ELSEVIER

Journal of Clinical Epidemiology 59 (2006) 478–484

Journal of Clinical Epidemiology

A 37-item shoulder functional status item pool had negligible differential item functioning

Paul K. Crane[a,*], Dennis L. Hart[b], Laura E. Gibbons[a], Karon F. Cook[c,d]

[a]*Division of General Internal Medicine, University of Washington School of Medicine, Seattle, WA, USA*
[b]*Focus On Therapeutic Outcomes, Inc., 551 Yopps Cove Road, White Stone, VA 22578, USA*
[c]*Veterans Affairs Measurement Excellence and Training Resource Information Center (METRIC), Houston, TX, USA*
[d]*University of Washington Center on Outcomes Research Rehabilitation, Seattle, WA, USA*

Accepted 2 October 2005

Abstract

Objective: Measures of shoulder function may differ by dominance of affected shoulder, surgical history, gender, or race. We present a technique for determining whether observed differences in function between groups are due to biased test items or real differences in function.

Study Design and Setting: Four hundred patients who were receiving rehabilitation for a variety of shoulder impairments completed a survey of shoulder function. Thirty-seven items measuring shoulder function were analyzed for differential item functioning (DIF) related to demographic characteristics using an ordinal logistic regression (OLR) and item response theory (IRT) approach. When DIF was identified in an item, we modified the IRT analysis to calibrate item parameters separately in appropriate demographic groups. We compared adjusted and unadjusted patient ability measures in each demographic group.

Results: Several items were found to have a modest amount of DIF related to the different demographic characteristics, especially gender; however, adjusting measures for DIF had little impact on overall measures of shoulder function and made almost no difference in average shoulder function across demographic groups.

Conclusion: In this pool of shoulder function items, adjustment for DIF made almost no difference in measures of function across demographic groups. © 2006 Elsevier Inc. All rights reserved.

Keywords: Differential item functioning; Ordinal logistic regression; Item response theory; Psychometrics; Shoulder function; Measurement invariance

1. Introduction

Clinicians need brief, reliable, valid, and responsive measures of a patient's function to assess the effects of rehabilitation. If differences are found in measures of function between different demographic groups, those differences should represent true differences in underlying function rather than item bias. Item bias erodes the construct validity of the functional scale.

Item bias or differential item functioning (DIF) is present when the relationship between item responses and the trait or ability measured by the test differs systematically between groups of patients after controlling for the patient's underlying ability [1]. A number of methods have been developed to determine whether items have DIF [1–3]. DIF may be insidious, and it is important to validly measure DIF, especially in a computerized adaptive testing (CAT) framework where relatively few items are presented to each patient.

Our objective was to determine whether items from a test of shoulder function have DIF and to determine whether any DIF that was found made a difference in estimates of shoulder function. The intended use of the items was for a CAT in which item response theory (IRT) scoring would be used [4].

2. Methods

2.1. Subjects

The sample analyzed has been detailed elsewhere [4]. Briefly, data from 400 patients who were receiving outpatient rehabilitation or seeing an orthopedic surgeon for a variety of shoulder impairments were analyzed. Patients, who represent a sample of convenience and who have been

* Corresponding author. Tel.: 206-744-1831.
E-mail address: pcrane@u.washington.edu (P.K. Crane).

0895-4356/06/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2005.10.007

described previously [5], completed the paper-and-pencil development pool of shoulder functional status items while they waited to be seen by their surgeon or therapist.

2.2. Data

Data for the study were responses to a set of 37 shoulder functional status items previously found to be unidimensional and locally independent [4]. For the present study, the latent trait of interest was shoulder functional status (SFS), which we operationally defined as the patient's perception of his or her ability to perform functional tasks described in the 37 SFS items. The items came from a larger pool of 60 items adapted from existing scales or developed based on patient interviews and input from an expert panel [5]. For each item, respondents indicated how much difficulty they had with the specified task. Response options and their corresponding scores were: "no difficulty" = 5, "little difficulty" = 4, "some difficulty" = 3, "much difficulty" = 2, "I can't do this" = 1, and "didn't do before shoulder problem" = N/A. The last response was considered missing data and was not scored.

2.3. Data analyses

For the purposes of DIF detection, we calibrated item responses to Samejima's graded response model (GRM) [6,7] using PARSCALE software [8]. We used SFS measures estimated using the GRM to look for DIF using DIFdetect [9], a DIF detection package for STATA [10]. Details of the analytic technique employed in DIFdetect have been published [11]. DIFdetect examines three ordinal logistic regression (OLR) models for each item and each demographic category selected for analysis:

$$f(\text{item response}) = \text{cut} + \beta_1 \times \text{shoulder function} + \beta_2 \times \text{group} + \beta_3 \times \text{shoulder function} \times \text{group} \quad \text{(model 1)}$$

$$f(\text{item response}) = \text{cut} + \beta_1 \times \text{shoulder function} + \beta_2 \times \text{group} \quad \text{(model 2)}$$

$$f(\text{item response}) = \text{cut} + \beta_1 \times \text{shoulder function} \quad \text{(model 3)}$$

where cut is the cutpoint for each level in the ordinal logistic regression model (as described by McCullagh and Nelder [12]) and shoulder function is the GRM estimate of the patient's SFS ability.

Two types of DIF are identified in the literature: uniform and nonuniform. In items with uniform DIF, the interference related to demographic groups between ability or trait level and item responses is the same across the entire range measured by the test (i.e., the same impact is seen from low to high shoulder function). In items with nonuniform DIF, the interference varies at different levels of the trait being measured. These concepts are analogous to confounding and effect modification relationships from epidemiology [11].

To detect nonuniform DIF, DIFdetect compares the -2 log likelihoods of models 1 and 2. The difference in log likelihoods is distributed as χ^2 with 1 degree of freedom. The user specifies the α-level associated with significant nonuniform DIF; the default is .05. To detect uniform DIF, the relative difference between the parameters associated with shoulder function (β_1 from models 2 and 3) is determined using the formula $|(\beta_{1(\text{model 2})} - \beta_{(\text{model 3})})/\beta_{1(\text{model 2})}|$. If the relative difference is large, group membership interferes with the relationship expected between ability and item responses. The user specifies the relative difference in β_1 associated with significant uniform DIF; the default is 10%.

We initially used criteria to determine whether items in the shoulder function scale had large amounts of DIF ("larger DIF criteria"). For nonuniform DIF, we used Bonferroni adjustment of the α-level based on the 37 items in the test, so that the critical α was .0014. For uniform DIF, we used a 10% change in β_1 criterion. We performed this analysis for four different demographic characteristics (gender, dominance of affected shoulder, history of shoulder surgery, and ethnic group) [4].

We repeated these analyses using criteria to determine whether there may be smaller amounts of DIF ("smaller DIF criteria"). We used unadjusted α-levels for nonuniform DIF (.05), and a 5% change in β coefficient for uniform DIF.

The smaller DIF criteria were used to adjust the GRM SFS estimates for DIF [13] and to address the issue of spurious DIF (see below). We evaluated each demographic category in turn, starting with gender, and created new variables when DIF was found. First, items that were found to have DIF related to gender were split into two new items. For the first new item, responses for females were as coded in the original dataset, while for males all responses were set to missing. For the second new item, responses for males were coded as in the original dataset, while for females all responses were set to missing. We thus calibrated item parameters independently in the two groups for items found with DIF. Items free of DIF served as anchor items, ensuring that the shoulder function measure was calibrated on the same metric for the two genders.

Next we addressed the issue of spurious DIF. Spurious DIF is a false identification of DIF (i.e., no DIF caused by DIF present in other items) [2,3]. To determine if any items had spurious DIF, we used the adjusted GRM SFS measure estimates in DIFdetect. If the items found with DIF were different from the items found in the previous round, we ascribed those differences to spurious DIF. If there was spurious DIF, we created a new dataset and

Table 1
Item mean scores for 37 shoulder functioning items by demographic groups

	Gender		Dominant arm		Surgery		Race	
Item	F	M	Yes	No	Yes	No	White	Other
Flush toilet	4.04	4.37	4.22	4.32	4.20	4.28	4.31	4.05
Apply deodorant	3.84	4.03	3.94	4.01	4.11	3.90	4.07	3.59
Don tie	3.50	3.85	3.69	3.81	3.77	3.73	3.89	3.19
Adjust collar	3.11	3.44	3.27	3.39	3.39	3.29	3.47	2.82
Take off glasses	4.31	4.40	4.37	4.39	4.45	4.31	4.47	4.02
Pull socks on	3.97	4.22	4.09	4.22	4.24	4.08	4.23	3.74
Put on underpants	3.80	4.30	4.07	4.22	4.23	4.07	4.24	3.70
Comb hair	3.04	3.58	3.37	3.43	3.50	3.32	3.51	2.98
Soup on overhead shelf	2.95	3.33	3.12	3.31	3.20	3.19	3.29	2.88
Soup on shoulder-level shelf	3.31	3.65	3.50	3.60	3.60	3.50	3.69	3.01
Reach string	3.00	3.46	3.29	3.32	3.38	3.25	3.40	2.91
Lower light object	2.58	3.16	2.94	2.97	2.93	2.95	2.99	2.82
Pick up water	3.86	4.20	4.08	4.11	4.12	4.06	4.21	3.63
Use skillet	2.61	3.58	3.23	3.27	3.28	3.20	3.37	2.76
Slide clothes	2.87	3.64	3.37	3.37	3.45	3.32	3.47	2.96
Carry object in crook of arm	3.27	3.71	3.48	3.68	3.59	3.54	3.68	3.10
Touch opposite ear	4.03	4.22	4.10	4.24	4.22	4.10	4.26	3.73
Reach back seat	2.12	2.52	2.37	2.40	2.42	2.35	2.40	2.31
Reach back pocket	2.99	3.63	3.36	3.47	3.51	3.32	3.52	2.96
Reach shoulder height shelf	3.11	3.58	3.35	3.52	3.45	3.39	3.54	2.94
Reach overhead shelf	2.54	3.05	2.79	2.98	2.92	2.83	2.95	2.51
Safety strap	3.10	3.60	3.43	3.43	3.47	3.41	3.55	3.03
Turn steering wheel	3.33	3.63	3.50	3.58	3.52	3.53	3.60	3.23
Turn faucets	4.01	4.35	4.18	4.33	4.33	4.18	4.35	3.84
Stir potatoes	2.80	3.66	3.33	3.38	3.47	3.30	3.43	3.07
Work overhead	2.14	2.35	2.25	2.32	2.27	2.30	2.30	2.24
Put arm on table	4.03	4.14	4.09	4.15	4.13	4.07	4.23	3.70
Reach salt shaker	3.52	3.86	3.70	3.82	3.75	3.73	3.81	3.51
Push chair	3.43	3.85	3.66	3.79	3.72	3.68	3.79	3.40
Pull chair	3.33	3.79	3.60	3.68	3.71	3.57	3.74	3.24
Throw ball underhand	3.02	3.43	3.21	3.42	3.14	3.36	3.39	2.96
Reach under bed	2.78	3.36	3.12	3.21	3.26	3.12	3.30	2.67
Pull box	3.12	3.58	3.38	3.49	3.50	3.37	3.56	2.95
Wash face	4.09	4.19	4.12	4.24	4.20	4.12	4.24	3.87
Lift chest lid	3.45	3.87	3.70	3.78	3.85	3.65	3.85	3.26
Tighten jar	3.19	3.75	3.56	3.55	3.60	3.53	3.69	3.10
Steady jar	3.46	3.88	3.68	3.81	3.74	3.74	3.88	3.21

Values are average response options scored from 1 (“I can’t do this”) to 5 (“no difficulty”).

new PARSCALE code to account for the items most recently found to have DIF, and repeated the procedure. If the items found with DIF were the same as the previous run, however, we concluded that there was no further spurious DIF. We continued these steps until additional adjustment did not affect the determination of which items had DIF related to gender.

We repeated the entire procedure in turn for dominance of affected shoulder (dominant or nondominant arm affected), surgical status (yes or no), and self-reported race (“Caucasian” vs. “other”). If an item had been found to have DIF related to a previous demographic category, it was examined separately in appropriate groups. Thus, items found with gender DIF were examined for DIF related to dominance of affected shoulder separately in males and females. The technique produced SFS estimates adjusted for DIF related to gender, dominance of affected shoulder, surgical status, and self-reported race. (PARSCALE code for all of the analyses performed is available on request from the first author.)

We compared the unadjusted and fully adjusted SFS measures in the four demographic comparisons to see whether adjusting for DIF had any impact on the mean SFS measure differences found between the groups. For ease of comprehension, both SFS measures were calibrated to a mean of 100 and SD of 15 with higher values representing greater SFS. Differences between the mean unadjusted and fully adjusted SFS measures were compared with t-tests for various demographic groups. We also examined the distribution of differences between unadjusted and fully adjusted SFS measures.

2.4. Visualization of DIFdetect item results

To confirm the relationship between the ordinal logistic regression approach embodied by DIFdetect and the IRT

Table 2
Differential item functioning results from the 37 shoulder functioning items using the iterative DIF detection and IRT adjustment procedure

	Gender		Dominant arm		Surgery		Race	
Item	N[a]	U[b]	N	U	N	U	N	U
Flush toilet	—	—	—	—	—	—	—	—
Apply deodorant	—	—	—	—	.021	—	—	—
Don tie	—	—	—	—	—	—	—	—
Adjust collar	—	—	—	—	—	—	—	—
Take off glasses	—	10.0%	—	—	—	—	—	—
Pull socks on	—	—	—	—	—	—	—	—
Put on underpants	—	—	—	—	—	—	—	—
Comb hair	.009	—	—	—	—	—	—	—
Soup on overhead shelf	—	—	—	—	—	—	—	—
Soup on shoulder-level shelf	—	—	—	—	—	—	—	—
Reach string	—	—	—	—	—	—	—	—
Lower light object	—	—	—	—	—	—	.021	7.7%
Pick up water	—	—	—	—	—	—	—	—
Use skillet	—	—	—	—	—	—	.035	—
Slide clothes	—	—	—	—	.002	—	—	—
Carry object in crook of arm	—	—	—	—	—	—	—	—
Touch opposite ear	—	6.0%	.010[c]	—	—	—	—	—
Reach back seat	—	—	—	—	—	—	—	6.0%
Reach back pocket	—	—	—	—	—	—	—	—
Reach shoulder height shelf	—	—	—	—	—	—	—	—
Reach overhead shelf	—	—	—	—	—	—	—	—
Safety strap	—	—	—	—	—	—	—	—
Turn steering wheel	—	—	—	—	—	—	—	—
Turn faucets	.003	—	—	—	.005[d]	—	—	—
Stir potatoes	—	—	—	—	—	—	—	—
Work overhead	—	—	—	—	—	—	—	9.4%
Put arm on table	—	7.1%	—	—	—	—	—	—
Reach salt shaker	—	—	—	—	—	—	—	—
Push chair	—	—	—	—	—	—	—	—
Pull chair	—	—	—	—	—	—	—	—
Throw ball underhand	.022	—	—	—	—	—	—	—
Reach under bed	—	—	—	—	—	—	—	—
Pull box	—	—	—	—	—	—	—	—
Wash face	—	9.5%	—	—	—	—	—	—
Lift chest lid	—	—	—	—	—	—	—	—
Tighten jar	—	—	—	—	—	—	—	—
Steady jar	—	—	—	—	—	—	—	—

Abbreviations: DIF, differential item functioning; N, nonuniform DIF; U, uniform DIF.
[a] Nonuniform DIF: *P*-values for the group × shoulder function interaction.
[b] Uniform DIF: percentages for the change in the shoulder function coefficient when the group term is in the model.
[c] In females.
[d] In males.

conceptualization of DIF, we plotted item boundary response functions in different demographic groups for selected items we found with DIF. These curves represent the probabilities for responding at or higher than each response category (y-axis) plotted against SFS measures (x-axis) for each demographic category. In the absence of DIF, curves for the two demographic categories should be superimposed. In the presence of DIF, there will be differences between the curves. If there is uniform DIF, the curves from one demographic group should be uniformly higher than the curves of the other demographic group (i.e., one curve in its entirety is shifted left or right on the plot). If there is nonuniform DIF, the relationship between the curves should be different at the extremes of the SFS ability scale. Thus, the two sets of curves will not be superimposed, implying there is DIF. Furthermore, the relationship between the two sets of curves will be in one direction at the right end of the scale (i.e., high functioning), but the other direction at the left end of the scale (i.e., low functioning). It is also possible for items to have both uniform and nonuniform DIF. In such an item, (a) the curves are not superimposed on top of each other, which implies the presence of DIF, (b) one set of curves is always higher than the other set of curves, which implies uniform DIF, and (c) the distance between the

Table 3
Descriptive statistics for the unadjusted and fully adjusted shoulder function estimates in four demographic groups

	Mean SFS (SD)	Median SFS	Interquartile range
Gender			
Female			
Unadjusted	96.3 (15.0)	93.8	86.4–105.7
Adjusted	96.0 (15.2)	93.0	85.5–105.8
Male			
Unadjusted	102.1 (14.6)	101.5	92.0–111.7
Adjusted	102.3 (14.4)	101.5	92.1–112.5
Dominance of affected shoulder			
Left			
Unadjusted	100.9 (15.2)	101.9	91.9–110.7
Adjusted	100.9 (15.3)	102.1	91.7–111.1
Right			
Unadjusted	99.6 (14.6)	98.4	90.2–108.8
Adjusted	99.6 (14.6)	98.5	89.6–109.2
Surgery status			
History of surgery			
Unadjusted	101.1 (15.9)	101.6	90.6–111.8
Adjusted	101.2 (16.0)	101.9	90.5–112.5
No history of surgery			
Unadjusted	99.5 (14.5)	98.4	90.7–109.2
Adjusted	99.5 (14.4)	98.3	90.4–109.9
Self-reported race			
White			
Unadjusted	101.7 (14.4)	101.1	91.7–111.1
Adjusted	101.8 (14.6)	101.3	91.9–111.3
Nonwhite			
Unadjusted	94.3 (15.8)	93.1	84.6–103.6
Adjusted	94.0 (15.1)	92.5	84.0–103.7

SFS measures have been calibrated to have a mean of 100 and a standard deviation of 15 for ease of comprehension.

Abbreviations: SD, standard deviation; SFS, measure of shoulder function status.

two sets of curves either increases or decreases from the left end of the scale to the right end of the scale, which implies nonuniform DIF.

3. Results

Item responses were available from 400 patients. There were 256 men (64%), 309 whites (78%), 134 with a history of surgery (34%), and 232 with the dominant shoulder affected (59%). Demographic data were missing on a small number of patients (0 for gender, 6 for ethnicity, 7 for surgery status, and 4 for dominance; overall $<2\%$ missing). Further demographic details are available (see Table 1 in [4]). Item means for each demographic category are shown in Table 1. Working overhead had the lowest mean score, implying that people had the most difficulty with this task, while taking off glasses or turning faucets had the highest mean score, implying that people had the least difficulty with these tasks.

None of the 37 items had DIF related to gender, dominance of affected shoulder, surgical status, or self-reported race using the larger DIF criteria. In all, 13 of 37 items

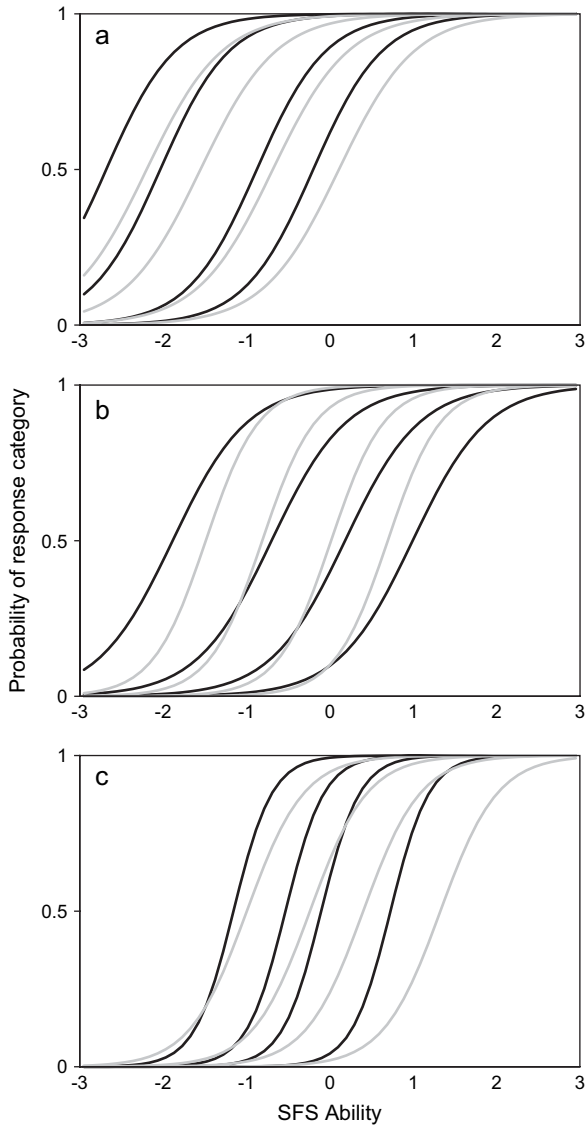

Fig. 1. Selected item boundary response functions demonstrating DIF findings. These curves represent the probability of each response category in the different groups. (a) ''Put arm on table'' in males (gray curves) and females (black curves): uniform DIF. (b) ''Comb hair'' in males (gray curves) and females (black curves): nonuniform DIF. (c) ''Lower light object'' in whites (gray curves) and nonwhites (black curves): both uniform and nonuniform DIF.

(35%) were found to have modest amounts of DIF related to at least one demographic covariate using the iterative DIF detection and IRT adjustment procedure with the smaller DIF criteria (Table 2).

There was no perceptible impact on shoulder function estimates when adjusting for DIF; the correlation between the unadjusted and fully adjusted SFS estimates was .997. Changes in the mean differences between the demographic groups were quite small (Table 3). There were increases in the differences by gender ($P < .001$) and ethnicity ($P = .003$), but these changes were not clinically relevant (i.e., $<5\%$ of a standard deviation in the scale). The distribution of differences between unadjusted and fully

adjusted scores showed a range of −3 points to +4 points; 99% of the observations had changes of 1 point or less. Item category characteristic curves for representative items with (a) uniform, (b) nonuniform, and (c) both uniform and nonuniform DIF are shown in Fig. 1.

4. Discussion

We found that none of the 37 shoulder function items had large amounts of DIF. Using criteria to detect small amounts of DIF, 13 of the 37 items had DIF related to at least one demographic covariate (gender, dominance of affected shoulder, surgery status, or race). The item category characteristic curves produced by these analyses confirmed the presence of modest amounts of DIF, as well as the nature of the DIF present (uniform or nonuniform); however, we adjusted SFS measures for DIF and found that unadjusted and adjusted SFS measures were almost identical. These results suggest that the amount of DIF in these items is negligible and that the items can be used in a CAT framework without adjusting for DIF.

In educational testing settings, items detected with DIF are often removed from consideration for inclusion in the test. This is not an attractive option for many medical testing settings, in which tests are often short and where removing items would threaten the scale's ability to differentiate patients with different functional abilities. It is advantageous for CAT item banks to be relatively large, because this allows the computer algorithm to be choosy in identifying the item that best targets the respondent's trait level. Removing a large number of items that display inconsequential DIF would impose limitations. It is thus important to use techniques for detecting both whether DIF is present and whether DIF makes any substantive difference in estimates of patient ability. Our results suggest that the modest amounts of DIF we detected in the shoulder function items had no practical consequence on estimates of patient SFS ability.

Items that have DIF may nevertheless be useful indicators of the attribute or ability the test is trying to measure [13]. If significant DIF were evident, it might be appropriate to use different parameters according to the respondent's demographic characteristics. One advantage of the technique for DIF detection outlined here is that the technique provides appropriate parameter estimates for each item within each subgroup. It may be possible to integrate subgroup-specific item parameters into a CAT algorithm, resulting in ability estimates that are adjusted for DIF. Our analysis did not support a need to adjust estimates of SFS measures for DIF.

An additional advantage of our iterative ordinal logistic regression and IRT technique for DIF detection, compared with traditional IRT-based DIF detection techniques, is flexibility in the selection of criteria for uniform DIF detection. IRT-based techniques developed to date rely exclusively on tests of statistical significance. In large samples, statistically significant DIF may represent practically irrelevant DIF. The DIFdetect program permits the user to select a test of statistical significance or a change in regression parameter criterion and lets the user identify critical values. For the present study, we used both a 10% and a 5% change in parameter criterion. None of the shoulder items showed uniform DIF at the 10% change level. When we modified scores to take into account the items that had uniform DIF at the 5% change level, adjusted SFS ability estimates were nearly identical to unadjusted SFS ability estimates. These techniques let the user determine whether detecting and accounting for smaller amounts of DIF have clinically relevant impact on patient ability estimates for groups or for individuals.

All models converged on solutions using the expected a posteriori [14] scoring algorithm with the 400 individuals in the present study. Larger samples would lead to more precision in the parameters estimated. The finding that none of these items had meaningful DIF should be confirmed with larger datasets. Additionally, other covariates such as age and literacy level should be examined for DIF as well.

In summary, we present an illustration of a technique for identifying items with DIF, adjusting IRT models for DIF, and determining the practical consequences of that adjustment on patient ability estimates. We argue that in the scaling of health outcome measures, the most important question regarding DIF is not whether it is present, but whether the DIF has practical consequences on either patient ability estimates or on substantive conclusions. In the case of the 37-item shoulder functional status test, although several items appeared to have modest amounts of DIF, this DIF did not make a difference in individual patient ability estimates or in the relationship between SFS measures across demographic groups.

Acknowledgments

P.K.C.'s time was supported by grant no. AG K08 022232 from the U.S. National Institute on Aging. L.E.G.'s time was supported by grants no. AG K08 022232 and 5-P50 AG05136-17 from the U.S. National Institute on Aging.

References

[1] Millsap RE, Everson HT. Methodology review: statistical approaches for assessing measurement bias. Appl Psychol Meas 1993;17: 297–334.

[2] Camilli G, Shepard LA. Methods for identifying biased test items. Thousand Oaks, CA: Sage; 1994.

[3] Holland PW, Wainer H, editors. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993.

[4] Hart DL, Cook KF, Mioduski JE, Teal CR, Crane PK. Simulated computerized adaptive test for patients with shoulder impairments was efficient and produced valid measures of function. J Clin Epidemiol 2006;59:290–8.

[5] Cook KF, Roddey TS, Gartsman GM, Olson SL. Development and psychometric evaluation of the Flexilevel Scale of Shoulder Function. Med Care 2003;41:823–35.
[6] Samejima F. Estimation of ability using a response pattern of graded responses. Psychometrika 1969 (Monogr Suppl 17).
[7] Samejima F. Graded response model. In: van der Linden WJ, Hambleton RK, editors. Handbook of modern item response theory. New York: Springer-Verlag; 1997. 85–100.
[8] PARSCALE for Windows [Computer program]. Version 4.1. Lincolnwood, IL: Scientific Software International; 2003.
[9] Crane PK, Jolley L, van Belle G. DIFdetect [Computer program]. University of Washington; 2002. Available at: http://www.alz.washington.edu/DIFDETECT/welcome.html.
[10] StataCorp. Stata statistical software. Release 8.0. College Station, TX: Stata Corporation; 2003.
[11] Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: differential item functioning in the CASI. Stat Med 2004;23:241–56.
[12] McCullagh P, Nelder JA, editors. Generalized linear models. 2nd ed. London: Chapman and Hall; 1989.
[13] Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol Bull 1993;114:552–66.
[14] Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. Appl Psychol Meas 1982;6:431–44.