007; 6 (Suppl 1): 43–68.
31. Wang W, Shih C, Sun G. The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Ed Psychol Meas*. 2012; 72: 687–708.
32. Woods CM. Empirical selection of anchors for tests of differential item functioning. *Appl Psychol Meas*. 2009; 33: 42–57.
33. Jin Y, Myers ND, Ahn S, Penfield RD. A comparison of uniform DIF effect size estimators under the MIMIC and Rasch models. *Educ Psychol Meas*. 2013; 73: 339–358.
34. Cole SR, Kawachi I, Maller SJ, Berkman LF. Test of item-response bias in the CES-D scale. Experience from the New Haven EPESE study. *J Clin Epidemiol*. 2000; 53: 285–289. PMID: [10760639](https://pubmed.ncbi.nlm.nih.gov/10760639/)
35. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ*. 1995; 310(6973): 170. PMID: [7833759](https://pubmed.ncbi.nlm.nih.gov/7833759/)
36. SAS Institute Inc. *SAS/STAT 9.3 User's Guide*. Cary, NC: SAS Institute Inc.; 2013.
37. Arthurs E, Steele RJ, Hudson M, Baron M, Thombs BD, Canadian Scleroderma Research Group. Are scores on English and French versions of the PHQ-9 comparable? An assessment of differential item functioning. *PLoS One*. 2012; 7(12): e52028. doi: [10.1371/journal.pone.0052028](https://doi.org/10.1371/journal.pone.0052028) PMID: [23251676](https://pubmed.ncbi.nlm.nih.gov/23251676/)
38. Kamata A, Bauer DJ. A note on the relation between factor analytic and item response theory models. *Struct Equ Modeling*. 2008; 15: 136–153.
39. Woods CM, Grimm KJ. Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Appl Psychol Meas*. 2011; 35: 339–361.
40. Hays RD, Liu H, Spritzer K, Cella D. Item response theory analyses of physical functioning items in the Medical Outcomes Study. *Med Care*. 2007; 45(5 Suppl1): S32–8. PMID: [17443117](https://pubmed.ncbi.nlm.nih.gov/17443117/)
41. Crins MH, Roorda LD, Smits N, de Vet HC, Westhovens R, Cella D, et al. Calibration and validation of the Dutch-Flemish PROMIS Pain Interference Item Bank in patients with chronic pain. *PLoS One*. 2015; 10(7): e0134094. doi: [10.1371/journal.pone.0134094](https://doi.org/10.1371/journal.pone.0134094) PMID: [26214178](https://pubmed.ncbi.nlm.nih.gov/26214178/)
42. Fleishman JA, Spector WD, Altman BM. Impact of differential item functioning on age and gender differences in functional disability. *J Gerontol B Psychol Sci Soc Sci*. 2002; 57: S275–S284 PMID: [12198107](https://pubmed.ncbi.nlm.nih.gov/12198107/)
43. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J Clin Epidemiol*. 2010; 63: 1179–1194. doi: [10.1016/j.jclinepi.2010.04.011](https://doi.org/10.1016/j.jclinepi.2010.04.011) PMID: [20685078](https://pubmed.ncbi.nlm.nih.gov/20685078/)