Construct validity and item response theory analysis of the promis-29 v2.0 in recipients of lumbar spine surgery

Cook, Chad E.; Rethorn, Zachary D.; Chiarotto, Alessandro; Garcia, Alessandra N.; Gottfried, Oren

published in
SPINE
2021

DOI (link to publisher)
10.1097/BRS.0000000000004085

document version
Publisher's PDF, also known as Version of record

Spine

SPINE Volume 46, Number 24, pp 1721–1728

HEALTH SERVICES RESEARCH

Construct Validity and Item Response Theory Analysis of the PROMIS-29 v2.0 in Recipients of Lumbar Spine Surgery

Chad E. Cook, PhD, PT, FAPTA,[a,b,c] Zachary D. Rethorn, DPT,[a] Alessandro Chiarotto, PhD, PT,[d,e] Alessandra N. Garcia, PhD, PT,[f] and Oren Gottfried, MD[g]

Study Design. Observational cohort design involving measurement property assessment.

Objectives. The aim of this study was to assess construct validity through hypothesis testing and to examine reliability and discrimination of the PROMIS-29 v2.0 using item response theory (IRT) analyses.

Summary of Background Data. Patient-Reported Outcomes Measurement Information System (PROMIS) 29.0 version 2.0 includes 28 questions for the seven domains of Physical function, Anxiety, Depression, Fatigue, Sleep disturbance, Social role, and Pain interference, and one item related to pain intensity. To date, the tool has not been tested for construct validity for selected concepts, nor has it been evaluated using IRT in a population of spine surgery recipients.

Methods. Using the Quality Outcomes Dataset lumbar registry, we evaluated the construct validity of the PROMIS-29 v2.0 against pain intensity measures for back and leg, the Oswestry Disability Index, the EQ5D 3L-visual analog scale (quality of life) and proxy measures of activities, mobility and self-care. IRT was used to test PROMIS-29 v2.0 assumptions and fit.

Results. The sample included 652 surgery recipients (mean age = 60.1, SD = 14.0) who had a high level of baseline disability. Hypothesis testing confirmed direction and magnitude of correlation between the PROMIS and legacy measures in 10 of the 12 hypotheses. IRT identified three misfit items, but otherwise adequate scale reliability and unidimensionality.

Conclusion. The PROMIS-29 v2.0 measures several different constructs pertinent to a patient's health and recovery during spine surgery. We feel that the PROMIS-29 v2.0 tool is a useful and effective outcome measure for populations who receive spine surgery.

Key words: item response theory, low back pain, measurement properties, PROMIS, spine surgery, validity.

Level of Evidence: 4

Spine 2021;46:1721–1728

Spine surgical candidates exhibit outcomes that are influenced by a litany of factors, including biological, psychological, and social components.[1,2] Surgical recipients tend to have higher levels of comorbidities, which increase their susceptibility to complications and fluctuations in overall health status.[3] Capturing the outcomes associated with surgery requires the judicious use of patient-reported outcome measures that exhibit psychomet-rically sound measurement properties. Patient-reported measures have become increasingly important with increased scrutiny of quality of care.[4]

Patient-Reported Outcomes Measurement Information System (PROMIS) measures were developed with support from the National Institutes of Health (NIH).[5] The PROMIS instruments are person-centered and designed to evaluate and monitor physical, mental, and social health in adults and children, with and without chronic conditions. An NIH initiative to facilitate research standards in patients with chronic low back pain endorsed, among other items and measures, the 29-item PROMIS profile.[6] The PROMIS-29 has included versions 1.0, 1.1., 2.0 and the most recent 2.1, each with minor changes from the previous iterations.

The PROMIS-29 v2.0 consists of 28 questions for the seven domains of Physical function, Anxiety, Depression, Fatigue, Sleep disturbance, Social role, and Pain interference;

From the [a]Department of Orthopedics, Duke University, Durham, NC; [b]Department of Population Health Sciences, Duke University, Durham, NC; [c]Duke Clinical Research Institute, Duke University, Durham, NC; [d]Department of Health Sciences, Faculty of Science, Vrije Universiteit Amsterdam, Amsterdam Public Health research institute, The Netherlands; [e]Department of General Practice, Erasmus MC, University Medical Center, Rotterdam, The Netherlands; [f]College of Pharmacy & Health Sciences, Doctor of Physical Therapy Program, Campbell University, Buies Creek, NC; and [g]Department of Neurosurgery, Duke University, Durham, NC.

Acknowledgment date: January 27, 2021. First revision date: February 22, 2021. Acceptance date: March 17, 2021.

The manuscript submitted does not contain information about medical device(s)/drug(s).

No funds were received in support of this work.

No relevant financial activities outside the submitted work.

Address correspondence and reprint requests to Chad E. Cook, PhD, PT, 317 Trent Drive, Duke University, Durham, NC, 27710;
E-mail: chad.cook@duke.edu

DOI: 10.1097/BRS.0000000000004085

and a single question related to pain intensity. At present, the physical function, pain interference, and depression PROMIS-29 measures have been concurrently validated (construct validity) against legacy outcomes measures such as the Oswestry Disability Index (ODI),[7] the numeric pain rating scale (NRS) for pain intensity, and the EuroQol five-dimensional (EQ5D3L) questionnaire in patients who received surgery of the lumbar spine.[8–10] To our knowledge, no studies have evaluated the PROMIS-29 tool in a spine surgery population using item response theory (IRT), nor have any studies concurrently evaluated the associations of the PROMIS-29 against the measures of self-care, activities of daily living (ADL), or mobility, which are standard measures in the Quality Outcome Database (QOD) lumbar registry.[11,12] IRT analyses are of particular importance, since they provide detailed information on item- and scale-level characteristics alongside the continuum of the latent construct measured by a specific measure. For example, the reliability of a patient-reported measure for different levels of the construct measured (*e.g.*, low-level *vs.* high level of physical functioning) cannot be evaluated with classical test theory.[13] Moreover, PROMIS measures were created using IRT in a general population,[14] but it remains unclear whether the item parameters behave similarly in different populations (such as those receiving spine surgeries). The measures of self-care, ADL, and mobility are standard questions across the 100 sites that use the registry data we evaluated, which is actually more sites than those that use the PROMIS-29 tool.

The objectives of this study were to assess construct validity and examine item-level characteristics and reliability of the PROMIS-29 v2.0 using IRT analyses in individuals who underwent lumbar spine surgery. It is intuitive to assume that different domains should exhibit different levels of correlation with comparative measures. Consequently, 12 independent hypotheses involving the PROMIS-29 v2.0 domains and their relationships with legacy and proxy measures (Table 1) were created. The study further hypothesized that all items and domains would have adequate psychometric properties following IRT analyses.

METHODS

Design and Data Resource

This observational study accessed the aggregate QOD lumbar registry for cases performed from 2016 to 2018. The QOD is a prospective observational registry that records patient baseline demographics, clinical, surgical, and patient-reported outcome measures data in both short and long-term follow-up from >100 active sites in North America.[11,12]

Participants

The dataset included individuals aged 18 years who underwent lumbar spine surgery performed for either primary or recurrent (revision) lumbar degenerative disorders. Lumbar spine surgery procedures included all codes for lumbar fusion, decompression, and discectomy, including situations involving multiple coding. Individuals diagnosed with severe spine disorders *(e.g.,* infection, tumor, fracture, and neurologic paralysis) were excluded from QOD registry.[12] Ethical approval for this study was obtained and exempt from the Institutional Review Board of Duke University.

Patient-reported Outcome Measurement Information System (PROMIS) 29 v2.0.

As stated, the PROMIS-29 v2.0 assesses seven domains, Physical function, Anxiety, Depression, Fatigue, Sleep

TABLE 1. Hypotheses Formulated A Priori to Assess Construct Validity of the PROMIS-29 v2.0 in Individuals Who Underwent Lumbar Spine Surgery

Hypothesis for construct validity
1. The correlation between the PROMIS physical function and the ODI is at least 0.2 higher than the correlation of the PROMIS physical function and the VAS LBP
2. The correlation between the PROMIS physical function and the ODI is at least 0.2 higher than the correlation of the PROMIS physical function and the EQ5D VAS
3. The correlation between the PROMIS fatigue and the ODI is <0.5
4. The correlation between the PROMIS fatigue and the VASLBP is <0.5
5. The correlation between the PROMIS sleep and the ODI is <0.5
6. The correlation between the PROMIS sleep and the VASLBP is <0.5
7. The correlation between the PROMIS social and the ODI is higher than the correlation of the PROMIS social with the VAS LBP
8. The correlation between the PROMIS social and the ODI is higher than the correlation of the PROMIS social with the EQ5D VAS
9. The correlation between the PROMIS pain interference and the ODI is at least 0.1 higher than the correlation between the PROMIS pain interference and the EQ5D VAS
10. The correlation between the PROMIS pain interference and the VAS LBP is at least 0.1 higher than the correlation between the PROMIS pain interference and the EQ5D VAS
11. The correlation between the PROMIS pain intensity and the VAS LBP is at least 0.3 higher than the correlation between the PROMIS pain intensity and the ODI
12. The correlation between the PROMIS pain intensity and the VAS LBP is at least 0.3 higher than the correlation between the PROMIS pain intensity and the EQ5D VAS

LBP indicates low back pain; NRS, the numeric pain rating scale; ODI, Oswestry Disability Index; VAS, visual analog scale.

disturbance, Social role (ability to participate in social roles and activities), and Pain interference (four questions for each domain), and a single question for pain intensity. PROMIS-29 v2.0 is scored using a T-score metric that requires individual conversion from a raw score to a final value. Raw scores were calculated and negatively oriented questions were inverted as recommended by the HealthMeasures group.[15] Tables that convert raw values to *t* scores are available on the HealthMeasures website and are unique for each domain.[15] For PROMIS measures, higher *t* scores are more reflective of the concept being measured (*e.g.,* sleep, anxiety) in which 50 is the mean of a relevant reference population and 10 is the standard deviation (SD) of that population. Consequently, scores of 40 or 60 are both one SD below and above the mean, which could be desirable or undesirable outcome, depending on how the item is coded.[15]

Comparator Instruments

Comparator instruments included the legacy measures of the NRS for pain intensity, the ODI v2.0, EQ5D3L-visual analog scale (VAS),[16] and three proxy single items related to the ability to perform usual ADLs, self-care, and self-described mobility. This set of patient-reported measures represents core domains in individuals who underwent lumbar spine surgery.[17]

The NRS for pain intensity is 11-item reliable scale used to measure back and leg pain intensity by asking patients to score their pain over the last 7 days from 0 (no pain) to 10 (worst possible pain).[18] The NRS for pain intensity has extensively been investigated in patients with LBP.[19] ODI v2.0 is a 10-item reliable and valid questionnaire used to measure how patients' back or leg pain affects their ability to manage in everyday life).[7] Higher scores on the ODI (>40%) indicate severe disability and lower scores (<20%) indicate minimal disability.[7] The EQ5D3L-VAS is a single item asking people to verbally self-report their health status.[20] The highest score (100) indicates "best imaginable health status," and the lowest score (0) indicates "worst imaginable health status." The three single items in the QOD[11,12] included: ability to perform usual activities using a self-reported single question "Describe your ability to perform your usual activities (*i.e.,* work, study, housework, leisure activities)" (with possible responses "no problems with performing usual activities," some problems with performing usual activities," or "unable to perform usual activities"); ability to self-care was assessed using a self-reported single question "Describe your self-care" (with possible responses "no problems with self-care," "some problems washing or dressing," "unable to wash or dress"); and self-described mobility, which was measured by a self-reported single question "Describe your mobility" (with possible responses "no problems to walk," "some problems to walk," or "confined to bed").

Additional Variables

For a description of our population, mean age in years, mean body mass index (BMI), and mean American Society of Anesthesiologists' (ASA) classification of Physical Health score (I to V, with higher values reflecting a more serious condition) were tabulated.[21] Primary *versus* revision surgeries, sex, and comorbidities (present or absent) such as diabetes, coronary artery disease, smokers, peripheral vascular disease, anxiety, depression, renal disease, chronic pulmonary disease, osteoporosis, and arthritis were also tabulated.

Statistical Analysis

Demographics and Patient Descriptors

All individuals with baseline measures of PROMIS-29 v2.0, as well as legacy and proxy measures from baseline, were included. Summative age, general means, and SDs for self-reported outcomes, and frequencies of descriptors such as sex, ASA scores, comorbidities, and outcomes measures were calculated for the full sample.

Objective One-construct Validity Assessment

Construct (concurrent) validity was assessed by testing *a priori* hypotheses regarding the expected correlation of the PROMIS-29 v2.0 t-score *values* to the legacy and proxy measures of pain, disability, and quality of life variables. The magnitude of expected correlations was formulated based on whether it was expected that two outcomes measures were measuring very similar (*e.g.,* two different single-item pain intensity scales), similar (*e.g.,* two different outcome measures associating physical function), related (*e.g.,* one outcome measuring physical function and one measuring pain interference), or unrelated constructs (*e.g.,* one outcome instrument measuring anxiety and one measuring pain intensity). This hypothesis testing approach is recommended by international initiatives such as COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and the International Society for Quality of Life Research (ISOQOL).[22,23] Continuous measures were correlated using a Pearson product correlation coefficient, whereas continuous to ordinal measures were calculated using a Kendall coefficient of rank, tau-sub-b.[24] First, correlational analyses were evaluated on the full sample, but also in primary surgery only (N = 505) and revision surgery only (N = 147) groups to determine if the hypotheses yielded similar results. Construct validity for each PROMIS domain was deemed satisfactory if at least 75% of the hypotheses were met.[25]

Objective Two-IRT

IRT models attempt to explain the relationship between abilities (unobservable characteristic or attribute) and their manifestations (outcomes, responses, or performance). IRT is a theory of testing that is based on the relationship between individuals' performances on a single test item and the test takers' levels of performance on an overall measure of the ability that item was designed to measure. IRT has the capacity to reduce scale burden by refining test

items into those that adequately reflect the construct it was designed to measure.

In this study, IRT was run on the fall sample of both primary and revision surgery recipients. Three assumptions must be met when IRT analyses are used—unidimensionality, local independence, and monotonicity. We followed the PROMIS IRT analytic plan to assess IRT assumptions and model fit.[26] This included the IRT assumptions of unidimensionality, local independence, and monotonicity. This also included model fit testing using Samejima's Graded Response Model (GRM) for polytomous data. Appendix 1, http://links.lww.com/BRS/B767 outlines the specific analytic plan using IRT.

RESULTS

Study Participant Characteristics

Six hundred fifty-two (652) individuals were included with a mean age was 60.1 (SD = 14.0), the mean BMI was 31.5 (SD = 6.7), and 52.5% were male. Nearly 77% were primary surgery recipients. The mean ASA grade was 2.52 (SD = 0.6) and intensity scores and variances for both the back and leg pain measures were 6.3/10 (SD = 2.6). The mean ODI was 44.8/100 (SD = 17.4) and the EQ5D3L-VAS score was 60.2/100 (SD = 21.0). With respect to comorbidities, 21% were diagnosed with diabetes, 9.5% had coronary artery disease, 17.8% were smokers, 2.8% had peripheral vascular disease, 23.8% were diagnosed with anxiety, 23.9% were diagnosed with depression and 22.5% were diagnosed with arthritis. Six percent were diagnosed with renal disease, 6.1% had chronic pulmonary disease, and 6.6% were diagnosed with osteoporosis. There were no missing values for the PROMISE-29 v2.0, legacy, or proxy measures.

Objective One-construct (Concurrent) Validity

All correlational analyses yielded statistically significant results ($P < 0.01$) (Table 2). The strongest correlations were present between the ODI percentile score and the PROMIS 29.0 v. 2.0 domains of physical function ($r = -0.61$), social role ($r = -0.62$), and pain interference ($r = 0.66$). For the full sample of primary and revision subjects, the *a priori* hypotheses were met in 10 of the 12 cases (>75%). There two instances where the hypotheses were incorrect included: *hypothesis 11*-The correlation between the PROMIS pain intensity and the NRS LBP is at least 0.3 higher than the correlation between the PROMIS pain intensity and the ODI; and *hypothesis 12*-The correlation between the PROMIS pain intensity and the NRS LBP is at least 0.3 higher than the correlation between the PROMIS pain intensity and the EQ5D3L-VAS.

When the same hypotheses were evaluated in primary-only and revision-only datasets, the same results were found with the exception on hypothesis 10 for the revision group: ''the correlation between the PROMIS pain interference and the VAS LBP is at least 0.1 higher than the correlation between the PROMIS pain interference and the EQ5D VAS.'' The *a priori* threshold for construct validity (≥75%) was met in each analysis: 83% of hypotheses were met for the full sample, 83% for the primary-only sample, and 75% for the revision sample.

Objective Two-IRT (Assumptions)

Each subscale of the PROMIS-29 v2.0 demonstrated adequate unidimensionality following confirmatory factor analysis (Table 3). All residual correlations were <0.2, indicating that each item was locally independent. All items demonstrated monotonicity.

Objective Two-IRT (Analyses)

Three items (In the past 7 days, I felt fearful [anxiety], In the past 7 days, my sleep was refreshing [sleep], and I have trouble doing all my usual work [including work at home] [social]) were identified as misfitting (S-χ^2 $P < 0.001$) (Table 4). This suggests a potential lack of content validity to reflect the item domain (*e.g.*, sleep, anxiety or social factors) in this sample. All items demonstrated slopes ≥1 and all item thresholds were ordered. Domain reliability

TABLE 2. Correlation of PROMIS-29 v2.0 T Score Values *Versus* Legacy Measures (N = 652

	ODI Percentage Score Baseline	NRS Pain Leg	NRS Pain in Low Back	EQ5D VAS	Ability to Perform ADLs Usually	Ability to Care for Self During Activities	Self-Described Mobility
PROMIS 29 Physical Function Domain	−0.609*	−0.246*	−0.347*	0.348*	−0.532*	−0.416*	−0.509*
PROMIS 29 Anxiety Domain	0.366*	0.179*	0.240*	−0.340*	0.204*	0.264*	0.205*
PROMIS 29 Depression Domain	0.439*	0.219*	0.284*	−0.339*	0.246*	0.291*	0.262*
PROMIS 29Fatigue Domain	0.452*	0.210*	0.295*	−0.382*	0.233*	0.280*	0.276*
PROMIS 29 Sleep Domain	0.493*	0.341*	0.318*	−0.216*	0.274*	0.245*	0.221*
PROMIS 29 Social Role Domain	−0.624*	−0.269*	−0.322*	0.381*	−.523*	−0.352*	−0.420*
PROMIS 29 Pain Interference Domain	0.663*	414*	0.469*	−0.319*	.511*	0.405*	0.439*
PROMIS 29 Pain Intensity Score	0.525*	0.283*	0.344*	−0.247*	.434*	0.350*	0.324*

ADL indicates activities of daily living; NRS, the numeric pain rating scale; ODI, Oswestry Disability Index; VAS, visual analog scale.
*$P < 0.01$.

TABLE 3. Results of Confirmatory Factor Analysis on PROMIS-29 v2.0

Subscale	Fit Index			
PROMIS domains	**CFI**	**TLI**	**RMSEA**	**SRMR**
Physical function	1.00	1.00	0.00	0.01
Anxiety	1.00	1.00	0.00	0.00
Depression	1.00	1.00	0.00	0.01
Fatigue	1.00	1.00	0.00	0.01
Sleep disturbance	0.99	0.99	0.05	0.03
Social role	1.00	1.00	0.00	0.02
Pain interference	1.00	1.00	0.00	0.01

CFI indicates comparative fit index; RMSEA, Root Mean Square Error of Approximation, Standardized root mean square residual; TU, Tucker-Lewis index.

TABLE 4. Item Response Theory Analysis of the PROMIS-29 v2.0

Item code and abbreviated text	LID	MON	H_i	α	β1	β2	β3	β4	*P* S-χ2
Physical Function									
PFA11 Do vacuuming or yard work	✓	✓	0.62	2.3	−2.7	−1.3	−0.2	0.8	0.069
PFA21 Go up and down the stairs	✓	✓	0.60	2.1	−2.0	−1.0	0.2	1.4	0.001
PFA23 Walk 15 minutes	✓	✓	0.64	2.7	−1.6	−1.0	−0.2	0.6	0.007
PFA53 Run errands and shop	✓	✓	0.68	3.6	−1.7	−0.8	0.2	1.2	0.234
Anxiety									
EDANX01 Felt fearful	✓	✓	0.69	2.9	0.1	0.6	1.7	2.6	<0.001
EDANX40 Hard to focus	✓	✓	0.75	4.0	0.1	0.7	1.6	2.4	0.070
EDANX41 Worries overwhelmed	✓	✓	0.74	4.1	0.1	0.5	1.5	2.3	0.177
EDANX53 Felt uneasy	✓	✓	0.77	4.2	−0.2	0.4	1.5	2.3	0.080
Depression									
EDDEP04 Felt worthless	✓	✓	0.77	4.4	0.5	0.9	1.7	2.6	0.311
EDDEP06 Felt helpless	✓	✓	0.75	3.5	0.2	0.6	1.5	2.2	0.007
EDDEP29 Felt depressed	✓	✓	0.77	3.7	0.1	0.6	1.4	2.4	0.004
EDDEP41 Felt hopeless	✓	✓	0.79	5.4	0.5	1.0	1.6	2.5	0.313
Fatigue									
HI7 Feel fatigued	✓	✓	0.81	3.4	−1.3	−0.4	0.4	1.4	0.043
AN3 Trouble starting	✓	✓	0.81	3.3	−0.7	0.0	0.8	1.7	0.466
FATEXP41 Run-down on average	✓	✓	0.83	5.9	−1.1	−0.2	0.6	1.5	0.212
FATEXP40 Fatigued on average	✓	✓	0.84	6.3	−1.2	−0.2	0.6	1.5	0.393
Sleep disturbance									
SLEEP109 Sleep quality was	✓	✓	0.66	3.0	−2.0	−0.9	0.3	1.3	0.285
SLEEP116 Sleep was refreshing	✓	✓	0.59	2.2	−2.2	−1.2	−0.1	0.8	<0.001
SLEEP20 Problem with sleep	✓	✓	0.68	3.5	−1.3	−0.5	0.4	1.3	0.116
SLEEP44 Difficulty falling asleep	✓	✓	0.59	1.9	−0.8	−0.1	0.8	1.8	0.168
Ability to participate in social roles and activities									
SRPPER11 Trouble with leisure activities	✓	✓	0.79	4.3	−1.4	−0.9	0.1	1.1	0.255
SRPPER18 Trouble with family activities	✓	✓	0.79	4.0	−1.3	−0.9	0.2	1.2	0.470
SRPPER23 Trouble with usual work	✓	✓	0.74	3.2	−1.7	−1.1	0.0	1.0	<0.001
SRPPER46 Trouble with activities with friends	✓	✓	0.81	5.8	−1.5	−1.0	0.0	0.9	0.158
Pain interference									
PAININ9 Pain interferes with day to day activities	✓	✓	0.84	4.5	−2.0	−1.2	−0.4	0.8	0.030
PAININ22 Pain interferes with work around home	✓	✓	0.87	9.5	−1.8	−1.1	−0.3	0.7	0.057
PAININ31 Pain interferes with work around home	✓	✓	0.83	3.8	−1.6	−1.0	−0.2	0.8	0.139
PAININ34 Pain interferes with chores	✓	✓	0.85	5.4	−1.7	−1.0	−0.3	0.7	0.304

ranged from theta values –4 to 4 (Figure 1). The fatigue and social role domains captured the highest theta ranges while the depression and pain interference captured the lowest theta ranges.

DISCUSSION

This study analyzed the construct (concurrent) validity of the PROMIS-29 v2.0 with legacy and proxy measures, as well as IRT of the PROMIS tool in a sample of individuals

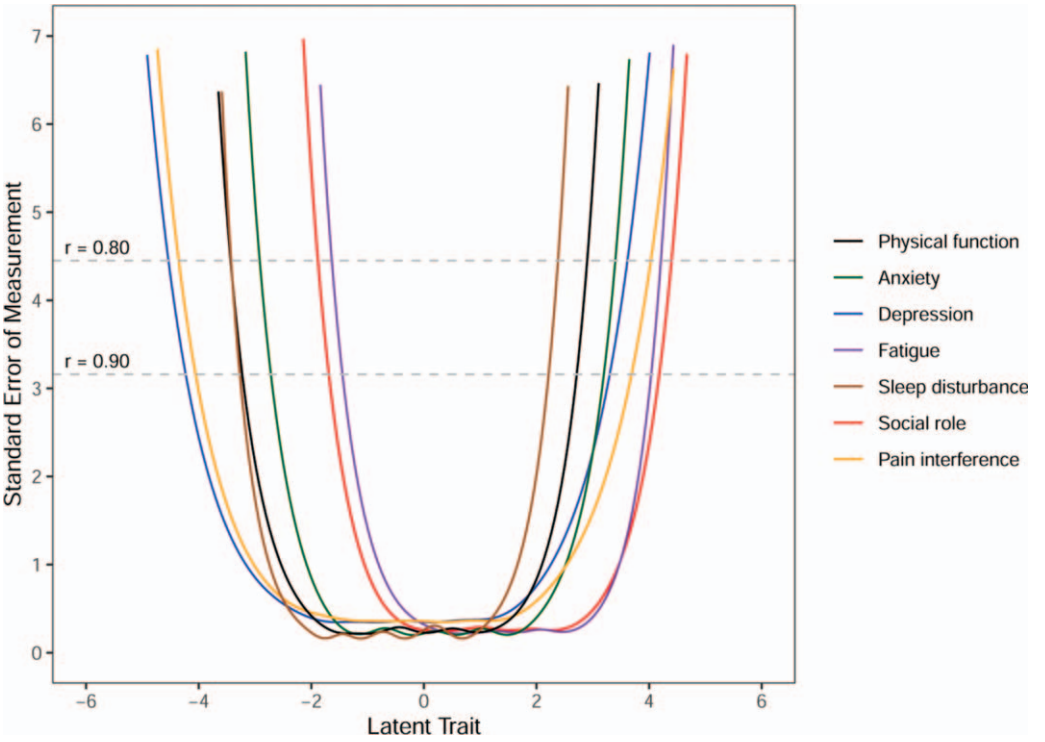

Figure 1. Scale information function for each domain of the PROMIS-29 v2.0.

who received spine surgery. Our study population had high levels of disability, and higher levels of disability influence the measurement properties of a tool.[27,28] In the study, it was hypothesized that variability in levels of correlation would be present between independently unique domains *(e.g.,* PROMIS function, pain interference, EQ5D3L-VAS, among others). The hypotheses were designed to estimate the levels of differences *a priori*. Furthermore, using IRT, it was hypothesized that all items and domains would have adequate psychometric properties, in line with their performance in the general population. This study is unique because it is the first to evaluated construct validity of the PROMIS 29.0 v2.0 instrument in a surgical population, the first to use IRT, and the first we are aware of to compare findings against the proxy measures of ADL, mobility and self-care, which are standard measures used within the spine registry.

The strength of the associations is similar to those hypothesized *a priori*. A hypothesis testing approach was selected because one should expect some variation in the correlation between two scales that address different constructs. Further, a hypothesis testing approach for construct validity is recommended by the COSMIN initiative and by the ISOQOL. This form of testing approach should evaluate the expected direction and magnitude of correlations or differences.[29] In two instances, the hypotheses failed to identify the a priori hypotheses, and interestingly, both involved the single-item, PROMIS pain intensity measure; this finding was consistent in the full sample, primary-only and revision-only. This may be explained by the fact that the measurement error of single-item pain scales tends to be quite large (e.g., 20%–30% of a scale range),[19] possibly leading to lower correlations than expected. In only one instance when the datasets were divided into primary and revision, did the study find differences in hypotheses.

With respect to the relative magnitude of the correlations, the strongest bivariate relationships were between the overall ODI percentile score and the PROMIS physical function score, the PROMIS social role domain, and the PROMIS pain interference domain. These strong correlations are consistent with past studies involving assessment against the ODI.[9,30,31] The weakest correlations included the relationships between the PROMIS anxiety domains and the NRS for Leg Pain and the individual single item measures within the QOD associated with ability to perform ADLs and Self-Described Mobility. This finding is likely a testament to the unique construct that the PROMIS anxiety items represent, constructs which not represented in the comparative measures or other PROMIS domains. This is supported by a past investigation that found strong correlations between the PROMIS anxiety items and legacy anxiety measures.[9]

IRT analyses revealed three misfit items based on the p-value of our fit statistics. When item misfit occurs, revision or deletion of the items may be an appropriate choice and should only be removed if they do not relevantly contribute to the inferences within a study. In all analyses, all model assumptions were supported by the data, the misfit items were ordered appropriately, and each misfit item demonstrated adequate discrimination. Moreover, because universally recognized fit statistics do not exist, strict adherence to model fit statistics in the presence of a valid model where all

assumptions are met is not vital.[13] Thus, we believe that these three items do not substantively affect model performance as it pertains to scaling individual differences.[32] Additionally, *P* values are strongly influenced by a study sample size, and this may have played a role in this study as well.

The IRT analyses demonstrated that the PROMIS-29 v2.0 had sufficient scale-level reliability within our population of individuals who had received spine surgery. These findings suggest that the PROMIS-29 v2.0 performs well and may be used to capture a variety of domains related to physical and mental health. The results add to the growing evidence base for the replacement of legacy disease-specific measures with generic instruments developed using modern psychometric methods,[31,33] and justify the use of the PROMIS-29 v2.0 in a population of individuals receiving lumbar spinal surgery.[34] Nevertheless, the PROMIS-29 v2.0 warrants further testing in other head-to-head comparisons in patients with spinal disorders, with other legacies (*e.g.*, Roland Morris Disability Questionnaire) or generic measures *(e.g.*, SF-12, PROMIS Global Health) that are broadly used in this patient population.[35]

Limitations

The study did not evaluate differential item functioning among subgroups of our population, and future studies should explore this measure. As the study was cross-sectional, and did not include longitudinal analyses on the domain scores' measurement error and responsiveness of PROMIS-29 v2.0, limiting the applicability of the results.

CONCLUSION

Hypothesis testing for construct validity of the PROMIS-29 v2.0 against legacy measures for back and leg pain, disability, and quality of life, and proxy measures of activities, mobility, and self-care yielded an expected direction and magnitude of correlations or differences. IRT analyses suggest that the PROMIS-29 v2.0 sufficiently captures a variety of domains related to physical and mental health. The PROMIS-29 v2.0 tool is a useful and efficient outcome measure for populations who receive spine surgery.

➢ Key Points

- ❑ Hypotheses testing of the relationship between the PROMIS 29.0 V2.0 to legacy measures for disability, pain, and quality of life yielded appropriate direction and magnitude estimates in >75% of *a priori* hypothesis, suggesting strong construct validity.
- ❑ IRT analyses suggest appropriate scale-level difficulty, discrimination, and reliability.
- ❑ These findings suggest that the PROMIS-29 V2.0 performs well and may be used to capture a variety of domains related to physical and mental health.

Supplemental digital content is available for this article. Direct URL citations appearing in the printed text are provided in the HTML and PDF version of this article on the journal's Web site (www.spinejournal.com).

References

1. Amaral V, Marchi L, Martim H, et al. Influence of psychosocial distress in the results of elective lumbar spine surgery. *J Spine Surg* 2017;3:371–8.
2. Rethorn ZD, Garcia AN, Cook CE, et al. Quantifying the collective influence of social determinants of health using conditional and cluster modeling. *PLoS One* 2020;15:e0241868.
3. Epstein NE. Spine surgery in geriatric patients: Sometimes unnecessary, too much, or too little. *Surg Neurology Internat* 2011;2:188.
4. McCormick JD, Werner BC, Shimer AL. Patient-reported outcome measures in spine surgery. *J Am Acad Orthop Surgeons* 2013;21:99–107.
5. Horn ME, Reinke EK, Couce LJ, et al. Reporting and utilization of Patient-Reported Outcomes Measurement Information System(R) (PROMIS(R)) measures in orthopedic research and practice: a systematic review. *J Orthop Surg Res* 2020;15:553.
6. Deyo RA, Dworkin SF, Amtmann D, et al. Report of the NIH Task Force on research standards for chronic low back pain. *J Pain* 2014;15:569–85.
7. Cook CE, Garcia AN, Wright A, et al. Measurement properties of the oswestry disability index in recipients of lumbar spine surgery. *Spine (Phila Pa 1976)* 2021;46:E118–25.
8. Bernstein DN, Greenstein AS, D'Amore T, et al. Do PROMIS physical function, pain interference,;1; and depression correlate to the oswestry disability index and neck disability index in spine trauma patients?. *Spine (Phila Pa 1976)* 2020;45:764–9.
9. Yee TJ, Smith BW, Joseph JR, et al. Correlation between the Oswestry Disability Index and the 4-item short forms for physical function and pain interference from PROMIS. *J Neurosurg Spine* 2019;1–6.
10. Khechen B, Haws BE, Patel DV, et al. PROMIS physical function score strongly correlates with legacy outcome measures in minimally invasive lumbar microdiscectomy. *Spine (Phila Pa 1976)* 2019;44:442–6.
11. Asher AL, Speroff T, Dittus RS, et al. The National Neurosurgery Quality and Outcomes Database (N2QOD): a collaborative North American outcomes registry to advance value-based spine care. *Spine (Phila Pa 1976)* 2014;39:S106–16.
12. McGirt MJ, Speroff T, Dittus RS, et al. The National Neurosurgery Quality and Outcomes Database (N2QOD): general overview and pilot-year project description. *Neurosurg Focus* 2013;34:E6.
13. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care* 2007;45 (5 suppl 1):S22–31.
14. Hays RD, Spritzer KL, Schalet BD, et al. PROMIS®-29 v2.0 profile physical and mental health summary scores. *Qual Life Res* 2018;27:1885–91.
15. HealthMeasures. PROMIS Scoring Manual. 2020. Available at https://www.healthmeasures.net/explore-measurement-systems/promis.
16. Chiarotto A, Terwee CB, Kamper SJ, et al. Evidence on the measurement properties of health-related quality of life instruments is largely missing in patients with low back pain: a systematic review. *J Clin Epidemiol* 2018;102:23–37.
17. Guzman JZ, Cutler HS, Connolly J, et al. Patient-reported outcome instruments in spine surgery. *Spine (Phila Pa 1976)* 2016;41:429–37.
18. Boonstra AM, Preuper HRS, Reneman MF, et al. Reliability and validity of the visual analogue scale for disability in patients with chronic musculoskeletal pain. *Int J Rehabil Res* 2008;31:165–9.
19. Chiarotto A, Maxwell LJ, Ostelo RW, et al. Measurement properties of visual analogue scale, numeric rating scale, and pain

severity subscale of the brief pain inventory in patients with low back pain: a systematic review. *J Pain* 2019;20:245–63.
20. Grandy S, Fox KM. EQ-5D visual analog scale and utility index values in individuals with diabetes and at risk for diabetes: Findings from the Study to Help Improve Early evaluation and management of risk factors Leading to Diabetes (SHIELD). *Health Qual Life Outcomes* 2008;6:18.
21. Daabiss M. American Society of Anaesthesiologists physical status classification. *Indian J Anaesth* 2011;55:111–5.
22. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–45.
23. Reeve BB, Wyrwich KW, Wu AW, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res* 2013;22:1889–905.
24. Khamis H. Measures of association: how to choose?. *J Diagnostic Med Sonography* 2008;24:155–62.
25. Prinsen CA, Vohra S, Rose MR, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. *Trials* 2016;17:449.
26. Maydeu-Olivares A, Liu Y. Item diagnostics in multivariate discrete data. *Psychol Methods* 2015;20:276–92.
27. Saltychev M, Mattie R, McCormick Z, et al. Psychometric properties of the Oswestry Disability Index. *Int J Rehabil Res* 2017;40:202–8.
28. Sheahan PJ, Nelson-Wong EJ, Fischer SL. A review of culturally adapted versions of the Oswestry Disability Index: the adaptation process, construct validity, test-retest reliability and internal consistency. *Disabil Rehabil* 2015;37:2367–74.
29. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol* 2010;10:22.
30. Pennings JS, Devin CJ, Khan I, et al. Prediction of Oswestry Disability Index (Odi) using PROMIS-29 in a national sample of lumbar spine surgery patients. *Qual Life Res* 2019;28:2839–50.
31. Purvis TE, Neuman BJ, Riley LH, et al. Comparison of PROMIS Anxiety and Depression, PHQ-8, and GAD-7 to screen for anxiety and depression among patients presenting for spine surgery. *J Neurosurg Spine* 2019;18:1–8.
32. Hambleton RK, Han N. Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In: Lenderking WR, Revicki D, editors. *Advances in Health Outcomes Research Methods, Measurement, Statistical Analysis, and Clinical Applications*. Washington, DC: International Society for Quality of life Research; 2005 pp. 57–78.
33. Chiarotto A, Roorda LD, Crins MH, et al. PROMIS physical function short forms display item- and scale-level characteristics at least as good as the roland morris disability questionnaire in patients with chronic low back pain. *Arch Phys Med Rehabil* 2020;101:297–308.
34. Vaishnav AS, McAnany SJ, Iyer S, et al. Psychometric evaluation of patient-reported outcomes measurement information system physical function computer adaptive testing in minimally invasive lumbar spine surgery: an analysis of responsiveness, coverage, discriminant validity, and concurrent validity. *J Am Acad Orthop Surg* 2020;28:717–29.
35. Chiarotto A, Boers M, Deyo RA, et al. Core outcome measurement instruments for clinical trials in nonspecific low back pain. *Pain* 2018;159:481–95.