Cross-walking the National Institutes of Health Impact Stratification Score to the PEG

Ron D. Hays PhD, Nabeel Qureshi MPHMPhil, Maria Edelen PhD, Anthony Rodriguez PhD, Mary Slaughter PhD, Patricia M. Herman NDPhD

PII: S0003-9993(22)00596-2
DOI: https://doi.org/10.1016/j.apmr.2022.08.006
Reference: YAPMR 58611

To appear in: Archives of Physical Medicine and Rehabilitation

Received date: 24 May 2022
Revised date: 3 August 2022
Accepted date: 8 August 2022

Please cite this article as: Ron D. Hays PhD, Nabeel Qureshi MPHMPhil, Maria Edelen PhD, Anthony Rodriguez PhD, Mary Slaughter PhD, Patricia M. Herman NDPhD, Cross-walking the National Institutes of Health Impact Stratification Score to the PEG, Archives of Physical Medicine and Rehabilitation (2022), doi: https://doi.org/10.1016/j.apmr.2022.08.006

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Inc. on behalf of the American Congress of Rehabilitation Medicine
Cross-walking the National Institutes of Health Impact Stratification Score to the PEG

Running head: PEG and ISS

Ron D. Hays, PhD,¹,² Nabeel Qureshi, MPH, MPhil,² Maria Edelen, PhD,³,⁴
Anthony Rodriguez, PhD,³ Mary Slaughter, PhD,² Patricia M. Herman, ND, PhD²

¹UCLA Department of Medicine, Los Angeles, CA
²RAND Corporation, 1776 Main Street, Santa Monica, CA
³RAND Corporation, Behavioral and Policy Sciences, 20 Park Plaza #920, Boston, MA
⁴Patient Reported Outcomes, Value and Experience (PROVE) Center, Department of Surgery, Brigham and Women’s Hospital, Boston, MA

Correspondence and reprint requests
Ron D. Hays, PhD
Division of General Internal Medicine and Health Services Research
UCLA Department of Medicine
1100 Glendon Avenue Suite 850
Los Angeles, CA, USA
310-794-2294  drhays@ucla.edu

Acknowledgement

Funded by the National Center for Complementary and Integrative Health (NCCIH). Grant No. 1R01AT010402-01A1. NCCIH had no role in the design; data collection, analysis, or interpretation; or writing of this manuscript.

This material has not been presented elsewhere.

The authors have no conflicts of interest in association with this manuscript.
Abstract

**Objective:** To crosswalk the National Institutes of Health (NIH) Pain Consortium’s Research Task Force proposed Impact Stratification Scale (ISS) to the PEG (Pain intensity, interference with Enjoyment of life, interference with General activity) scale.

**Design:** Cross-sectional data collected in 2021. Ordinary least squares regression analyses of ISS and PEG.

**Setting:** Amazon Mechanical Turk workers.

**Participants:** 1931 adults with back pain with an average age of 41 (range 19-77), 48% were female, 16% Hispanic, 7% non-Hispanic Black, 5% non-Hispanic Asian, and 71% non-Hispanic White.

**Interventions:** Not applicable.

**Main Outcome Measures:** The Patient-Reported Outcomes Measurement Information System (PROMIS®)-29+2 v2.1 survey that includes the ISS and the 3-item PEG.

**Results:** The ISS and PEG correlated 0.74 with one another. The ISS accounted for 55% of the adjusted variance in the PEG and the standardized average deviation between observed and predicted scores (normalized mean absolute error, NMAE) was 0.53. Likewise, the PEG explained 55% of the variance in the ISS with a NMAE of 0.52.

**Conclusions:** This study provides a crosswalk between the ISS and PEG that can be used to predict one from the other. The regression equations can facilitate comparisons in studies that use different measures.
Keywords: PEG, PROMIS®, ISS, crosswalks

Abbreviations:
BPI – Brief Pain Inventory
CLBP - chronic low back pain
HITs – Human Intelligence Tasks
ISS - Impact Stratification Score
MTurk – Mechanical Turk
NIH - National Institutes of Health
NMAE - Normalized mean absolute error
OLS – Ordinary least squares
PEG - Pain intensity, interference with Enjoyment of life, interference with General activity scale
PROMIS - Patient-Reported Outcomes Measurement Information System
PROPr - Patient reported outcome preference score
RTF - Research Task Force
SD – Standard deviation
Introduction

An extensive body of research has evaluated interventions directed at adults with chronic pain using patient-reported outcomes [1]. It is challenging to synthesize findings across studies because of the plethora of measures. The National Institutes of Health (NIH) Pain Consortium’s Research Task Force (RTF) on chronic low back pain noted that because of variations in study design and measures used it is “difficult to compare epidemiologic data and studies of similar or competing interventions, replicate findings, pool data from multiple studies, resolve conflicting conclusions, develop multidisciplinary consensus, or even achieve consensus within a discipline regarding interpretation of findings” [2].

The NIH RTF focused on impact in terms of pain intensity and interference with activities, and physical function [2,3]. They proposed an Impact Stratification Score (ISS) for chronic low back pain (CLBP) consisting of the Patient-Reported Outcomes Measurement Information System (PROMIS®)-29 physical function, pain interference, and pain intensity measures. Hays, Edelen et al. [4] found internal consistency reliability estimates of 0.92–0.93 for the ISS. In addition, the ISS correlated 0.75 to 0.84 with the Roland-Morris Disability Questionnaire, 0.51 to 0.75 with a single-item rating of average pain, and 0.64 to 0.71 with the PROMIS-29 v1.0 satisfaction with social role participation. The ISS was also found to be responsive to change. The area under the curve for the ISS predicting improvement on the retrospective rating of change from baseline to 6 weeks later was 0.83.

The responsiveness of the ISS to change was shown in a prospective comparative effectiveness clinical trial of 750 active-duty U.S. military personnel with low back pain [5]. As hypothesized, ISS scores improved for a substantial proportion of the sample. Thirty-seven percent of the sample improved significantly on the ISS over these 6 weeks and 59% reported on...
a retrospective change item that they were better (16% a little better, 14% moderately better, 23% much better, and 6% completely gone). Among those who improved significantly on the ISS, 89% reported they were better on the retrospective rating item. Thirty-three percent of the sample improved significantly and reported improvement on the retrospective change item, 4% improved significantly but did not report that they were better on the retrospective change item), 26% did not improve significantly but reported improvement on the change item, and 37% did not improve significantly on the ISS or report improvement on the retrospective change item.

One measure increasingly used to assess the pain experience is the PEG (Pain intensity, interference with Enjoyment of life, interference with General activity), a 3-item subset of the Brief Pain Inventory (BPI). The developers reported internal consistency reliability of 0.73 and 0.89 in two samples, and comparable construct validity to the full BPI [6]. In a subsequent clinical trial of 244 patients with persistent musculoskeletal pain of moderate severity, the PEG was better able to detect symptom change than the SF-36 Bodily Pain and PROMIS Pain Interference measures [7].

The usefulness of the ISS and PEG will be enhanced with the availability of empirical crosswalks from one to the other so that researchers can interpret their results in the context of the other measure. Crosswalks also serve to help with the integration of results from studies using only the ISS or only the PEG and can be used for meta-analyses. This study provides regression equations to predict the ISS from the PEG and vice versa.

**Methods**

**Data Source**

The data were collected in 2021 from Amazon Mechanical Turk (MTurk). MTurk is a source of temporary workers who are paid to complete tasks. The job or tasks are referred to as Human Intelligence Tasks (HITs) and include completing surveys, writing product descriptions,
coding, or identifying content in images or videos. Eligible study participants had to complete a minimum of 500 previous HITs on MTurk with a successful completion rate of at least 95%.

All participants provided electronic consent starting the survey. Those who completed a general health survey and reported currently having back pain were asked to complete a back pain survey. Those who completed the general health and back pain survey were paid $4 for participation. The study was designed to administer the general health survey to approximately 6,000 adults in order to obtain about 2,000 completed back pain surveys. All procedures were reviewed and approved by the research team's Institutional Review Board (RAND Human Subjects Research Committee FWA00003425; IRB00000051).

Measures

ISS. The PROMIS-29+2 v2.1 was administered. The ISS is made up of 9 PROMIS-29 items including 4 physical function items, 4 pain interference items and 1 pain intensity item. Physical function (without any difficulty = 1 to unable to do = 5) and pain interference (not at all = 1 to very much = 5) each contribute from 4 to 20 points, and pain intensity (0-10 rating) contributes from 0-10 points. The ISS has a possible range of 8 (least pain impact) to 50 (greatest pain impact).

PROMIS-29+2 v2.1. In addition to the 9 ISS items, the PROMIS-29+2 includes 5 multi-item scales with 4 items each (fatigue, sleep disturbance, depression, anxiety, ability to participate in social roles and activities) and a 2-item cognitive function scale [8]. In addition, physical health and mental health summary scores [9] and a single preference-based score, the PROPr, can be estimated [10].

PEG. The three PEG items are: 1) What number best describes your pain on average in the past week? 2) What number best describes how, during the past week, pain has interfered
with your enjoyment of life? 3) What number best describes, how, during the past week, pain has interfered with your general activity? PEG response options range from 0 to 10, with 10 indicating the most severe pain. The PEG scale score is the mean of the 3 items and has a possible range of 0-10.

Analysis Plan

We summarize demographic and health characteristics of the sample. Next, we report product-moment correlations of the PEG with the PROMIS-29+2 v2.1 measures. Then we estimate item-scale correlations (corrected for overlap of each item with the scale score) and internal consistency reliability [11] for the PEG and the ISS and report their means and standard deviations (SDs). A minimum bivariate correlation of 0.87 between the ISS and PEG is considered necessary for use of optimal methods such as item response theory co-calibration [12]. For correlations less than that, ordinary least squares (OLS) regression models have been used [13]. OLS models were evaluated in terms of $R^2$ and the normalized mean absolute error (NMAE). The NMAE statistic indicates the average deviation between the observed and predicted scores divided by the SD of the observed score. Lower NMAE values indicate better performance. There is no absolute rule of thumb for an acceptable NMAE but close to 0.50 is what was previously found for associations of PROMIS-29 scales with targeted disability measures [13]. We examined the PEG in predicting the ISS and vice versa. Finally, we report correlations between the PEG and ISS by gender (female, male), ethnicity (non-Hispanic, Hispanic), race (non-White, White), and education (high school or less, more than high school).

Results

As seen in Table 1, the sample of 1931 adults with back pain had an average age of 41 (range: 19-77). Forty-eight percent were female; 16% Hispanic, 7% non-Hispanic Black, 5%
non-Hispanic Asian, and 71% non-Hispanic White; 90% had more than high school education; 69% were married or living with a partner; and 69% were working full-time. The most common conditions reported were depression (49%), hypertension (41%), and anxiety (38%).

The mean PEG score was 4.02 (SD = 2.12). Internal consistency reliability of the PEG was 0.89 and item-scale correlations corrected for overlap of each item with the scale ranged from 0.71 to 0.84. Table 2 reports PROMIS-29+2 v2.1 score means and SDs. The sample of respondents had worse physical function and cognitive function, and more pain interference, pain intensity, fatigue, sleep disturbance, anxiety, depression, and worse health overall (physical health summary, mental health summary, and PROPr) than the general U.S. population. The largest differences were medium (pain intensity, pain interference, anxiety, depression) or large (PROPr) effect sizes [14]. The mean ISS score was 20.68 (SD = 8.06), falling within the “mild” range of severity [2]. Internal consistency reliability of the ISS was 0.79 and item-scale correlations (corrected for overlap) ranged from 0.59 to 0.75.

Table 3 shows that the correlations of the PEG with PROMIS-29+2 v2.1 scales ranged from -0.27 (cognitive distress) to 0.74 (ISS). Because they are all less than 0.80, these correlations support an OLS approach to cross-walking. The ISS was chosen for the crosswalk with the PEG because it had the largest correlation with it.

The ISS accounted for 55% of the variance (adjusted R-squared) in the PEG (NMAE = 0.53) and the regression equation was: PEG = -0.043982 + 0.19620 x ISS. The PEG accounted for 55% of the variance (adjusted R-squared) in the ISS (NMAE = 0.52) and the regression equation was: ISS = 9.33684 + 2.82342 x PEG. The correlations between the PEG and ISS were similar for females (r = 0.75) and males (r = 0.73) and between those with more than high school education (r =0.74) versus high school or less (r = 0.77). But the correlation differed for
Hispanics ($r = 0.60$) versus non-Hispanics ($r = 0.76$), and those who were non-White ($r = 0.68$) versus those who were White ($r = 0.77$).

**Discussion**

That the ISS had the strongest correlation of all the PROMIS-29 v2.1 measures with the PEG provides further support for the value of the pain impact measure recommended by the NIH Pain Consortium’s RTF on chronic low back pain. The study provides useful preliminary crosswalks between the PEG and ISS using regression to predict one from the other. The NMAE estimates indicate that the average deviation between the observed and predicted scores is about a half of a standard deviation. The NMAE values are like those obtained in a previous study predicting the Oswestry Disability Index from the PROMIS-29 physical function, pain interference, and pain intensity measures that are used to create the ISS [13].

**Study Limitations**

Nonetheless, this study has limitations. The correlation between the ISS and the PEG of 0.74 was below the threshold to use item response theory equating. Moreover, the MTurk sample from which the respondents with back pain were selected differs in pain, mental health, age, education, and income from that of the U.S. general population [15]. Hilton et al. [16] found that a MTurk sample was more likely to report chronic low back pain than a clinic-based sample but less average and worst pain and lower Oswestry Disability Index scores (indicating less disability). It is unclear how well the back pain subgroup focused on in this study represents adults with back pain in general. The results of this study are based on only one sample and results may vary in other samples. Finally, the correlation between the PEG and the ISS was smaller for Hispanics and those who were non-White compared to non-Hispanics and those who
were White. Hence, there is less accuracy in prediction from the crosswalks among these subgroups of the population.

Conclusions

Researchers can use the regression equations reported here to estimate one score (PEG or ISS) from the other. Regression to the mean can be accounted for using linear equating [17] that adjusts the regression model predictions to have the same mean and standard deviation as the observed dependent variable scores (see Appendix). These estimates can be used to facilitate comparisons across interventions and enhance interpretation of study results.

The prediction equations can be used for group-level comparisons, but there is too much error for use in estimating individual’s scores. Further studies are needed to evaluate the generalizability of the prediction equations derived in this study given the unknown representativeness of the sample of adults with back pain to the overall population.
References


Table 1. Characteristic of the Sample (n = 1931)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age means (range)</td>
<td>41 (19-77)</td>
</tr>
<tr>
<td>Female (%)</td>
<td>48%</td>
</tr>
<tr>
<td>Race/ethnicity</td>
<td></td>
</tr>
<tr>
<td>Hispanic (%)</td>
<td>16%</td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td></td>
</tr>
<tr>
<td>White (%)</td>
<td>71%</td>
</tr>
<tr>
<td>Black (%)</td>
<td>7%</td>
</tr>
<tr>
<td>Asian (%)</td>
<td>5%</td>
</tr>
<tr>
<td>Other (%)</td>
<td>1%</td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>&lt; High school (%)</td>
<td>.2%</td>
</tr>
<tr>
<td>High school graduate (%)</td>
<td>10%</td>
</tr>
<tr>
<td>Some college (%)</td>
<td>16%</td>
</tr>
<tr>
<td>AA degree (%)</td>
<td>8%</td>
</tr>
<tr>
<td>Bachelor’s degree (%)</td>
<td>49%</td>
</tr>
<tr>
<td>Master’s degree (%)</td>
<td>15%</td>
</tr>
<tr>
<td>Ph.D. or professional degree (%)</td>
<td>2%</td>
</tr>
<tr>
<td>Working full-time (%)</td>
<td>69%</td>
</tr>
<tr>
<td>Marital Status</td>
<td></td>
</tr>
<tr>
<td>Married or living with partner</td>
<td>69%</td>
</tr>
<tr>
<td>Never married</td>
<td>22%</td>
</tr>
</tbody>
</table>
Separated, divorced or widowed | 9%
---|---
Hypertension (%) | 41%
Arthritis (%) | 23%
Depression (%) | 49%
Anxiety (%) | 38%
Cancer (%) | 7%
Asthma (%) | 22%
Diabetes (%) | 17%
Chronic obstructive pulmonary disease (%) | 8%
Angina (%) | 7%
Heart disease (%) | 8%
Myocardial infarction (%) | 6%

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Function (4 items)</td>
<td>46 (8)</td>
</tr>
<tr>
<td>Pain Interference (4 items)</td>
<td>56 (8)</td>
</tr>
<tr>
<td>Pain Intensity (1 item)</td>
<td>57 (9)</td>
</tr>
<tr>
<td>Ability to participate in social roles and responsibilities</td>
<td>50 (9)</td>
</tr>
<tr>
<td>Fatigue</td>
<td>53 (9)</td>
</tr>
<tr>
<td>Sleep disturbance</td>
<td>52 (9)</td>
</tr>
<tr>
<td>Cognitive function</td>
<td>48 (8)</td>
</tr>
<tr>
<td>Anxiety</td>
<td>57 (9)</td>
</tr>
<tr>
<td>Depression</td>
<td>56 (9)</td>
</tr>
<tr>
<td>Physical health summary score</td>
<td>46 (8)</td>
</tr>
<tr>
<td>Mental health summary score</td>
<td>46 (8)</td>
</tr>
<tr>
<td>PROPr</td>
<td>0.35 (0.21)</td>
</tr>
</tbody>
</table>

Table 2. PROMIS-29+2 v2.1 Scale Scores for the Sample (n = 1931)

Note: Higher scores mean better physical function, and ability to participate in social roles and activities. Higher scores mean better health on the physical health and mental health summary scores and on the PROPr (patient reported outcome preference score). Higher scores on the other measures indicate worse health. The PROPr is scored so that 0 is dead or as bad as being dead and 1 is perfect health. The general population mean of the PROPr = 0.52 (Hanmer, 2021). The other scales are scored on a T-score metric with a mean of 50 and SD of 10 in the U.S. general population.
Table 3. Product-moment correlations of the PEG with the PROMIS-29+2 v 2.1 scales and the Impact Stratification Score (n = 1931)

<table>
<thead>
<tr>
<th>PROMIS-29+2 Measures</th>
<th>PEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impact Stratification Score (ISS)</td>
<td>0.74</td>
</tr>
<tr>
<td>Pain intensity</td>
<td>0.70</td>
</tr>
<tr>
<td>Pain interference</td>
<td>0.68</td>
</tr>
<tr>
<td>Physical health summary score</td>
<td>-0.62</td>
</tr>
<tr>
<td>PROPr</td>
<td>-0.59</td>
</tr>
<tr>
<td>Mental health summary score</td>
<td>-0.58</td>
</tr>
<tr>
<td>Physical function</td>
<td>-0.57</td>
</tr>
<tr>
<td>Ability to participate in social roles and activities</td>
<td>-0.56</td>
</tr>
<tr>
<td>Fatigue</td>
<td>0.43</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.42</td>
</tr>
<tr>
<td>Depression</td>
<td>0.39</td>
</tr>
<tr>
<td>Sleep disturbance</td>
<td>0.28</td>
</tr>
<tr>
<td>Cognitive Distress</td>
<td>-0.27</td>
</tr>
</tbody>
</table>

Note: PROPr = patient reported outcomes preference score. Higher scores mean better physical function, ability to participate in social roles and activities, physical health summary score, mental health summary score, and PROPr. Higher scores on the other measures indicate worse health.