ORIGINAL ARTICLE

The Symptom Inventory Disability-Specific Short Forms for Multiple Sclerosis: Construct Validity, Responsiveness, and Interpretation

Carolyn E. Schwartz, ScD, Rita K. Bode, PhD, Brian R. Quaranto, BS, Timothy Vollmer, MD

ABSTRACT. Schwartz CE, Bode RK, Quaranto BR, Vollmer T. The Symptom Inventory Disability-Specific Short Forms for multiple sclerosis: construct validity, responsiveness, and interpretation. Arch Phys Med Rehabil 2012;xx:xxx.

Objectives: To test the cross-sectional and longitudinal construct validity of the disability-specific short forms of the Symptom Inventory for multiple sclerosis, to compare its internal consistency reliability and construct validity with those of the original (1999) 29-item short form of the Symptom Inventory, and to provide for the new disability-specific short forms interpretation metrics for use in sample size calculation for future research.

Design: A Web-based longitudinal study, with data collected at baseline and 6 months after baseline. Correlations evaluated the overlap among disease-specific and generic patient-reported outcome measures. Responsiveness was assessed by using symptom transition scores and modified standardized response means. Interpretation guidelines were provided by using Cohen's effect size and crosswalks to the disease-specific and generic quality-of-life measures.

Setting: National Multiple Sclerosis Registry.

Participants: People with multiple sclerosis (N=1142) who participated in the North American Research Committee on Multiple Sclerosis Registry.

Interventions: Not applicable.

Main Outcome Measures: The Symptom Inventory; the disease-specific Performance Scales and the Patient-Determined Disease Steps; the generic Short Form 12.

Results: The Symptom Inventory evidenced convergent and divergent validity, and the disability-specific short forms evidenced similar psychometric performance as the 1999 short form but had slightly better alpha reliability. They also evidenced moderate responsiveness to clinically important change, with more responsiveness among those with mild and moderate disabilities than among those with severe disabilities. Effect sizes were larger among patients who reported symptom improvement, rather than deterioration, suggesting that the tool would be more useful in clinical research focused on testing whether an intervention improves symptom experience, particularly for patients with mild and moderate disabilities. Crosswalks provided graphic and numeric links between the Symptom Inventory and other patient-reported outcomes.

Conclusions: The Symptom Inventory can be useful for elucidating the patient's experience, but it varies considerably across and within disability groupings. Directions for future research are discussed.

Key Words: Measurement; Outcomes; Psychometrics; Quality of life; Rehabilitation; Symptoms.

THE VALIDATION PROCESS for a new health-related quality-of-life measure should be extensive and iterative. In addition to testing for and demonstrating reliability and a number of aspects of validity, a number of other considerations have become increasingly prominent as measurement science has evolved. Measures must also demonstrate responsiveness to clinically important change[1-5] and must provide effect size estimates to guide the interpretation of scores over time.[6,7] With the successful application of item response theory methods to measurement validation, the benefits of developing linear scales improve precision, responsiveness, and effect size assessment. There has consequently been an increasing focus on meeting the assumptions of these methods (ie, unidimensionality and conditional independence) as well as characterizing item parameters (eg, threshold difficulty and discrimination).[8] Finally, increased emphasis on the Federal Drug Administration Guidance[9] for patient-reported outcomes[8] has led to more interest in symptom measures, which have certain caveats that must be considered in the validation process.[10]

Variability or change in the characteristics of the disease or condition being measured can also extend the validation process required for a health-related quality-of-life measure. The disease or condition may be variable and somewhat unpredictable and may be changed by disease-modifying agents. In multiple sclerosis (MS), for example, 80% to 90% of patients experience a disease course characterized by periods of symptom exacerbation followed by symptom remission (ie, relapsing–remitting disease).[11] Both these considerations play a substantial role in MS, an autoimmune disease that affects the central nervous system and can have variable and diverse symptom effects.[12]

List of Abbreviations

MS	multiple sclerosis
NARCOMS	North American Research Committee on Multiple Sclerosis
SF-12	Medical Outcomes Survey – 12 item short form

From DeltaQuest Foundation, Inc, Concord, MA (Schwartz, Bode, Quaranto); Departments of Medicine and Orthopaedic Surgery, Tufts University Medical School, Boston, MA (Schwartz); Department of Physical Medicine and Rehabilitation, Feinberg School of Medicine, Northwestern University, Chicago, IL (Bode); Department of Neurology, University of Colorado Denver, Denver, CO (Vollmer); and Rocky Mountain Multiple Sclerosis Center, Aurora, CO (Vollmer).

Funded in part by a Metric Development and Validation Award (USAMRAA grant number MS090018) and by a CMSC/Global MS Registry Visiting Scientist Fellowship, which was supported through a Foundation of the Consortium of Multiple Sclerosis Centers grant from EMD Serono, Inc. CMSC/Global MS Registry is supported by the Consortium of Multiples Sclerosis Centers and its Foundation.

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit on the authors or on any organization with which the authors are associated.

Reprint requests to Carolyn E. Schwartz, ScD, DeltaQuest Foundation, Inc, 31 Mitchell Rd, Concord, MA 01742, e-mail: *carolyn.schwartz@deltaquest.org*.

0003-9993/12/xxx-01286$36.00/0
doi:10.1016/j.apmr.2012.01.012

Our group has been further developing the Symptom Inventory measure of the MS symptom experience to be used in clinical research and clinical trials. The original measure had 99 items, and its subscales were derived on the basis of expert neurologist judgment.[13] The initial validation used logistic regression to identify a 29-item short form that was able to discriminate known groups as defined by clinician-rated disability scores.[13] The measure demonstrated good internal consistency reliability, test-retest stability, construct validity, and discriminant validity.[13]

More than a decade has passed since the initial validation study, and the landscape of MS therapeutics has substantially been transformed along with the natural history of MS.[11,14] There are a number of partially effective disease-modifying agents for MS that are particularly efficacious when used early in the course of disease.[14] If clinicians and clinical researchers have a useful symptom measure that does not present an excessive burden for patients, then earlier intervention may be more feasible. In addition, psychometric methods have evolved in that time. Item response theory[15] methods have been applied to health-related quality-of-life measures and utilized to create static short forms and dynamic computerized adaptive tests, both of which have increased precision with less subject burden.[16] Finally, psychometric theory has also advanced in that time, particularly our understanding of the distinction and importance of *effect versus causal indicators*.[10,17-19] Effect indicators (eg, anxiety, depression, and physical functioning) *reflect* the level of quality of life, whereas causal indicators (eg, symptoms and treatment side effects) *cause* a change in quality of life.[19] This understanding lends itself to different ways of thinking about the construct validity of a symptom measure, as well as distinct analytic approaches to assessing an instrument's validity.[17,20] For example, effect indicators such as quality-of-life items would have similar relationships with each other over time. This would be reflected in a stable factor structure. In contrast, causal indicators such as fatigue and pain symptom items would have different relationships with each other across different patient samples (eg, MS vs cancer) because they would be reflecting distinct clinical entities. Furthermore, correlations over time with evaluative functional health scores (eg, quality-of-life outcomes) could differ if the 2 measurement time points corresponded to a relapsing state in one instance and a remission state in another; if, for example, pain symptoms were a constant experience but fatigue symptoms varied as a function of whether the patient was in a relapse. The intercorrelation of the 2 types of symptoms would be unstable over time due to this third variable.

In our companion article,[21] we describe work done to develop such a set of tools that would build upon the information provided by a brief MS-specific disability measure used as a screening tool, the Performance Scales.[13] That work focused on factor structure and item selection for mild, moderate, and severe disability levels. Building on that work, the present article aimed to test the cross-sectional and longitudinal construct validity of the disability-specific short forms of the Symptom Inventory for MS, to compare its internal consistency reliability and construct validity with those of the original (1999) 29-item short form of the Symptom Inventory,[13] and to provide for the new disability-specific short forms interpretation metrics for use in sample size calculation for future research.

We hypothesized that the correlation between the Symptom Inventory Short Form Summary score would be highest (ie, convergent validity) for the other MS-specific measures (ie, the Performance Scales Sum and Patient-Determined Disease Steps), moderate for the generic quality-of-life measure (Medical Outcomes Survey – 12 item short form [SF-12] Physical Component score and Mental Component score), and lowest (ie, divergent validity) for a non–quality-of-life measure, the Godin Leisure Time Exercise Questionnaire.[22] We hypothesized that the convergent validity correlations would be highest among respondents who provided both sets of data within 1 week (Week-Lag subgroup), as compared with the whole sample. We also expected that the new Symptom Inventory Disability-Specific Short Forms would perform psychometrically similarly to the original Symptom Inventory Disability-Specific Short Forms in terms of the internal consistency reliability and the intercorrelations with other patient-reported outcomes. We expected that there would be significant differences on the Symptom Inventory Short Form Summary score and the Symptom Inventory Short Form subscales by transition group, supporting the measure's responsiveness to clinically important change.

METHODS

This article builds on work reported in our companion article.[21] For the sake of clarity and completeness, methods will be briefly summarized below and the interested reader is referred to the above-mentioned companion article for more detail.

Sample and Design

This project involved longitudinal data collected at baseline and 6 months after baseline from 1142 people with MS who participated in the North American Research Committee on Multiple Sclerosis (NARCOMS) Registry and an add-on survey implemented by our group. Eligible participants resided in the United States, were at least 18 years old, and as of spring 2010 had completed the latest 2 semiannual NARCOMS update surveys online. The sample was stratified to have an equal distribution by sex, age, course of disease (relapsing or not), and level of disability as measured by the most recent Patient-Determined Disease Steps[23] score (0–2 mild, 3–4 moderate, 5–8 severe), a tool with demonstrated construct validity,[24] showing a similar magnitude of correlation with patient-reported outcomes as clinician-reported indices. In addition to the data collected twice over a 6- to 9-month window of time, NARCOMS provided patient-reported outcome data on the study participants collected semiannually over the past 5 years. These data were used to compute trajectory scores as described in early responsiveness analyses.

Procedure

Supplementary data were collected by using the Web-based survey engine SurveyGizmo.com[a] (www.surveygizmo.com), a user-friendly and Health Insurance Portability and Accountability Act–compliant interface for collecting data in a secure environment. The SurveyGizmo questionnaire began with an online (written) consent form and the measures described under Measures. Data from the NARCOMS Registry spring and fall updates and the SurveyGizmo data were then linked by the NARCOMS Registry Coordinating Center by using a unique identifier, and data were de-identified prior to data analysis. The project was reviewed and approved by the institutional review boards associated with NARCOMS Registry (the Western Institutional Review Board, Olympia, WA) and DeltaQuest Foundation (the New England Institutional Review Board, Newton, MA).

Measures

In addition to the Symptom Inventory Disability-Specific Short Forms, participants completed disease-specific and ge-

neric health-related quality-of-life measures. The disease-specific measures included the Patient-Determined Disease Steps[23] and the Performance Scales. The generic SF-12[25] was also completed, and scored to generate Physical and Mental Component scores. A transition item querying perceived symptom change over the past 6 months was included for the planned responsiveness analyses. A full description of other patient-reported outcome measures is provided in the companion article.[21]

Statistical Analysis

Selection bias. To understand possible selection biases in the baseline and follow-up samples, we implemented logistic

Table 1: Comparison of Baseline and Follow-up Sample Demographic Characteristics*

Variable	Provided Spring SI 2010 Data (%) (N=1510)	Provided Fall SI 2010 Data (%) (N=1142)	Odds Ratio	95% Confidence Interval
Sex: % Female	74.9	75.1	1.06	0.81–1.38
Mean age ± SD (y)	54.2±9.4	54.4±9.2	1.01	0.99–1.02
Marital status				
Never married	9.00	8.60	0.8	0.53–1.22
Married	71.00	71.40	1.12	0.85–1.47
Divorced	11.00	10.90	0.91	0.62–1.35
Widowed	3.40	3.60	1.43	0.66–3.08
Separated	1.40	1.30	0.84	0.30–2.34
Cohabitation/domestic partner	4.20	4.20	0.92	0.50–1.70
Employment status				
Full-time	25.60	25.60	1.02	0.77–1.36
Part-time	12.90	13.20	1.19	0.81–1.76
Not employed	61.50	61.10	0.91	0.70–1.17
Annual income				
<$15,000	6.00	5.40	0.63	0.39–1.02
$15,001 to $30,000	11.60	12.30	1.39	0.91–2.12
$30,001 to $50,000	15.40	15.00	0.88	0.63–1.23
$50,001 to $100,000	29.70	30.00	1.04	0.79–1.37
Over $100,000	19.10	19.80	1.26	0.90–1.75
Do not wish to answer	18.10	17.40	0.83	0.61–1.14
Income change in past 6 months				
Yes	19.40	20.90	1.59*	1.12–2.25
No	75.40	73.70	0.64*	0.47–0.88
Do not wish to answer	5.20	5.30	1.22	0.67–2.22
If income changed, how?				
Increased	31.20	29.20	0.63	0.32–1.21
Decreased	63.40	64.80	1.33	0.69–2.54
Lost all income	5.40	5.90	2.80	0.35–21.85
Residence status				
Private home	96.80	97.10	1.45	0.76–2.82
Private home with home health aid	2.50	2.30	0.63	0.31–1.30
Assisted living	0.30	0.20	0.28	0.04–2.00
Nursing home†	0.30	0.40		
Mean BMI ± SD	27.4±6.5	27.4±6.5	1.01	0.98–1.03
% Underweight (BMI<18.5)	3.10	3.00	0.86	0.43–1.72
% Normal weight (BMI 18.5–25)	38.10	37.10	0.85	0.66–1.09
% Overweight (BMI 25.1–30)	30.70	31.10	1.10	0.84–1.45
% Obese (BMI >30)	28.00	28.70	1.13	0.85–1.50
Alcohol use				
Never	31.00	30.50	0.89	0.68–1.16
Monthly or less	26.80	27.20	1.08	0.81–1.43
2–4 times per month	19.60	19.00	0.89	0.66–1.22
2–3 times per week	10.90	11.20	1.11	0.74–1.68
4 or more times a week	11.70	12.00	1.22	0.81–1.83
Smoking				
No, not at all	88.70	88.50	0.99	0.66–1.47
Yes, some days	3.30	3.60	1.23	0.59–2.56
Yes, every day	7.90	7.90	0.94	0.59–1.48
Cigarettes per day (among smokers only, n=163), mean ± SD	12.9±8.1	12.9±8.2	0.99	0.94–1.03
Secondhand smoke exposure: % Yes	10.10	9.90	0.90	0.60–1.36

Abbreviations: BMI, body mass index; SI, Symptom Inventory.
*P=.006; Bonferroni adjustment: 0.05/43=0.0012; therefore, detected difference is not significant.
†This statistical comparison was not possible because the sample size was too small. There was no difference in the number of people (n=5) who endorsed living in a nursing home in the Spring and Fall data collections.

regression analyses to evaluate whether any demographic or health behavioral attribute differentiated the 2 samples. If the odds ratio was significant, we concluded that the 2 samples were different in this regard and that this difference could reflect bias.

Cross-sectional construct validity. The cross-sectional construct validity of the Symptom Inventory Short Forms was evaluated by examining the overlap between the Symptom Inventory Short Forms scores and the other patient-reported outcomes. We computed Pearson correlation coefficients in (a) the whole sample and (b) the subsample of respondents whose NARCOMS and supplemental survey were conducted within 8 days of one another (the "Week-Lag" subgroup). We also compared the original Symptom Inventory Short Form[13] with the newly derived Symptom Inventory Short Forms, in terms of both their internal consistency reliability and their intercorrelations with other patient-reported outcomes testing their convergent and divergent validity as shown in earlier work by our group.[13,26]

Longitudinal construct validity. The longitudinal construct validity[1] was evaluated by assessing responsiveness, using patient-reported symptom transition scores over 6 months. We computed the modified standardized response mean, which is the mean change in scores divided by the standard deviation of change scores in patients defined as stable.[27] We computed this responsiveness index separately for patients who reported better versus worse symptoms based on previous prospect-theory–based research suggesting that people value gains differently than they value losses.[28,29] Better-versus-worse symptom change groups were compared by using linear regression predicting the Symptom Inventory Short Form subscale change score (dependent variable) with dummy variables representing better or worse symptom transition (referent is no-change group). These comparisons were done separately for Patient-Determined Disease Steps–derived disability subgroups (mild, moderate, severe), for the Symptom Inventory Short Forms disability-specific subscales, and for the Performance Scales items. We also compared the original Symptom Inventory Short Form[13] with the newly derived Symptom Inventory Short Forms in terms of how responsive their subscales were to change on the Performance Scales items.

Interpretation. To facilitate the use and interpretation of the Symptom Inventory Disability-Specific Short Forms in future research, 2 metrics for comparison were computed. First, Cohen's effect size[30] was computed separately for patients who reported better and worse symptom change. This statistic was the better (or worse) mean minus the same mean divided by better (or worse) standard deviation. The magnitude of this statistic is interpretable according to Cohen's standard effect size guidelines, with a small effect size of 0.20 to 0.49, a medium effect size of 0.50 to 0.79, and a large effect size of 0.80 and greater. Second, we created a crosswalk by using equipercentile equating[31,32] to facilitate linkage with other patient-reported outcomes. This crosswalk illustrates how scores on the Symptom Inventory Disability-Specific Short Forms relate to scores on the other disease-specific and generic measures.

RESULTS

Sample Characteristics and Selection Bias

This longitudinal study had a 76% follow-up rate, with 150 (13%) participants providing NARCOMS and SurveyGizmo data within 1 week of each other (ie, the Week-Lag subsample). The sample had 75% women, with a mean age of 54 years (see table 1). Most participants were married and not employed and lived in a private home. Their median income ranged between $50,000 and $100,000, and it had not changed in the past 6 months. The mean body mass index was in the overweight range, with about 38% of the sample being of normative weight. Most people reported minimal alcohol use (monthly or less or never), did not smoke, and were not exposed to secondhand smoke.

Table 1 summarizes the results of the univariate logistic regressions examining selection bias. Only 1 variable of the 12 examined (and 43 comparisons) showed a significant difference between the baseline and follow-up groups: patients who did not provide follow-up data were slightly more likely to report having had an income change in the past 6 months (20.9% vs 19.4%). However, after adjusting our P value for the number of comparisons computed,[33] this comparison was not significant. From these analyses, we conclude that the baseline and follow-up sample show no signs of bias across the demographic characteristics examined.

Cross-Sectional Construct Validity

Figure 1 shows the intercorrelations and 95% confidence-interval error bars of the Symptom Inventory Short Form Summary score with the disease-specific and generic measures by disability grouping on the Patient-Determined Disease Steps

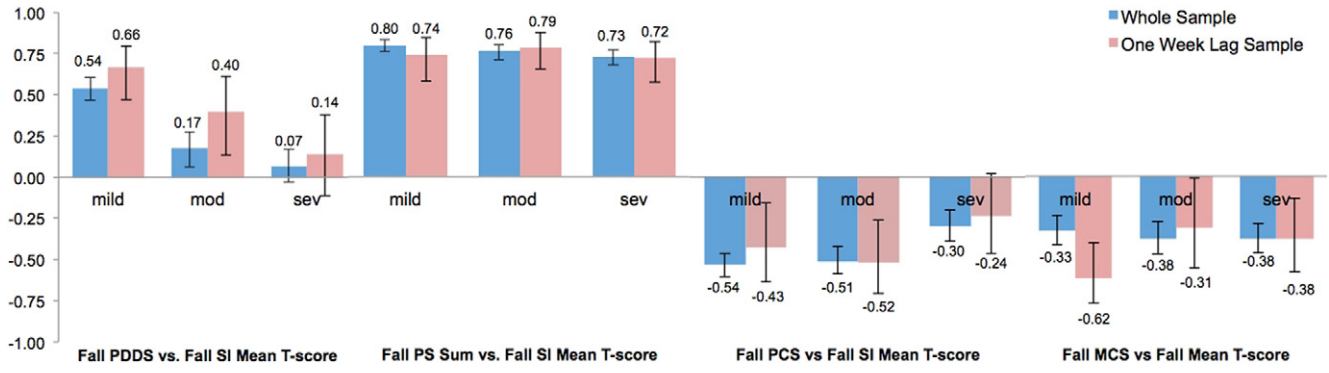


Fig 1. Convergent construct validity correlations. A comparison of correlation coefficients is shown by disability groupings by using the PDDS. The intercorrelations and 95% confidence-interval error bars of the SI Short Form Summary score with the disease-specific and generic measures are shown by disability grouping on the PDDS (mild, moderate, and severe), and among the whole sample in the blue bar (n=992), as compared with the Week-Lag subsample in the pink bar (n=150). In addition to the PDDS, the disease-specific measures included the PS. The generic measures included the SF-12 PCS and MCS. Abbreviations: MCS, Mental Component score; mod, moderate; PCS, Physical Component score; PDDS, Patient-Determined Disease Steps; PS, Performance Scales; sev, severe; SI, Symptom Inventory.

(mild, moderate, and severe), and among the whole sample, as compared with the Week-Lag subsample. The correlations between the Symptom Inventory Summary score and the other outcomes appeared to differ by severity grouping, with the highest correlations among the subgroup with mild disability, lower correlations among the subgroup with moderate disability, and lowest correlations among the subgroup with severe disability. The pattern was similar among the Week-Lag respondents, with regard to scores on the Patient-Determined Disease Steps, Physical Component score, and Mental Component score. As hypothesized, the Symptom Inventory Summary score was most highly correlated with the Performance Scales disease-specific measure and less highly correlated with the generic SF-12 Physical Component score and Mental Component score. Of note, the Patient-Determined Disease Steps was highly correlated with the Symptom Inventory Summary score only in the group with mild disability. In addition, the correlation between the measures was more attenuated for respondents with severe disabilities.

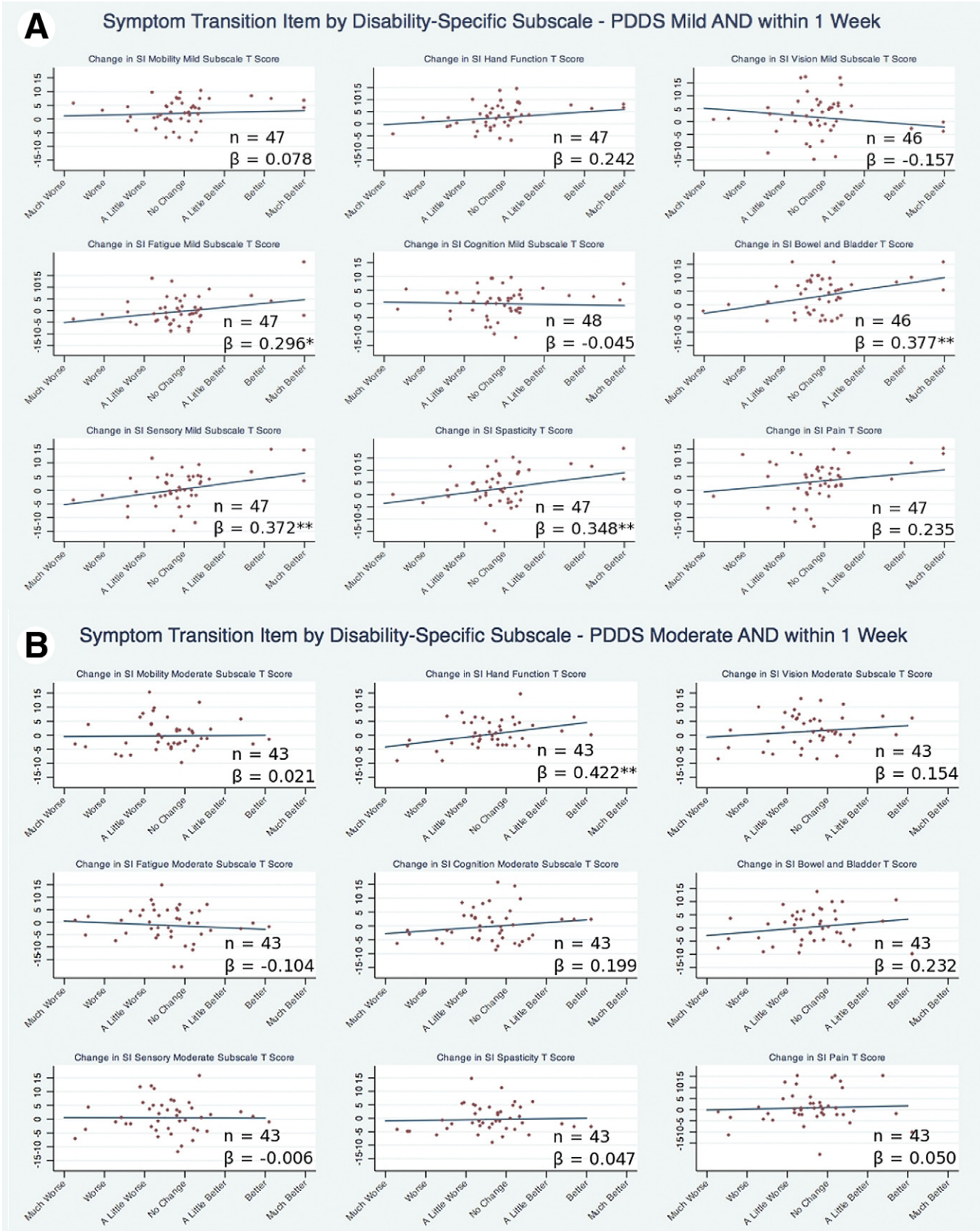
Comparison of the new and original Symptom Inventory Short Form. The next set of analyses compared the 29-item short form of the Symptom Inventory that was originally developed in 1999[13] with the newly developed disability-specific short forms described in this article and the companion article.[21] The former was developed by using logistic regression to discriminate known groups on the basis of a clinician-observed measure. The latter was developed by using item response theory analyses using patient-reported outcome measures for construct validation.

The 2 measures performed similarly psychometrically. The correlation between the new Symptom Inventory Summary score and the 29-item version score as originally published[13] was high (r=.90 in spring, .92 in fall) but different enough to conclude that the 2 scores were not identical. The alpha reliability of the original

Table 2: Mean Score Change and MSRM for Symptom Inventory Subscales by Patient-Determined Disease Steps Disability Level and 6-Month Symptom Transition Question

				Mild				
	Better (n=47)			Same (n=231)	Worse (n=95)			
Subscale	Mean ± SD	MSRM	95% CI	Mean ± SD	Mean ± SD	MSRM	95% CI	*P*
Mobility	0.72±4.60	0.15	−1.16 to 1.47	1.44±4.66	−0.10±4.23	−0.02	−0.87 to 0.83	0.04
Hand function	2.45±5.65	0.46	−1.15 to 2.08	3.20±5.30	−0.14±4.67	−0.03	−0.97 to 0.91	0.00
Vision	1.60±6.23	0.24	−1.54 to 2.02	2.62±6.69	1.22±5.71	0.18	−0.97 to 1.33	0.15
Fatigue	−0.50±6.25	−0.09	−1.88 to 1.69	−0.26±5.29	−1.17±5.40	−0.22	−1.31 to 0.87	0.42
Cognition	−0.84±5.76	−0.13	−1.78 to 1.52	0.76±6.54	−1.00±5.99	−0.15	−1.36 to 1.05	0.10
Bowel and bladder	0.50±6.22	0.09	−1.69 to 1.87	2.14±5.53	1.23±5.31	0.22	−0.84 to 1.29	0.22
Sensory	1.81±6.31	0.31	−1.49 to 2.12	−0.08±5.78	−1.04±4.66	−0.18	−1.12 to 0.76	0.02
Spasticity	2.39±6.62	0.39	−1.51 to 2.28	2.31±6.17	0.68±5.21	0.11	−0.94 to 1.16	0.07
Pain	2.80±5.74	0.47	−1.18 to 2.11	2.41±6.01	1.37±6.02	0.23	−0.98 to 1.44	0.26
Vasomotor	0.15±7.44	0.02	−2.10 to 2.15	1.42±6.17	−0.67±5.59	−0.11	−1.23 to 1.02	0.01
				Moderate				
	Better (n=21)			Same (n=115)	Worse (n=158)			
Mobility	−0.40±5.42	−0.09	−2.41 to 2.23	−0.18±4.38	−0.82±4.26	−0.19	−0.85 to 0.48	0.83
Hand function	1.86±4.05	0.40	−1.33 to 2.14	0.97±4.60	−0.84±4.67	−0.18	−0.91 to 0.54	0.00
Vision	2.23±5.44	0.36	−1.97 to 2.68	1.57±6.26	1.38±6.45	0.22	−0.79 to 1.22	0.80
Fatigue	−0.99±6.75	−0.18	−3.06 to 2.71	−1.34±5.63	−1.34±5.39	−0.24	−1.08 to 0.60	0.92
Cognition	1.70±5.66	0.32	−2.10 to 2.74	0.82±5.40	−0.39±6.18	−0.07	−1.04 to 0.89	0.19
Bowel and bladder	0.90±7.09	0.16	−2.87 to 3.20	0.24±5.48	−1.88±5.51	−0.34	−1.20 to 0.52	0.01
Sensory	0.28±6.30	0.05	−2.65 to 2.74	1.49±5.80	0.07±5.77	0.01	−0.89 to 0.91	0.33
Spasticity	−0.03±9.12	0.00	−3.91 to 3.90	0.46±5.49	−1.17±4.90	−0.21	−0.98 to 0.55	0.09
Pain	0.96±7.96	0.15	−3.26 to 3.55	2.83±6.54	−0.58±7.00	−0.09	−1.18 to 1.00	0.00
Vasomotor	1.14±6.12	0.19	−2.42 to 2.81	0.71±5.94	−0.26±6.64	−0.04	−1.08 to 0.99	0.01
				Severe				
	Better (n=21)			Same (n=114)	Worse (n=233)			
Mobility	−1.78±5.83	−0.34	−2.84 to 2.15	−0.65±5.19	−0.10±4.89	−0.02	−0.66 to 0.62	0.23
Hand function	0.04±3.80	0.01	−1.62 to 1.63	−0.55±4.75	−1.80±4.74	−0.38	−1.00 to 0.24	0.02
Vision	1.18±7.67	0.16	−3.12 to 3.44	1.52±7.57	1.98±6.65	0.26	−0.61 to 1.13	0.61
Fatigue	−0.42±5.31	−0.06	−2.33 to 2.21	0.66±6.54	−1.91±5.85	−0.29	−1.06 to 0.48	0.00
Cognition	5.00±6.36	0.69	−2.03 to 3.41	3.75±7.24	2.72±6.33	0.38	−0.46 to 1.21	0.21
Bowel and bladder	−1.65±5.85	−0.29	−2.79 to 2.21	−0.46±5.71	−1.96±5.57	−0.34	−1.07 to 0.39	0.06
Sensory	0.45±5.55	0.07	−2.31 to 2.44	1.64±6.97	1.05±6.01	0.15	−0.64 to 0.94	0.68
Spasticity	−0.71±4.77	−0.12	−2.16 to 1.92	−1.17±5.92	−1.25±6.51	−0.21	−1.07 to 0.64	0.89
Pain	0.87±5.34	0.14	−2.15 to 2.42	0.63±6.28	0.11±7.47	0.02	−0.96 to 1.00	0.75
Vasomotor	−0.70±5.85	−0.12	−2.63 to 2.38	0.14±5.73	0.37±6.81	0.07	−0.83 to 0.96	0.65

NOTE. *P* value of F statistic comparing better and worse groups. When available, disability-specific subscales are used. The formula for MSRM is the mean change in scores divided by the SD of change scores of stable patients.
Abbreviations: CI, confidence interval; MSRM, Modified Standardized Response Mean.


A
Symptom Transition Item by Disability-Specific Subscale - PDDS Mild AND within 1 Week
Change in SI Mobility Mild Subscale T Score
n = 47
β = 0.078
Change in SI Hand Function T Score
n = 47
β = 0.242
Change in SI Vision Mild Subscale T Score
n = 46
β = -0.157
Change in SI Fatigue Mild Subscale T Score
n = 47
β = 0.296*
Change in SI Cognition Mild Subscale T Score
n = 48
β = -0.045
Change in SI Bowel and Bladder T Score
n = 46
β = 0.377**
Change in SI Sensory Mild Subscale T Score
n = 47
β = 0.372**
Change in SI Spasticity T Score
n = 47
β = 0.348**
Change in SI Pain T Score
n = 47
β = 0.235
B
Symptom Transition Item by Disability-Specific Subscale - PDDS Moderate AND within 1 Week
Change in SI Mobility Moderate Subscale T Score
n = 43
β = 0.021
Change in SI Hand Function T Score
n = 43
β = 0.422**
Change in SI Vision Moderate Subscale T Score
n = 43
β = 0.154
Change in SI Fatigue Moderate Subscale T Score
n = 43
β = -0.104
Change in SI Cognition Moderate Subscale T Score
n = 43
β = 0.199
Change in SI Bowel and Bladder T Score
n = 43
β = 0.232
Change in SI Sensory Moderate Subscale T Score
n = 43
β = -0.006
Change in SI Spasticity T Score
n = 43
β = 0.047
Change in SI Pain T Score
n = 43
β = 0.050
-15 -10 -5 0 5 10 15
Much Worse
Worse
A Little Worse
No Change
A Little Better
Better
Much Better

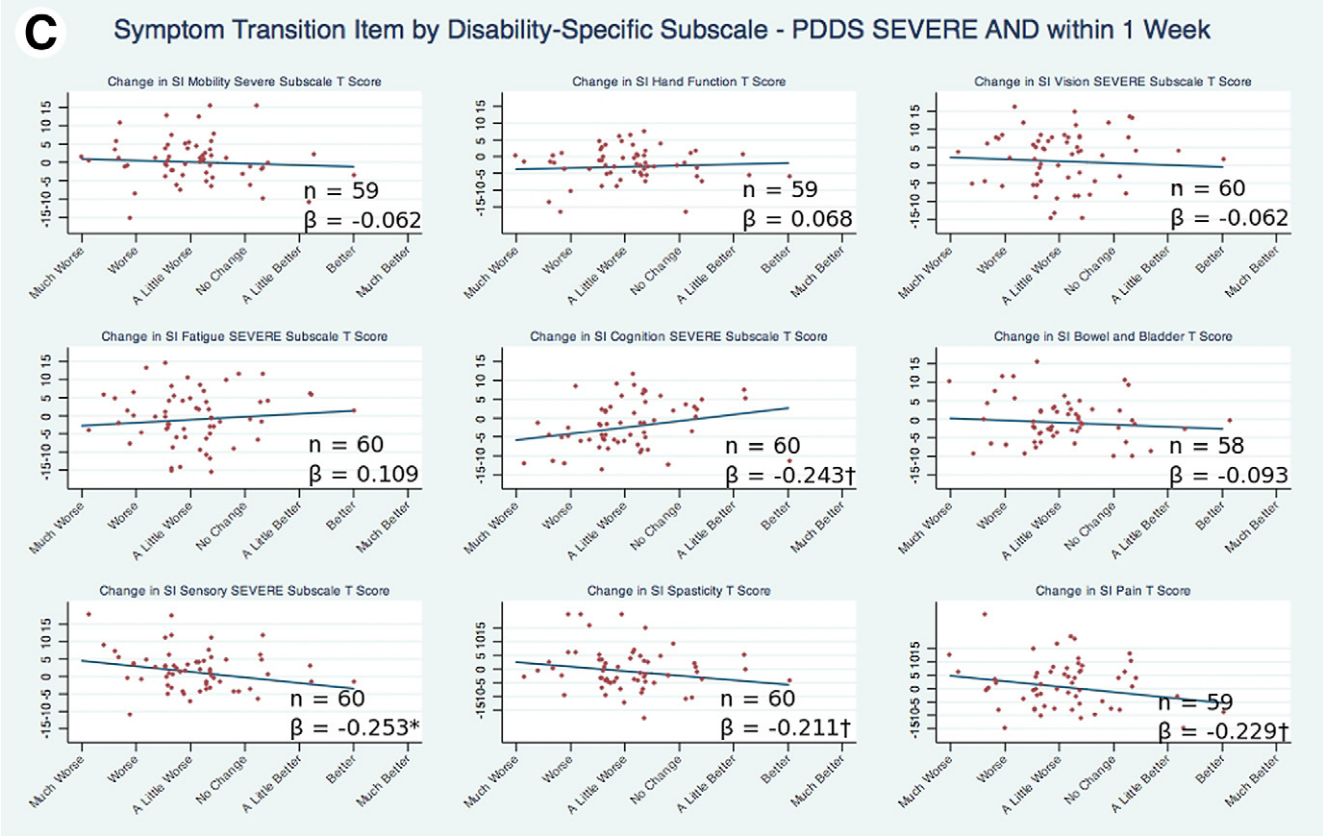


Fig 2. Symptom transition item by disability-specific subscale by disability severity groupings. Jitter plots illustrate the relationship between the symptom transition scores (*x* axis) and change on the Symptom Inventory Short Forms (*y* axis) for the Week-Lag respondents with mild (A), moderate (B), and severe (C) disabilities. Jitter plots allow coincident data points to be visualized by adding a small, random, perturbation to each point so that the discrete nature of the data remains apparent while the point density emerges to view. The jitter plots include linear fit lines, the beta coefficient for which was generated from a linear regression model predicting symptom subscale change score (dependent variable) from symptom transition score (independent variable). **P*<.05; ***P*<.001; †*P*<.10. Abbreviations: PDDS, Patient-Determined Disease Steps; SI, Symptom Inventory.

subscales as computed in this data set was slightly lower, ranging from 0.73 to 0.92, as compared with 0.77 to 0.95 for the new subscales, as reported in our companion article.[21] The magnitude of the associations with the other patient-reported outcomes was similar, with the highest correlations among the disease-specific quality-of-life measures (Patient-Determined Disease Steps: 0.59 and 0.62; Performance Scales Sum: 0.84 and 0.80, for the new and original Symptom Inventory Summary score, respectively) and lower but still significant correlations with the generic quality-of-life measures (SF-12 Physical Component score: −.68 and −.69; SF-12 Mental Component score: −.33 and −.28, for the new and original Symptom Inventory Summary score, respectively; spring scores given as examples in both cases). The correlations were lowest with the Godin Leisure Time Exercise Questionnaire (r= −.21 and −.22, for the new and original Symptom Inventory Summary score, respectively) supporting the divergent validity of the Symptom Inventory Summary score.

Longitudinal Construct Validity

Responsiveness. Table 2 shows the modified standardized response means for the patients reporting better and worse symptom change for the whole sample in the past 6 months. The Symptom Inventory produced the largest levels of responsiveness for hand function and pain among patients with mild disability who improved, for hand function among patients with moderate disability who improved, and for cognition among patients with severe disability who improved. Table 2 also summarizes the results of the 3 linear models comparing better versus worse groups at mild, moderate, and severe levels of disability. Better, same, and worse were defined on the basis of response to the transition item among those respondents who endorsed symptom change in the past 6 months, as compared with those expressing no change (ie, better vs same, worse vs same), for groups with mild, moderate, and severe disabilities. The F tests comparing better versus worse suggested that the Symptom Inventory subscales showing short-term responsiveness in the group with mild disability were mobility, hand function, sensory, and pain. In the group with moderate disability, the responsive subscales were hand function, bowel/bladder, pain, and vasomotor. In the group with severe disability, the responsive subscales were hand function and fatigue. A similar comparison among the Week-Lag subsample was not viable given the very small sample sizes within subgroups.

To better understand the relationship between the symptom transition scores and change on the Symptom Inventory Short Forms, we examined jitter plots of the symptom transition score (x axis) by change scores on the disability-specific Symptom Inventory Short Forms subscales (y axis) for the Week-Lag respondents with mild (fig 2A), moderate (fig 2B), and severe (fig 2C) disabilities. Jitter plots allow coincident data points to

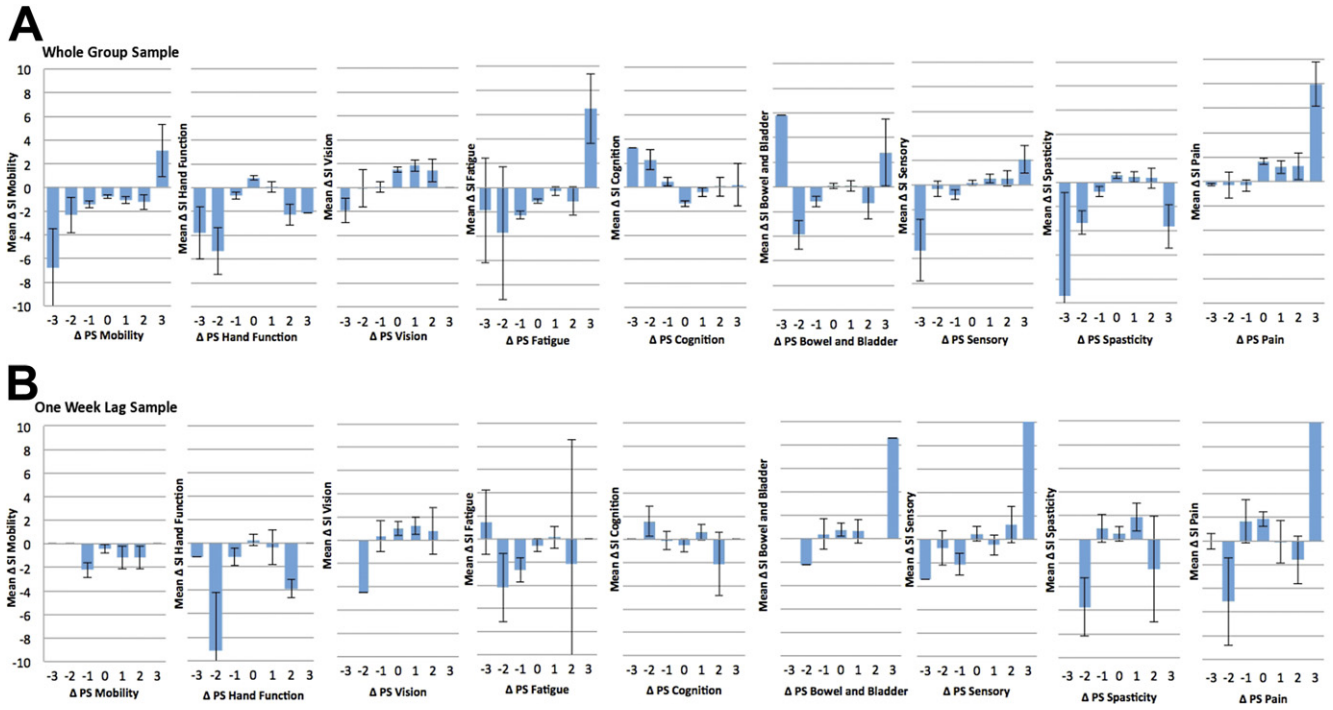


Fig 3. Change in PS by point and corresponding change in SI subscales. The mean change and standard-error-of-the-mean bars are shown for each SI Short Form subscale by change on the corresponding PS item within the whole sample (A) (n=992) and the Week-Lag subsample (B) (n=150). The correspondence between change scores on the PS and SI Disability-Specific Short Forms appears to be relatively linear for the mobility, fatigue, sensory, and pain subscales; it appears to be nonlinear for the hand function, vision, cognition, bowel/bladder, and spasticity subscales. In the Week-Lag subsample, the range of change scores is relatively truncated, and all the relationships appear to be nonlinear. Note that in cases of very small sample (ie, n=1), error bars are not shown. Abbreviations: PS, Performance Scales; SI, Symptom Inventory.

be visualized by adding a small, random, perturbation to each point so that the discrete nature of the data remains apparent while the point density emerges to view. The jitter plots include linear fit lines, the beta coefficient for which was generated from a linear regression model predicting symptom subscale change score (dependent variable) from symptom transition score (independent variable). One would expect that the linear fit lines would have a positive slope, indicating that respondents who endorse that their symptoms have improved (eg, symptom transition score≥5) would also show a positive difference in time 2 − time 1 scores (eg, *y* axis ≥0). These plots demonstrate that the responsiveness of the Symptom Inventory Short Forms is best for the subgroup with mild disability (4 of 9 linear fit lines have statistically significant positive slopes) and weak for the subgroup with moderate disability (1 of 9 linear fit lines has a statistically significant positive slope) and for the subgroup with severe disability (1 of 9 linear fit lines is statistically significant and negative). These figures suggest that the Symptom Inventory Disability-Specific Short Forms are somewhat responsive for mild disability but that there may be a ceiling effect for people with moderate or severe disability.

Symptom Inventory change and Performance Scales change. Since the Performance Scales is intended to be used as a screener for the Symptom Inventory Disability-Specific Short Forms, we also evaluated the responsiveness for the Symptom Inventory Summary score and Performance Scales Summary score and domain-specific items. The Performance Scales Summary score and domain-specific items showed substantially less responsiveness in comparison to the Symptom Inventory Summary score. The Symptom Inventory Summary score showed responsiveness in groups with both mild and moderate disability, whereas the Performance Scales Summary score was responsive to change only in the group with severe disability. Among the Performance Scales domain-specific items, only the "sensory" item was responsive in the group with mild disability, none of the items was responsive in the group with moderate disability, and the "vision, fatigue, and sensory" items showed responsiveness among the group with severe disability (data not shown; see supplemental table 1, available online only at the Archives website: www.archives-pmr.org).

To better understand this variable responsiveness, figure 3 shows the mean change (with standard-error-of-the-mean bars) for each Symptom Inventory Short Forms subscale by change on the corresponding Performance Scales item within the whole sample (top row) and the Week-Lag subsample (bottom row). These plots suggest that the correspondence between change scores on the Performance Scales and Symptom Inventory Disability-Specific Short Forms appears to be relatively linear for the mobility, fatigue, sensory, and pain subscales; it appears to be nonlinear for the hand function, vision, cognition, bowel/bladder, and spasticity subscales. In the Week-Lag subsample, the range of change scores is relatively truncated and all the relationships appear to be nonlinear.

Because the Symptom Inventory is a measure of symptoms and thus a causal indicator, we examined a similar scatter plot for the Performance Scales, which should be an effect indicator and should behave more predictably. These scatter plots suggested no relationship between the change in Performance

Table 3: Effect Sizes for Patients Reporting Better and Worse Symptom Transition Scores

Scale or Subscale	Cohen's Effect Size*					
	Better vs Same	95% CI		Worse vs Same	95% CI	
SI Summary	1.25	1.01	1.50	−0.57	−0.91	−0.24
SI Mobility	1.05	0.80	1.29	−0.57	−0.90	−0.23
Mobility mild	1.12	0.87	1.38	−0.49	−0.83	−0.14
Mobility moderate	1.01	0.76	1.25	−0.54	−0.87	−0.21
Mobility severe	0.88	0.64	1.12	−0.64	−0.98	−0.31
SI Hand function	0.77	0.53	1.00	−0.42	−0.75	−0.09
SI Vision	0.61	0.37	0.84	−0.18	−0.51	0.15
Vision mild	0.56	0.33	0.80	−0.19	−0.53	0.14
Vision moderate	0.45	0.22	0.69	−0.17	−0.51	0.16
Vision severe	0.47	0.23	0.70	−0.04	−0.37	0.30
SI Fatigue	1.27	1.02	1.51	−0.68	−1.02	−0.35
Fatigue mild	1.15	0.91	1.40	−0.67	−1.00	−0.33
Fatigue moderate	1.27	1.02	1.52	−0.86	−1.20	−0.53
Fatigue severe	0.97	0.73	1.21	−0.55	−0.89	−0.22
SI Cognition	0.42	0.19	0.65	−0.24	−0.57	0.09
Cognition mild	0.29	0.06	0.52	−0.19	−0.52	0.14
Cognition moderate	0.38	0.15	0.61	−0.13	−0.46	0.20
Cognition severe	0.37	0.14	0.61	−0.12	−0.45	0.21
SI Bowel and bladder	0.88	0.64	1.12	−0.55	−0.88	−0.22
SI Sensory	1.12	0.87	1.36	−0.44	−0.77	−0.11
Sensory mild	0.92	0.68	1.15	−0.39	−0.72	−0.06
Sensory moderate	1.08	0.83	1.32	−0.51	−0.84	−0.17
Sensory severe	0.63	0.39	0.86	−0.39	−0.73	−0.06
SI Spasticity	0.94	0.70	1.18	−0.32	−0.66	0.01
SI Pain	0.70	0.46	0.93	−0.22	−0.55	0.11
SI Vasomotor	0.74	0.50	0.97	−0.42	−0.75	−0.09

NOTE. Effect size calculation: (better or worse mean − same mean)/(better or worse SD).
Abbreviations: CI, confidence interval; SI, Symptom Inventory.

Scales summary score and the symptom transition score (data not shown). These plots suggest that the symptom transition scores do not covary in a meaningful way with this disease-specific effect indicator measure.

Interpretation

Effect size estimates for sample size planning in future studies. Table 3 presents Cohen's effect size statistic[30] separately for patients who reported improved versus worsened symptoms in the past 6 months. This table suggests that patients reporting symptom improvement generally have large effect sizes (mean=.82), whereas patients reporting worsening symptoms generally have moderate effect sizes (mean=−.40). The Symptom Inventory Summary score and overall (not disability-specific) subscales evidenced the largest effect sizes among patients reporting symptom improvement than did patients reporting symptom worsening. The effect sizes for the disability-specific subscales were generally smaller than those for the overall subscales. Using Cohen's[30] sample size estimations, the above information would suggest that for intervention studies anticipating improvement in overall symptom burden, a sample of 26 per group after attrition would be sufficient to detect change as compared with a stable patient group. For intervention studies anticipating prevention of worsening of overall symptom burden, a sample of 64 per group after attrition would be sufficient to detect change as compared with a stable patient group. If an intervention is focused on a particular symptom domain, the requisite sample size will vary, depending on the domain and the level of disability. Interventions focusing on cognitive symptoms will need to be powered for small effect sizes (ie, 393 per group after attrition), whereas interventions focused on sensory symptom reduction would need to be powered for moderate effect sizes in groups with severe disability (ie, 64 per group after attrition) as compared with large effect sizes in groups with mild and moderate disability (ie, 26 per group after attrition). These estimates are provided primarily to illustrate the possible utility of table 3.

Crosswalks. A graphic presentation of the *crosswalk* among the Symptom Inventory Summary score and the disease-specific and generic quality-of-life measures is shown in figures 4A and 4B, respectively. The *disease-specific crosswalks* suggest that the correspondence between the Symptom Inventory Summary score and the Performance Scales Sum covers the whole range of scores and is evenly distributed (fig 4A). The linkage with the Patient-Determined Disease Steps suggests that a relatively broad range of symptom burden is captured by some levels of disability (eg, scores of 0 and 6) but a very narrow range of symptom burden is captured by the other scores on the Patient-Determined Disease Steps (fig 4A).

The *generic crosswalk* (fig 4B) denoting the linkage with the SF-12 physical and mental health functioning scores suggests that the correspondence between SF-12 and Symptom Inventory Summary score is different at low and high ends of functioning than in the middle. This suggests that the Symptom Inventory Summary scores have a more equal interval represented by a given 5-point change on the SF-12. Of note, the worst levels of symptoms are associated with higher SF-12 mental health scores than with SF-12 physical health scores (ie, 13 vs 5, respectively).

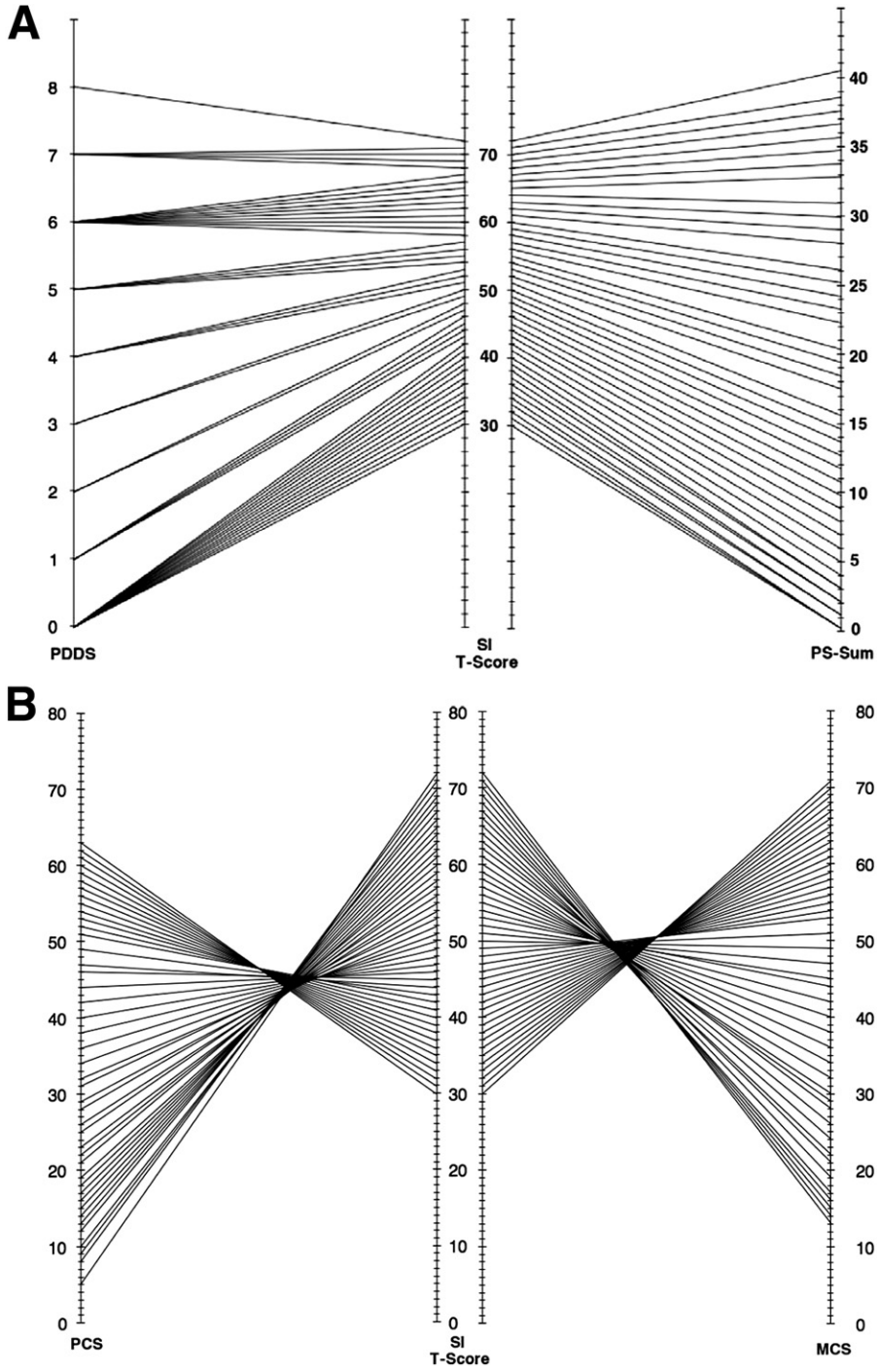


Fig 4. Disease-specific and generic crosswalk nomograms. Graphic presentations of the correspondence between scores are shown in crosswalks among the SI Summary score and the disease-specific (A) and generic (B) quality-of-life measures. Crosswalk tables for all the disease-specific SI Short Forms subscales are available at www.deltaquest.org/tools/Symptom_Inventory_Crosswalk. Abbreviations: MCS, Mental Component score; PCS, Physical Component score; PDDS, Patient-Determined Disease Steps; PS, Performance Scales; SI, Symptom Inventory.

DISCUSSION

The Symptom Inventory evidenced convergent and divergent validity in the present study, and the new item response theory–derived disability-specific short forms evidenced similar psychometric performance as the original 29-item short form in terms of intercorrelations with other patient-reported outcomes but with slightly better alpha reliability. The Symptom Inventory also evidenced moderate responsiveness to clinically important change, with more responsiveness on specific subscales among the groups with mild and moderate disability

than among the group with severe disability. Effect sizes were generally larger among patients who reported improvement in their symptoms, rather than deterioration, suggesting that the tool would be more useful in clinical research focused on testing whether an intervention improves symptom experience (as compared with preventing deterioration), and particularly for patients with mild and moderate disability.

Study Limitations

Our results also suggest that symptom measures can be useful for elucidating the patient's experience, but vary considerably across and within disability groupings, making the psychometric evidence of validity and responsiveness less straightforward than among functional status measures. Because symptom measures are causal indicators, there are caveats in applying standard approaches to assessing psychometric performance and responsiveness. For example, Fayers and colleagues[10,17-19] caution against the use of correlational and other linear approaches because symptoms *cause* changes in quality of life but do not indicate changes in quality of life. Consequently, there is an inherent asymmetry in the relationships between the causal indicators (self-reported symptoms) and the latent construct (quality of life): that is, having a particular symptom will cause worse quality of life (a sufficient cause) but not having it will not necessarily result in better quality of life (a necessary cause). For example, if someone's symptoms got worse in 1 area, which is a small subset of the whole, then the Symptom Inventory Summary score or subscale score might not correlate highly with an omnibus score (Performance Scales) or with most of the other symptom subscales. To have a high correlation with Performance Scales Sum, the symptom burden would have to be consistent in all domains related to disability. A clinimetric scale, such as a symptom measure, selects items because they are thought to be related to an underlying concept that defies explicit measurement.[17] This makes testing the validity of such a tool more complex, and from some perspectives inappropriate.[17]

Another limitation of the present work relates to the use of transition ratings. There is a growing body of research that suggests that transition questions can be affected by patients' ability to recall prior health states and unduly affected by their health: if they are feeling well, they rate themselves as improved; if unwell, as deteriorated.[34] Guyatt et al[34] recommend using transition ratings for investigating a questionnaire's responsiveness and noted that such ratings are more useful if they are grounded on a significant event for the patient, such as a clinic visit. While our work followed these recommendations using transition ratings to aid in assessing responsiveness, there was no salient grounding event and there was substantial variability in the data collection window for the quality-of-life measures as compared with the Symptom Inventory. This variability is a caveat of the present work.

CONCLUSIONS

Our goal in this work was to create a series of short symptom measures that would be clinically useful for MS clinical research and practice. The end result of the numerous analyses is not a definitive statement that these are reliable and valid psychometric tools but rather that these clinimetric measures are complex, differentially responsive, and differentially sensitive to clinically important change. Furthermore, large samples would be needed for future research using the disability-specific subscales because of the small effect sizes documented in the present work among patients who reported improving versus deteriorating. The tool also appears to be more sensitive to improvements in patients with mild and moderate diabilities than to deterioration in patients with severe disabilities.

Indeed, we emerge from this series of analyses with more questions about what drives the disconnection between symptom change and quality-of-life outcomes. Future research might address such questions by utilizing the Symptom Inventory in combination with qualitative data that explore the patients' cognitive processes underlying their answers. In addition, other statistical modeling approaches, such as structural equation modeling and data mining techniques, could be useful in addressing such research questions, particularly if the distinction between causal and effect indicators can be modeled. Future research might also focus on empirically linking symptom experience and the concepts of prognostic factors, treatment effect modifiers/mediators, and treatment mediators proposed by Hill and Fritz.[35] For example, reflective[36] as compared with formative[37] structural equation models can be used to test this theoretical distinction, particularly if longitudinal data with more than 2 time points are available.[38] Formative structural equation models, such as Multiple Indicator Multiple Cause structural equation models, would be a way to evaluate whether symptoms also qualify as prognostic factors, treatment effect modifiers/mediators, and treatment mediators.

Acknowledgments: We thank Gary Cutter, PhD, and Stacey Cofield, PhD, for data management services early in the project; Tamara Fong, MD, PhD, for helpful suggestions in graphic displays of the crosswalks; and Ruth Ann Marrie, MD, PhD, and Robert Fox, MD, for helpful comments on an earlier draft of this manuscript.

References

1. Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Med Care 2000; 38:II84-90.
2. Hays RD, Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. Qual Life Res 1992;1:73-5.
3. Beaton DE. Understanding the relevance of measured change through studies of responsiveness. Spine 2000;25:3192-9.
4. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003;12:349-62.
5. Revicki DA, Hays RD, Cella D, Sloan J. Recommended methods for determining responsivenes and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61: 102-9.
6. Sprangers MAG, Moinpour CM, Moynihan TJ, Patrick DL, Revicki DA; Clinical Significance Consensus Meeting Group. Assessing meaningful change in quality of life over time: a user's guide for clinicians. Mayo Clin Proc 2002;77:561-71.
7. Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T; Clinical Significance Consensus Meeting Group. Estimating clinically significant differences in quality of life outcomes. Qual Life Res 2005;14:285-95.
8. Embretson S, Reise SP. Item response theory for psychologists. Mahwah: Erlbaum; 2000.
9. U.S. Department of Health and Human Services, F.A.D.A., Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH). Guidance for industry: patient-reported outcome measures: use in medical product development to support

labeling claims. Washington (DC): U.S. Department of Health and Human Services; 2009.
10. Fayers PM, Hand DJ. Factor analysis, causal indicators, and quality of life. Qual Life Res 1997;6:139-50.
11. Tremlett H, Zhao Y, Rieckmann P, Hutchinson M. New perspectives in the natural history of multiple sclerosis. Neurology 2010; 74:2004-15.
12. Compston A, McDonald IR, Noseworthy J, et al, editors. McAlpine's multiple sclerosis. 4th ed. New York: Churchill Livingstone; 2005.
13. Schwartz CE, Vollmer T, Lee H. North American Research Consortium on Multiple Sclerosis Outcomes Study Group. Reliability and validity of two self-report measures of impairment and disability for MS. Neurology 1999;5263-70.
14. Rammohan KW, Shoemaker J. Emerging multiple sclerosis oral therapies. Neurology 2010;74S47-53.
15. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah: Erlbaum; 2000.
16. Schwartz CE, Welch G, Santiago-Kelley P, Bode R, Sun X. Computerized adaptive testing of diabetes impact: a feasibility study of Hispanics and non-Hispanics in an active clinic population. Qual Life Res 2006;15:1503-18.
17. Fayers PM. Quality-of-life measurement in clinical trials—the impact of causal variables. J Biopharm Stat 2004;14:155-76.
18. Fayers PM, Hand DJ. Causal variables, indicator variables and measurement scales: an example from quality of life. J Royal Stat Soc 2002;165:233-61.
19. Fayers PM, Hand DJ, Bjordal K, Groenvold M. Causal indicators in quality of life research. Qual Life Res 1997;6:393-406.
20. Schwartz CE, Merriman MP, Reed G, Byock I. Evaluation of the Missoula-VITAS Quality of Life Index–Revised: research tool or clinical tool? J Palliat Med 2005;8:121-35.
21. Schwartz CE, Bode RK, Vollmer T. The Symptom Inventory disability-specific short forms for multiple sclerosis: reliability and factor structure. Arch Phys Med Rehabil. 2012. In press.
22. Godin G, Shephard RJ. A simple method to assess exercise behavior in the community. Can J Appl Sport Sci 1985;10:141-6.
23. Hohol MJ, Orav EJ, Weiner HL. Disease steps in multiple sclerosis: a simple approach to evaluate disease progression. Neurology 1995;45:251-5.
24. Marrie RA, Goldman M. Validity of performance scales for disability assessment in multiple sclerosis. Mult Scler 2007; 13:1176-82.
25. Ware JE Jr, Kosinski M, Keller SD. A 12-item Short-Form Health Survey. Med Care 1996;34:220-33.
26. Motl RW, Schwartz CE, Vollmer T. Continued validation of the Symptom Inventory in multiple sclerosis. J Neurol Sci 2009;285: 134-6.
27. Haywood KL, Garratt AM, Jordan K, Dziedzic K, Dawes PT. Disease-specific, patient-assessed measures of health outcome in ankylosing spondylitis: reliability, validity, and responsiveness. Rheumatology (Oxford) 2002;41:1295-302.
28. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. Econometrica 1979;47:263-91.
29. Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002;11:207-21.
30. Cohen J. A power primer. Psychol Bull 1992;112:155-9.
31. Fong TG, Fearing MA, Jones RN, et al. Telephone interview for cognitive status: creating a crosswalk with the Mini-Mental State Examination. Alzheimers Dement 2009;5:492-7.
32. Wu AW, Huang IC, Gifford AL, Spritzer KL, Bozzette SA, Hays RD. Creating a crosswalk to estimate AIDS Clinical Trials Group quality of life scores in a nationally representative sample of persons in care for HIV in the United States. HIV Clin Trials 2005;6:147-57.
33. Miller RGJ. Simultaneous statistical inference. New York: Springer-Verlag; 1991.
34. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002;55:900-8.
35. Hill JC, Fritz JM. Psychosocial influences on low back pain, disability, and response to treatment. Phys Ther 2011;91:712-21.
36. Oort FJ. Using structural equation modeling to detect response shifts and true change. Qual Life Res 2005;14:587-98.
37. Donaldson GW. Structural equation models for quality of life response shifts: promises and pitfalls. Qual Life Res 2005;14: 2345-51.
38. Jones RN, Manly J, Glymour MM, Rentz DM, Jefferson AL, Stern Y. Conceptual and measurement challenges in research on cognitive reserve. J Int Neuropsychol Soc 2011;17:593-601.

Supplier

a. Survey-Gizmo, 4888 Pearl East Cir, Ste 300W, Boulder, CO 80301.

Supplemental Table 1: Analysis Showing Responsiveness of the Symptom Inventory Summary Score, Performance Scale Summary Score, and Performance Scale Items by Disability Groupings

	Mild			
	Better (n=47)	Worse (n=95)		
Scale or Subscale	Mean ± SD	Mean ± SD	F	*P*
SI Mean T difference	−1.0±3.6	−0.2±2.7	9.5	0.000
PS Summary	−0.1±3.5	−0.2±2.7	0.3	0.744
Mobility	−0.1±0.7	0.0±0.7	0.0	0.956
Hand function	−0.1±0.7	−0.1±0.6	0.1	0.895
Vision	0.1±0.5	−0.1±0.6	2.3	0.097
Fatigue	−0.2±1.0	−0.1±0.7	0.3	0.716
Cognition	0.0±0.6	0.0±0.8	1.0	0.375
Bowel and bladder	0.1±0.8	0.0±0.6	0.9	0.427
Sensory	−0.2±0.9	0.1±0.8	3.0	0.049
Spasticity	0.1±0.6	−0.1±0.7	1.4	0.243
Pain	0.2±0.9	0.0±0.8	2.1	0.120
	Moderate			
	Better (n=21)	Worse (n=158)		
SI Mean T difference	−0.68±3.53	0.83±3.07	6.29	0.002
PS Summary	−0.11±3.29	−0.30±3.02	0.19	0.828
Mobility	−0.04±0.69	−0.05±0.68	0.22	0.800
Hand function	−0.04±0.64	−0.07±0.76	0.04	0.960
Vision	−0.04±0.64	0.03±0.71	0.50	0.609
Fatigue	−0.07±1.12	−0.13±0.82	0.07	0.931
Cognition	−0.18±0.82	−0.02±0.71	0.94	0.392
Bowel and bladder	−0.07±0.60	−0.03±0.79	0.13	0.875
Sensory	0.00±0.94	−0.02±0.94	0.63	0.534
Spasticity	0.07±0.86	0.03±0.77	0.29	0.750
Pain	0.25±0.84	−0.04±0.92	1.21	0.299
	Severe			
	Better (n=21)	Same (n=224)		
SI Mean T difference	0.47±2.90	0.79±2.78	0.84	0.434
PS Summary	0.39±2.79	−0.38±3.85	3.95	0.020
Mobility	0.00±0.38	0.07±0.57	0.14	0.871
Hand function	−0.07±0.81	−0.08±0.84	2.06	0.129
Vision	0.11±0.57	−0.09±0.84	3.39	0.034
Fatigue	0.29±0.85	−0.14±0.87	4.50	0.012
Cognition	0.25±0.75	−0.06±0.85	2.02	0.134
Bowel and bladder	0.14±0.52	−0.02±0.84	0.97	0.379
Sensory	−0.36±0.78	0.01±1.05	3.67	0.026
Spasticity	0.04±0.79	0.00±0.95	1.06	0.347
Pain	0.00±1.09	−0.08±1.02	0.54	0.585

Abbreviations: PS, Performance Scales; SI, Symptom Inventory.