Is less more? A preliminary investigation of the number of response categories in self-reported pain

Karon F Cook
David Cella
Erin L Boespflug
Dagmar Amtmann

Abstract: The purpose of this study was to conduct a preliminary investigation of the number of response options for self-reports of pain interference. Responses to interference items of the 11-category Brief Pain Inventory (BPI) were obtained in a sample of 434 persons from two sites and modeled using the partial credit model. In successive calibrations, response categories were collapsed and new scores were generated. Scores based on two to three categories produced poor results. Four to five categories yielded better results. However, scoring using more than five categories did not appreciably improve the reliability, person separation, or validity of scores. These results suggest that fewer response categories—as few as five or six—may function as well as the 11 response categories that are conventionally used. The results are preliminary since the number of response categories actually presented was not manipulated in the study design. Future research should compare the reliability and validity of scores based on the BPI interference items when items are presented with the conventionally 11-response format, versus presentation with fewer response options.

Keywords: psychometrics, outcomes, quality of life, measurement, pain

Background
Pain is defined as a subjective and unpleasant sensory/emotional experience that is associated with actual or potential tissue damage.1,2 The measurement of pain presents difficult challenges, and issues regarding the scaling of pain have yet to be thoroughly conceptualized. In addition to informing pain measurement, psychometric evaluations of self-reported pain instruments potentially contribute substantively to understanding how pain is perceived.

Clinical tradition has established a ‘standard’ zero-to-ten rating scale for pain and related functional problems. A numerical rating scale with 11 response categories was recommended by IMMPACT3 and has become a gold standard in research applications. In their recommendations, the IMMPACT group did not directly address the number of response categories and did not reference any studies to support the choice of 11 categories. The common use of 11 or more response categories for the measurement of symptoms suggests an assumption that ‘more is better.’

Much of the research on the impact of the number of response categories was conducted two or more decades ago and published outside the field of health outcomes, mostly in psychology and marketing research. Taken together, a manuscript by Preston and Colman4 and another by Weng5 provide a thorough review of this literature. To summarize, Symonds6 argued in 1924 for the use of more than 20 response categories, but research since then has not supported the use of this many. Bending7,8
found similar reliability with two- to nine-category scales, but decreased reliability with 11-category scales. Using a simulation study, Cicchetti and colleagues observed increases in reliability from two- to seven-point scales, but concluded that seven response categories were ‘at least functionally interchangeable with as many as 100 such ordered categories’ (p. 35). Consistent with these results are the conclusions of others who argue that seven is the optimal number of response categories for maximizing reliability.10–12 Others have indicated a preference for fewer than seven categories.13–16

Though including a large number of categories may not add substantially to the reliability of scores, too few can be detrimental. A study by Preston and Colman and another by Weng suggested that adequate reliability is not achieved with less than three or four response categories. Reliability, validity and discrimination improved as more response categories were added, up to about seven categories.4,5

One possible explanation for the psychometric results obtained in the studies described above is that there are substantial limits in how many levels of a trait persons are able to discriminate. In 1956, George Miller published an influential article titled, ‘The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information’. In it, he summarized several psychological experiments that evaluated the number of levels of a stimulus persons could discriminate (referred to as ‘channel capacity’). With few exceptions, the upper limit of this discrimination was ‘seven, plus or minus two’ levels.

The purpose of the current study was to use a Rasch measurement model to conduct a preliminary investigation into the use of different numbers of response options for scoring items of the interference subscale of the Brief Pain Inventory (BPI). Rasch models enable evaluations of scale structure at both the item and the response category level. The BPI interference scale is a seven-item measure of pain interference that employs an 11-point scale ranging from 0 (‘Does not interfere’) to 10 (‘Completely interferes’).18 We compared the results of scoring the BPI using all 11 categories with results obtained by collapsing adjacent categories and scoring items with two to 10 response categories.

Methods

Brief Pain Inventory

The BPI is a multidimensional instrument that includes items evaluating pain location, intensity, quality, and interference.18 It has been translated into many languages including Korean, Spanish and Norwegian, and its psychometric properties have been evaluated extensively in many clinical populations.19,21–32

Data

The data used in this study were archival. Responses (N = 434) to the BPI Interference items were collected in two geographical locations and several clinical populations. Data from several unrelated studies were combined for the dataset that was used. Persons with cancer were recruited through Evanston Northwestern Healthcare Center on Outcomes Research and Education (ENH-CORE) in Chicago, Illinois (N = 202). Persons with multiple sclerosis and amputation were recruited through the University of Washington Center on Outcomes Research in Rehabilitation (UW-CORR) in Seattle, Washington (N = 232). In addition to the BPI responses, information was collected on demographic and clinical variables.

Item calibrations

Using WINSTEPS software, item responses were modeled to the partial credit model (PCM), a Rasch model appropriate for calibrating items that have more than two response categories. With the PCM Masters conceptualized the category responses associated with a given polytomous item as a series of successive ‘steps.’ An examinee either succeeds or fails each step within an item. The total number of steps passed serves as an examinee’s category score. Masters defined the probability of a given category score as:

\[
P_{x}(\theta) = \frac{\exp\left[\sum_{k=0}^{x} (\theta b_{k})\right]}{\sum_{k=0}^{m_{i}} \exp\left[\sum_{k=0}^{x} (\theta b_{k})\right]} \quad (x = 0, \ldots, m_{i}),
\]

where

- \(b_{k}\) = the difficulty of the step associated with the category score, \(x\), for item \(i\), and,
- \(m_{i}\) = the highest possible score on item \(i\).

Though the response categories must be ordered when using the PCM, the step difficulties do not have to be ordered; ie, reversals are allowed.

Datasets

Using the original data, a total of 11 datasets were created, and calibrated to the PCM. One of these was the dataset with the original 0–10 coding. For an additional nine datasets, adjacent categories were collapsed uniformly across items in order to obtain datasets with 2–10 categories per
item. For example, for the 10-response category dataset, the last two response categories (‘9’ and ‘10’) were collapsed so that persons endorsing either ‘9’ or ‘10’ were assigned the same category score. For the sake of clarity, we refer to the nine datasets developed to have an equal number of categories across all items as the ‘Uniform Response’ datasets.

Table 1 describes how responses in each dataset were recoded. For the Uniform Response datasets, we favored collapsing those adjacent categories which, across items, had the least responses. Because the scores were skewed positively, we more often collapsed adjacent higher response categories. Because the items within a given uniform response dataset all had the same number of response categories, we attempted to make decisions regarding which categories to collapse based on the characteristics of all items within a dataset. However, except for favoring the collapsing of sparsely-populated response categories, decisions regarding which categories to collapse were somewhat arbitrary, and a large number of other combinations of recodings would have been equally defensible.

In a final dataset, decisions about which pairs of response categories to collapse were made at the item level, and the number of recoded categories was not the same for all items. We refer to this as the ‘Variable Response’ dataset. As a first step in developing this dataset, we collapsed categories so that there would be at least 20 observations in each category. The resulting data were then recalibrated using the PCM, and the process was repeated. For each item, the collapsing of categories continued until every remaining item category was the most probable response for persons in some range of the trait being measured. Based on plots of the CCFs we identified response categories for which this was not the case. As the next step in selecting response categories to collapse, we chose a single pair of adjacent categories, neither of which were ever the most likely response at any level of theta, and collapsed these into a single category. The data were then recalibrated using the PCM, and the process was repeated. For each item, the collapsing of categories continued until every remaining item category was the most probable response for at least some range of theta. In the completed dataset, one item had four response categories, and all other items had five response categories. Table 2 reports how each item was recoded to obtain the final dataset.

For all datasets, fit to the partial credit model was evaluated based on WINSTEPS-generated fit statistics. The InFit Mean Square reflects the degree to which response patterns are similar and is based on the ratio of observed residual variance to expected residual variance. Values outside the range of 0.6 to 1.4 have conventionally been viewed as indicating inadequate model fit. Person separation and reliability estimates were calculated based on calibrations of each dataset. The person reliability statistic estimates the degree to which the instrument discriminates among peoples’ trait levels and the degree to which the order of persons on the trait continuum can be replicated using a different instrument measuring the same construct. Person reliability is approximately equivalent to more traditional reliability estimates such as KR-20 and Cronbach’s alpha. Values of the person reliability range from 0 to 1, with values closer to 1 indicating better discrimination at extreme levels of the trait (eg, low pain and high pain).

To compare the validity of scores obtained based on different numbers of response categories, we conducted nonparametric significance tests using the scores as the predictor variables. We hypothesized that BPI interference scores would be related to self-reported general health,

**Table 1** Recoding of 11 category Brief Pain Inventory responses

<table>
<thead>
<tr>
<th>Number of categories in recoding</th>
<th>How response categories were recoded</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 (original)</td>
<td>0 1 2 3 4 5 6 7 8 9 10</td>
</tr>
<tr>
<td>10</td>
<td>0 1 2 3 4 5 6 7 8 9 9</td>
</tr>
<tr>
<td>9</td>
<td>0 1 2 3 4 5 6 7 7 8 8</td>
</tr>
<tr>
<td>8</td>
<td>0 1 2 3 4 5 6 6 6 7 7</td>
</tr>
<tr>
<td>7</td>
<td>0 1 2 3 3 4 4 5 5 6 6</td>
</tr>
<tr>
<td>6</td>
<td>0 1 1 2 2 3 3 4 4 5 5</td>
</tr>
<tr>
<td>5</td>
<td>0 1 1 2 2 3 3 3 4 4 4</td>
</tr>
<tr>
<td>4</td>
<td>0 0 1 1 1 2 2 2 3 3 3</td>
</tr>
<tr>
<td>3</td>
<td>0 0 0 1 1 1 2 2 2 2 2</td>
</tr>
<tr>
<td>2</td>
<td>0 0 0 0 0 1 1 1 1 1 1</td>
</tr>
</tbody>
</table>

**Table 2** Coding of dataset developed by inspection of category characteristic functions

<table>
<thead>
<tr>
<th>How response categories were recoded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
</tr>
<tr>
<td>Item 2</td>
</tr>
<tr>
<td>Item 3</td>
</tr>
<tr>
<td>Item 4</td>
</tr>
<tr>
<td>Item 5</td>
</tr>
<tr>
<td>Item 6</td>
</tr>
<tr>
<td>Item 7</td>
</tr>
</tbody>
</table>
average pain in the past seven days, and worst pain in the past seven days. We also compared BPI interference scores obtained from the various datasets with respect to their ability to distinguish among levels of pain interference in the three clinical populations: cancer, MS, and limb amputation (upper or lower).

Results

Study population

The population was predominantly female (66%) and Caucasian (90%) and relatively evenly split between the two geographic sites (54% from ENH, Chicago). The predominant diagnosis was cancer (47%). Among those for whom the stage of cancer was reported (82%), there was a relatively even spread among stages (Stage 1: 21%, Stage 2: 31%, Stage 3: 24%, and Stage 4 and higher: 24%). Thirty-one percent of the population had MS, and eight percent had received an (upper/lower) limb amputation.

Score distribution

Figure 1 presents the distribution of scores by item and response category. As expected, scores were skewed negatively; that is, more persons endorsed lower response categories indicating lower levels of pain interference.

Data/model fit

The most misfitting item in every calibration was the item that queried respondents about how much their pain interfered with sleep (BPI-6). Infit values for this item ranged between 1.29 and 1.58 and fell within the preferred 0.6 to 1.4 range for only two datasets: the dataset in which all items were scored with two categories, and the dataset in which all items were scored with three categories (infit values = 1.29 and 1.38, respectively).

Person reliability and separation

Variable Response datasets

The person reliability estimates for the Variable Response dataset was 0.86. The person separation estimate for the dataset with variable numbers of response categories was 2.45.

Uniform Response datasets

For the Uniform Response datasets, person reliability estimates ranged from 0.38 (2 categories) to 0.88 (6 and 7 categories). However, there was little variation in reliability among the datasets scored with 5 to 11 categories (0.86 to 0.88). The only exceptionally-poor result was for the two-category dataset.

The three- and four-category datasets had values of 0.81 and 0.84, respectively. Figure 2 portrays these results graphically.

Person separation estimates for the Uniform Response datasets ranged from 0.77 (3 categories) to 2.66 (6 and 7 categories). Values dropped off slightly for the datasets with the largest number of categories. For the 10- and 11-category datasets, the values were 2.47 and 2.45, respectively. The biggest decrement in separation, however, was for the datasets with two and three categories (0.77 and 0.79, respectively). These results are displayed in Figure 3.

Self-reported general health and diagnoses

Self-reports of general health (collected at University of Washington site) and diagnoses (collected at both sites) were used to evaluate how well the scores from different datasets discriminated among levels of health and among the three clinical populations. University of Washington participants responded to the 36-item Medical Outcomes Study Short-Form Health Survey (SF-36). One item asked respondents to rate their general health as ‘excellent’, ‘very good’, ‘good’, ‘fair’, or ‘poor’. Respondents were classified into five groups based on their responses to this item. To evaluate pain interference scores with respect to discrimination among levels of self-reported general health, we calculated the nonparametric Kruskal–Wallis H statistic. The probability of the null was taken as an indicator of how well the scores were discriminating. Our assumption was that lower P-values would indicate more successful discrimination among groups. A second set of analyses compared pain interference scores from different datasets with respect to their ability to discriminate among the three clinical groups in our study population: MS, amputation, and cancer.

Variable Response datasets

For the Variable Response dataset, the P-value obtained in the comparison of self-reported general health was 0.0004; while for the comparison of scores by diagnosis, the value was $5.4 \times 10^{-10}$.

Uniform Response datasets

Figure 4 presents results for the Uniform Response datasets. For all except the two-category dataset, statistical comparisons between pain interference scores and General Health scores resulted in $P$-values less than 0.001. All but the scores from the two-, three-, and four-category
datasets distinguished diagnostic groups at P-values less than 0.002. The two-, three-, and four-category datasets had P-values of 0.25, 0.07, and 0.10, respectively. In all but one comparison, the amputation group had the highest pain interference scores, followed by MS, and then cancer. In the two-category scoring, the amputation group had the highest pain interference scores, followed by cancer, and then MS.

**Figure 1** Distribution of observed item responses to the seven items of the Brief Pain Inventory (BPI) Interference scale.
Self-reported average and worst pain in past seven days

In addition to the pain interference subscale, the BPI includes items that query participants regarding their ‘worst’ and ‘average’ pain during the past seven days. Responses range from 0–10 with 0 indicating ‘no pain,’ and 10 indicating ‘pain as bad as you can imagine.’ Persons were grouped into 11 categories based on their responses to these items. The Kruskal–Wallis H statistic was calculated to evaluate the degree to which pain interference scores from each dataset differentiated among levels of average and worst pain. These results are summarized in Figure 5.

Variable Response datasets

For the dataset with variable numbers of response categories, $P$-values obtained from the comparisons of pain interference and levels of average and worst pain were $6.8 \times 10^{11}$ and $9.6 \times 10^{12}$, respectively.
Uniform Response datasets

Among the Uniform Response datasets, the comparison of groups classified based on responses to the ‘worst pain’ item yielded $P$-values substantially less than 0.001 for every dataset except the two-category one. For this dataset, the $P$-value for the worst pain comparison was 0.06. In the comparisons based on reported average pain, the value was 0.08 for the two-category dataset and 0.004 for the three-category dataset. All other $P$-values were well below 0.001.

Figure 4 Probability of the null (Kruskall–Wallis) in comparisons of Brief Pain Inventory Interference scores and self-reported general health and diagnosis.

Figure 5 Probability of the null (Kruskall–Wallis) in comparisons of self-reported average and worst pain.
Conclusions

In the current study, measuring pain interference scores obtained after collapsing down to as few as five categories was roughly equivalent to or slightly more effective than scoring based on all original 11 categories. The fact that, in all comparisons, scores obtained based on five to seven categories performed at least as well and often slightly better than did scores based on more categories raises the question of what is the optimum number of response categories for self-reporting pain interference. The design for the current study is not adequate for answering this question. Studies need to be conducted that systematically vary the number of responses presented for persons self-reporting pain interference. The resulting scores should be compared with regard to their concurrent validity and success in discriminating among known groups. However, our results do provide preliminary evidence that respondents may not distinguish 11 levels of pain interference, and asking them to attempt to do so may increase measurement error. Presenting a large number of response categories also increases the ‘cognitive load’ of self-report and may add appreciably to response burden.

Our results highlight the need for investigations regarding pain perception and self-report. Though it is common in pain assessment to present items with eleven categories or more, it is an empirical question whether people are actually able to discriminate this many levels of pain. The items used in the current study were developed to measure pain interference, not a ‘simple sensory attribute’. However, the results suggest that self-perceptions of pain interference may be similarly limited in channel capacity. Research regarding persons’ cognitive representations and discriminations of pain outcomes would inform efforts to measure it.

Acknowledgments/disclosures

The authors report no conflicts of interest in this work. This study was supported by grants from the National Institutes of Health: National Institute of Child Health and Human Development (PO1 HD33988); National Cancer Institute (CA60068); and Patient-Reported Outcomes Measurement Information System (PROMIS), a National Institutes of Health (NIH) Roadmap initiative (U01AR52171). Support also was received from the National Institute of Disability and Rehabilitation Research, Department of Education (H133B031129 and H133B031120) and the American Cancer Society (#RSG-01-221-01-PBP).

References

17. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev. 1956;63:81–97.


Appendix

Brief Pain Inventory Interference subscale items

During the past seven days, how much has pain interfered with your:
- General activity
- Mood
- Walking ability
- Normal work (includes both work outside the home and housework)
- Relations with other people
- Sleep
- Enjoyment of life

Response scale: 0–10, where 0 = ‘Does not interfere’ and 10 = ‘Interferes completely’.