Advancing Measurement for Low Back Function

Expanding the FOTO Low Back Functional Status (Low Back FS) Item Bank:

A Brief Report

Michael A. Kallen, PhD, MPH
Psychometric Research Scientist, Net Health Systems, Inc., Pittsburg, PA
Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Daniel Deutscher, PT, PhD
Senior Research Scientist, Net Health Systems, Inc., Pittsburgh, PA
Maccabitech Institute for Research & Innovation, Maccabi Healthcare Services, Tel-Aviv, Israel

Deanna Hayes, PT, DPT, MS
Director Clinical Outcomes and Research, Net Health Systems, Inc., Pittsburg, PA

October 2022

Table of Contents

EXECUTIVE SUMMARY

The item bank seeding process and evaluation were part of a planned item bank development and maintenance effort. This report has two parts.

PART 1: New seeded items. The *FOTO Low Back Functional Status (FS)*, also known as "the Low Back CAT," is an item response theory patient-reported outcomes measure that began as an item bank consisting of 25 items,[1-4] to which were added three new "seeded" items. Items added were those with content identified by clinicians and patients as important to include in the bank, and that were successfully calibrated on the existing Low Back FS metric. The calibration process maintained ***the same metric, enabling score compatibility between versions.*** Briefly, with the original item calibrations established to serve as anchorable item parameters for subsequent item seeding efforts and analyses, new items were then seeded in a single phase. Using the original 25 items and their calibrations as anchoring item parameters, three new items ***were successfully calibrated and added to the item bank, creating the final 28-item bank.***

PART 2: Analyses to determine value added. After including three new items, we assessed the specific value they added to the item bank, comparing the 28-item to the original 25-item bank. This part reviews the assessment of the improvements following the addition of new item content that clinicians and patients felt was (a) missing from the original item bank and (b) important for patient self-evaluation when reporting on low back functional status. The assessment focused on improved reliability, improved computer adaptive test (CAT) performance, and expanded score coverage.

Internal consistency reliability, as well as score-level reliability, increased slightly with the original plus new seeded items. Score-level reliability increased mostly for lower and higher Low Back FS scores, i.e., lower and higher levels of FS.

CAT performance was also improved, needing a slightly lower average number of items, measured with a slightly lower average measurement error, across a slightly extended score range. We observed ***impressive new seeded item usage*** with the CAT. For the 28-item CAT, all three of the new seeded items were administered. For the new seeded items, two had usage rates $\geq 20\%$ and $< 30\%$ (i.e., 22.6% and 24.0%), and one had a usage rate $\geq 10\%$ and $< 20\%$ (i.e., 15.5%). Approximately one tenth (10.3%) of the CAT items administered were new seeded items.

Finally, the Low Back FS ***measurement continuum was extended*** within several item parameter threshold maximum value ranges, which provides proof of an ***improved range and density of item coverage*** of the now expanded measurement continuum of the 28-item bank.

PART 1: New seeded item analyses (samples, anchoring, seeding process)

1.1 Background

As part of ongoing measure maintenance, the FOTO Low Back FS item bank was expanded for the purpose of evolving measurement properties, including the bank's clinical content and measurement coverage, reliability, and the CAT administration process.

This maintenance process included several analytical steps that are described below.

All seeding analyses were conducted using a 1-parameter item response theory (IRT)-related model, the *Rasch rating scale model (RSM).* IRT is a method for scoring items that considers one or more parameters on which items are characterized. The Rasch model considers the level of difficulty represented by each item.

1.2 Establish baseline item calibrations as anchoring parameters

Method

Prior to adding new seeded items to the item bank, the original item parameters needed to be obtained to serve as an overall anchor for the new item parameters. This process ensures that the updated measure's metric remains the same as compared to the original metric, enabling score comparison between measure versions.

The original item bank included 25 items. Their item labels, descriptions, and item locations (difficulty levels) are shown in the **Appendix. N**ew seeded items are marked in **bold**. Items are sorted by location.

The 6-category response options (and scoring values) for nine items (WORK, HOBBY, HEAVY, BENDING, SHOES, LIFT, STAND, STAIRS, and DRIVE) were:

Unable to perform activity (1)
Extreme difficulty (2)
Quite a bit of difficulty (3)
Moderate difficulty (4)
A little bit of difficulty (5)
No difficulty (6)

The 3-category response options (and scoring values) for 16 items (VIGOR, MODERATE, LIFTGROC, STAIRSPF, STAIR1PF, MILE, BLOCKS, ONEBLOCK, BATHING, LIFTOVER, SPORT, VACATION, CULTURAL, CHAIR, WALKRM, and BED) were:

Yes, limited a lot (1)
Yes, limited a little (2)
No, not limited at all (3)

The 5-category response options (and scoring values) for the three new seeded items (BROOM, SIT TO STAND, and GETUP&DOWN FLOOR) were:

Extreme difficulty or unable to perform (1)
Quite a bit of difficulty (2)
Moderate difficulty (3)
A little bit of difficulty (4)
No difficulty (5)

Results

The 25 original item calibrations were obtained and coded as "constrained" to retain their original values in any subsequent new-item calibration; thus they became available to serve as anchorable item parameters for all subsequent item seeding efforts and analyses.

1.3 Item seeding

Development of content for item seeding

Following a period of clinical use of the original 25-item bank, clinician users provided feedback that the item bank was missing certain concepts that their patients said were relevant and meaningful when experiencing low back impairments. In response to clinician/patient feedback, three new items were developed by a small panel of physical therapist researcher scientists with experience treating adults with low back pain; the new items were formally added to the Low Back item bank in 2013.

The original seeding data collection took place during 2006-2009. A single seeded item was administered to all patients with low back impairments for a 3-month period and then removed from the system. This process was repeated for each seeded item, one at a time. Analytic work took place in a single phase as described below.

Step 1 method: Seeding three new items

Phase 1 addressed a total of 45,752 Low Back FS CAT surveys collected with the three new seeded items that were administered to all patients during the data collection period.

Response data requirement for items to be seeded:
A minimum of 10 responses for each item response category option was required for items to be included in seeding analyses.

Response options and scoring values were unique for the new seeded items. Note that the *Rasch RSM* is perfectly capable of analyzing and estimating item parameters across multiple response category sets within a measure.

Step 1 results:

Using the 25 original item calibrations as anchoring item parameters, the three new items, which all met the minimum of 10 responses per category data requirement, underwent item response theory-based analyses, were successfully calibrated, and, therefore, were added to the Low Back FS item bank.

Scaling

For the original 25-item bank, the measure was scaled to have a score range of 0-100, with higher scores indicating better functional status. For the updated measure, including the original plus seeded items, this range was slightly expanded. ***This is a unique strength of IRT approaches over older methods utilizing classical test theory approaches, as it allows scores from the original scale to be compared directly to scores from the expanded item bank.***

Note: The possible score range using CAT may differ slightly from the possible range of scores when the full item bank is administered. Further, CAT administration parameters, such as stopping rules, may be adjusted from time to time while maintaining the same scoring continuum (i.e., metric).

PART 2: How this work added value to the Low Back FS item bank: Psychometric evidence

2.1 Introduction

The following is a description of the results of analyses providing evidence of the added value of the new Low Back FS items and the expanded Low Back FS item bank. The Low Back FS measure began as an item bank consisting of 25 items, to which were subsequently added a total of three new "seeded" items. Seeded items were those (a) with content identified by clinicians and patients as important to include in the Low Back FS item bank, (b) whose item-level data met minimum $n \geq 10$ responses per category requirements, and (c) that successfully passed IRT-based analytic requirements, were successfully calibrated on the existing Low Back FS metric, and, thus, were available to be added to the item bank.

The item bank item seeding process and evaluation were part of a planned item bank development and maintenance effort. After including new items, we then assessed the specific value added by the three seeded Low Back FS items, comparing the now 28-item Low Back FS item bank's performance to the original 25-item Low Back FS item bank's performance.

2.2 Added important new item content

First, we considered the original vs. original plus seeded item content of the Low Back FS item bank. Clinicians and patients had identified specific content they felt was (a) missing from the original item bank, and (b) important for patient self-evaluation when reporting on their low back function status. For this unaddressed content, new "seeded" items were written, tested, and, when appropriate, calibrated and included in the expanded Low Back FS item bank. Item content for the original and new items (in **bold**) are described in the **Appendix**.

2.3 Improved overall reliability

Second, we considered the Low Back FS item bank's overall reliability.

To obtain reliability estimates, we simulated N=2000 full item bank responses, assuming a normally distributed population centered on an approximate Low Back FS score of 50. We estimated classical test theory's Cronbach's alpha and IRT's standard error-based reliability, both internal consistency-type estimates of overall measure reliability.

Internal consistency reliability, already considered excellent at approximately 0.966 and 0.978, increased slightly with the original plus new seeded items, as compared to the original items only (**Table 1**).

Table 1: Improved internal consistency reliability: Original vs. Original Plus Seeded Items

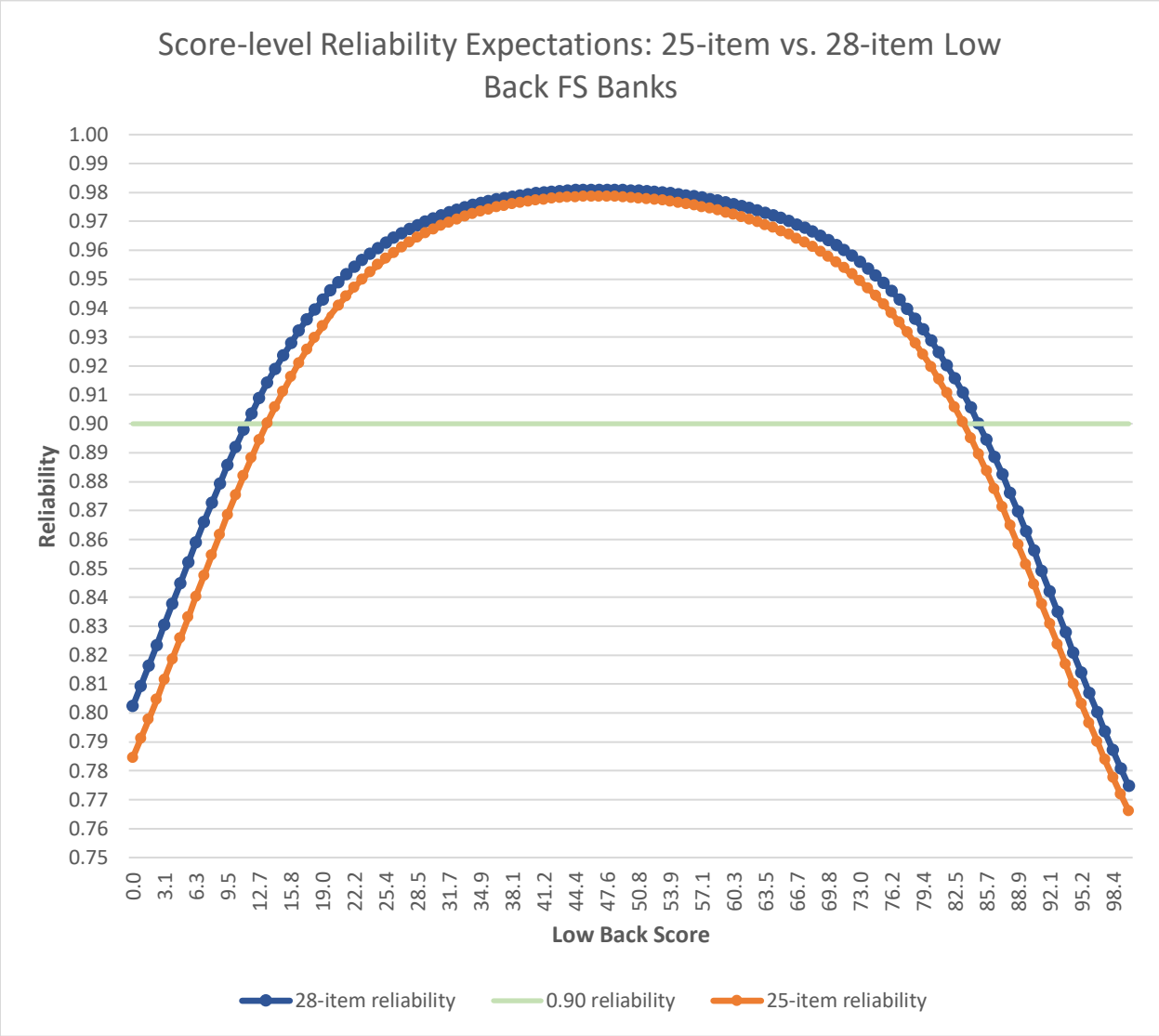
Added Value	Original Only	Original Plus Seeded
Cronbach's alpha	0.966	0.971
IRT-based reliability	0.978	0.981

2.4 Enhanced score-level reliability

Third, we considered the Low Back FS's score-level-specific reliability. We estimated score-level reliabilities across the Low Back FS measurement continuum.

Score-level-specific reliabilities increased with the original plus seeded items bank, as compared to the original items only bank. This increase in the 28-item Low Back FS item bank score-level-specific reliabilities is particularly noteworthy at scores ≤ 30 and ≥ 60 (**Figure 1**), that is, for patients with lower than average and higher than average low back functional status.

Figure 1: Enhanced Score-level Reliability: Original Only Items vs. Original Plus Seeded Items



2.5 Increased reliable score range

Fourth, we evaluated the reliable score range of the 25-item vs. the 28-item Low Back FS item banks.

For low back functional status, reliability-level-defined (e.g., ≥ 0.90) score ranges increased in width with the original plus seeded items, compared to the original items only (**Table 2**).

Table 2: Increased Reliable Score Ranges: Original Only vs. Original Plus Seeded Items

Added Value	**Original Only**	**Original Plus Seeded**
Reliability standard	***Reliable score range***	***Reliable score range***
≥ 0.80	1.6 to 96.0	-0.8 to 96.8
≥ 0.85	7.1 to 89.7	5.5 to 91.3
≥ 0.90	13.5 to 84.1	11.1 to 84.9
≥ 0.95	22.2 to 73.8	19.8 to 76.2

2.6 Improved CAT performance

Fifth, we considered the CAT performance of the 25-item vs. the 28-item Low Back FS item bank.

A CAT is a type of dynamic assessment whereby an item selection algorithm identifies the specific items a particular person should answer (i.e., items are tailored or customized per person). That is, the item selection algorithm picks each item to administer to a person in order to locate and then refine that person's estimated score; thus, the CAT can measure most precisely that specific person's status in the domain of interest while simultaneously minimizing error.

CAT performance is a useful way to understand the practical value of an item bank's items by identifying with whom and how often the CAT selects specific items for administration, with the understanding that the CAT item selection algorithm identifies and administers the most informative items targeted to the level of the domain trait being measured.

We employed the following specific CAT administration parameters: (a) start with the 25-item Low Back FS's previously identified starting item (i.e., the item having the maximum information at theta=0, which was item #1 – WORK (see the **Appendix**); (b) the minimum # of items to administer=4; (c) the maximum # of items to administer=25 or 28, depending on the item bank version being assessed; (d) stop when the CAT standard error (SE) < 0.30; and (e) stop when the theta change across three consecutive items < 0.1259.

We employed a simulated response sample, as described above (i.e., N=2000 full item bank responses, assuming a normally-distributed population centered on an approximate Low Back FS score of 50).

CAT administration performance improved slightly with the original plus seeded items, compared to the original items only (**Table 3**). Improvements were small but measurable, particularly for reduced average number of items required for patient responses, and reduced measurement error. The notable increase in maximum observed score illustrates the increased score coverage for measuring higher low back functional status.

Table 3: Improved FOTO CAT Performance: Original Only Items vs. Original Plus Seeded Items

Added Value	Original Only	Original Plus Seeded
Average # of items	6.09	6.02
Mean SE	0.536	0.531
Correlation with Full Bank	0.983	.981
Minimum observed score	6.63	6.38
Maximum observed score	93.80	98.08

2.7 New seeded item usage in the CAT administration context

Sixth, we evaluated the number of new seeded items used as well as their frequency of use in the context of a CAT administration of the Low Back FS item bank. Frequency of use of the new items is important to evaluate, because it assesses their usefulness with the intended patient population. For the 25-item Low Back FS CAT, 22 of the 25 original items were administered; three items (ONEBLOCK, VACATION, and CULTURAL) were not administered.

For the 28-item Low Back FS CAT, each of the three new seeded items was administered (BROOM, SIT TO STAND, and GETUP&DOWN FLOOR), while 20 of the 25 original items were also administered. Combined with the results above related to the original 25-item bank, this suggests that some of the new items functioned better than some of the original items. With the 28-item Low Back FS CAT, five original items were not administered, including the same three items not administered for the 25-item bank, and two additional original items (STAIR1PF and LIFTOVER). For the new seeded items, the two items SIT TO STAND and GETUP&DOWN FLOOR had usage rates (i.e., the percent of N=2000 cases who were administered the item) of ≥ 20% and < 25% (i.e., 22.6% and 24.0%, respectively); the item BROOM had a usage rate of 15.5%.

Approximately one tenth (10.3%) of the items administered by the 28-item Low Back FS CAT for the simulated sample were new seeded items (see the **Appendix**). The relatively heavy use of the new items considering their low percentage of total item bank items supports their added value.

2.8 Expanded range and more dense coverage of the Low Back FS measurement continuum

Seventh, we examined the range and density of the item parameters of the new seeded Low Back FS items by analyzing item threshold estimates. Thresholds are item specific and refer to the ability level at which a respondent has the same probability of selecting the lower or higher response category. For example, when there are six response categories, there are five thresholds for each item.

Across all threshold estimates, the minimum threshold remained at 18.34, while the maximum threshold remained at 75.82.

Within Threshold 1 estimates, the minimum threshold remained at 18.34, while the maximum threshold remained at 59.22.

Within Threshold 2 estimates, the minimum threshold remained at 26.18, while the maximum threshold remained at 77.98.

Within Threshold 3 estimates, the minimum threshold remained at 33.75, while the maximum threshold increased from 56.05 to 59.37.

Within Threshold 4 estimates, the minimum threshold remained at 41.70, while the maximum threshold increased from 64.00 to 71.14.

Finally, within Threshold 5 estimates, the minimum threshold remained at 53.52, while the maximum threshold remained at 75.82.

The increased range across several item maximum thresholds provides proof of the expanded measurement continuum of the 28-item Low Back FS item bank. In conjunction with the overall expanded score range (see **Table 3**) and the improved score-level-specific reliabilities (see **Figure 1**), this attests to the improved ability of the 28-item Low Back FS item bank to measure both lower and higher low back functional status. This finding affirmed clinician/patient feedback that the original bank was lacking in certain items representing unique and meaningful levels of functional ability. Only trivial floor (0.39%) and ceiling (0.85%) effects were observed in the item seeding sample of N=45,752 Low Back FS CAT surveys; note that floor and ceiling effects below 15% are considered acceptable.[5-8]

In addition, the increased range of several item thresholds within Thresholds 1-5, as well as the increased overlap of threshold locations across thresholds, provides proof of an improved density of item coverage of the now expanded measurement continuum of the 28-item Low Back FS item bank (**Table 4**).

Table 4: Range and density of item parameters

Threshold	Item bank version	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5
Minimum	25-item	18.339	26.183	33.746	41.698	53.520
	28-item	18.339	26.183	33.746	41.698	53.520
Maximum	25-item	59.219	77.982	56.049	64.001	75.820
	28-item	59.219	77.982	59.369	71.137	75.820

REFERENCES

1. Hart DL, Mioduski JE, Werneke MW, Stratford PW. Simulated computerized adaptive test for patients with lumbar spine impairments was efficient and produced valid measures of function. *J Clin Epidemiol.* Sep 2006;59(9):947-956.

2. Hart DL, Werneke MW, Wang YC, Stratford PW, Mioduski JE. Computerized adaptive test for patients with lumbar spine impairments produced valid and responsive measures of function. *Spine (Phila Pa 1976).* Nov 15 2010;35(24):2157-2164.

3. Wang YC, Hart DL, Werneke M, Stratford PW, Mioduski JE. Clinical interpretation of outcome measures generated from a lumbar computerized adaptive test. *Phys Ther.* Sep 2010;90(9):1323-1335.

4. Hart DL, Stratford PW, Werneke MW, Deutscher D, Wang YC. Lumbar computerized adaptive test and Modified Oswestry Low Back Pain Disability Questionnaire: relative validity and important change. *J Orthop Sports Phys Ther.* Jun 2012;42(6):541-551.

5. Brodke DS, Goz V, Lawrence BD, Spiker WR, Neese A, Hung M. Oswestry Disability Index: a psychometric analysis with 1,610 patients. *Spine J.* Mar 2017;17(3):321-327.

6. Brodke DS, Goz V, Voss MW, Lawrence BD, Spiker WR, Hung M. PROMIS PF CAT Outperforms the ODI and SF-36 Physical Function Domain in Spine Patients. *Spine (Phila Pa 1976).* Jun 15 2017;42(12):921-929.

7. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* Jan 2007;60(1):34-42.

8. Wamper KE, Sierevelt IN, Poolman RW, Bhandari M, Haverkamp D. The Harris hip score: Do ceiling effects limit its usefulness in orthopedics? *Acta Orthop.* Dec 2010;81(6):703-707.

Appendix: Item content, CAT usage, and item location (difficulty)

Version	Item label	Item description	CAT usage: n (%) from N=2000	Location
Original	SHOES	Today, because of your back problem, do you or would you have any difficulty at all putting on your shoes or socks?	593 (29.7%)	34.7
Original	BATHING	Does or would your back problem limit: BATHING or DRESSING?	162 (8.1%)	36.8
Original	DRIVE	Today, because of your back problem, do you or would you have any difficulty at all driving for 1 hour?	778 (38.9%)	37.6
Original	CHAIR	Does or would your back problem limit: Getting in and out of a CHAIR?	99 (5.0%)	38.0
Original	WALKRM	Does or would your back problem limit: WALKING around a room?	50 (2.5%)	38.0
Original	BED	Does or would your back problem limit: Getting in and out of BED?	7 (0.4%)	40.0
Seeded item	**BROOM**	**Because of your back, how much difficulty do you have using a broom?**	**310 (15.5%)**	**40.8**
Original	STAIR1PF	Does or would your back problem limit: Climbing 1 flight of STAIRS?	0 (0.0%)	42.7
Original	LIFTOVER	Does or would your back problem limit: LIFTING OVERHEAD to a cabinet?	0 (0.0%)	44.3
Original	VACATION	Does or would your back problem limit: Going on VACATION?	0 (0.0%)	45.0
Original	ONEBLOCK	Does or would your back problem limit: WALKING one BLOCK?	0 (0.0%)	45.9
Original	LIFT	Today, because of your back problem, do you or would you have any difficulty at all lifting a box of groceries from the floor?	916 (45.8%)	46.1
Original	CULTURAL	Does or would your back problem limit: Attending SOCIAL EVENTS?	0 (0.0%)	47.5
Original	STAIRS	Today, because of your back problem, do you or would you have any difficulty at all going up or down 2 flights of stairs (about 20 stairs)?	1095 (54.8%)	47.6
Original	WORK[a]	Today, because of your back problem, do you or would you have any difficulty at all performing any of your usual work, housework, or school activities?	2000 (100.0%)	48.9
Seeded item	**SIT TO STAND**	**Because of your back, how much difficulty do you have changing positions quickly like sitting to standing?**	**451 (22.6%)**	**49.4**

Original	STAND	Today, because of your back problem, do you or would you have any difficulty at all standing for 1 hour?	1258 (62.9%)	50.1
Original	BENDING	Today, because of your back problem, do you or would you have any difficulty at all bending or stooping?	1162 (58.1%)	51.8
Original	HOBBY	Today, because of your back problem, do you or would you have any difficulty at all performing your usual hobbies, recreational or sporting activities?	1084 (54.2%)	51.9
Seeded item	**GETUP&DOWN FLOOR**	**Because of your back, how much difficulty do you have getting down to and up from the floor?**	**479 (24.0%)**	**52.1**
Original	BLOCKS	Does or would your back problem limit: WALKING several BLOCKS?	1 (0.1%)	53.2
Original	STAIRSPF	Does or would your back problem limit: Climbing several flights of STAIRS?	1 (0.1%)	54.3
Original	LIFTGROC	Does or would your back problem limit: LIFTING or CARRYING items like groceries?	6 (0.3%)	54.7
Original	MODERATE	Does or would your back problem limit: MODERATE ACTIVITIES like moving a table, pushing a vacuum cleaner, bowling, or playing golf?	11 (0.6%)	56.1
Original	MILE	Does or would your back problem limit: WALKING more than a mile?	23 (1.2%)	56.2
Original	HEAVY	Today, because of your back problem, do you or would you have any difficulty at all performing heavy activities around your home?	1113 (55.7%)	57.0
Original	SPORT	Does or would your back problem limit: Participating in RECREATION?	181 (9.1%)	62.6
Original	VIGOR	Does or would your back problem limit: VIGOROUS ACTIVITIES like running, lifting heavy objects, participating in strenuous sports?	268 (13.4%)	68.6

Items are sorted by ascending location (item difficulty level).
Seeded items are marked in **bold**.
[a]The CAT starting item