Chapman
Journal of Patient-Reported Outcomes (2022) 6:59
https://doi.org/10.1186/s41687-022-00464-9

Journal of Patient-Reported Outcomes

REVIEW Open Access

Expected a posteriori scoring in PROMIS®

Robert Chapman*

Abstract

Background: The Patient-Reported Outcome Measurement Information System® (PROMIS®) was developed to reliably measure health-related quality of life using the patient's voice. To achieve these aims, PROMIS utilized Item Response Theory methods in its development, validation and implementation. PROMIS measures are typically scored using a specific method to calculate scores, called Expected A Posteriori estimation.

Body: Expected A Posteriori scoring methods are flexible, produce accurate scores and can be efficiently calculated by statistical software. This work seeks to make Expected A Posteriori scoring methods transparent and accessible to a larger audience through description, graphical demonstration and examples. Further applications and practical considerations of Expected A Posteriori scoring are presented and discussed. All materials used in this paper are made available through the R Markdown reproducibility framework and are intended to be reviewed and reused. Commented statistical code for the calculation of Expected A Posteriori scores is included.

Conclusion: This work seeks to provide the reader with a summary and visualization of the operation of Expected A Posteriori scoring, as implemented in PROMIS. As PROMIS is increasingly adopted and implemented, this work will provide a basis for making psychometric methods more accessible to the PROMIS user base.

Introduction

The Patient-Reported Outcome Measurement Information System® (PROMIS®) [1], is a disease-agnostic measurement system of health-related quality of life which utilizes Item Response Theory (IRT). PROMIS was originally created to leverage the benefits of IRT and Computer Adaptive Testing (CAT) to minimize patient response burden while maximizing measurement reliability. PROMIS measures have been shown to be reliable, valid and accurate in a variety of conditions and contexts [2–7]. Over the past fifteen years, there has been substantial development, adoption and implementation of PROMIS [8, 9]. Such efforts have leveraged IRT to increase the accessibility of and aid their interpretation, including T-score maps [10] and "linking" between non-PROMIS and PROMIS measures [11].

This paper aims to make PROMIS IRT scoring methods accessible to a broader audience of users who have a basic statistical background by supplementing foundational psychometric literature with non-technical descriptions and illustrative graphics. To the same end, this paper was created in the reproducibility framework of R Markdown [12]. An R Markdown document (.rmd) contains both commented statistical code and the explanatory text in this document. Both the text and statistical code for scoring is intended to be reviewed and implemented by the reader. Included in the appendices of this paper are a set of annotated statistical programming scripts for scoring PROMIS measures.

IRT foundations

The IRT methods employed in PROMIS and their foundations were developed 70–90 years ago [13–17] and have been used extensively in the educational field. Over the past two decades, researchers have also shown how IRT can be applied to patient-centered outcomes generally [18, 19] and documented how IRT has been applied in PROMIS specifically [8, 9]. This paper briefly reviews foundations of IRT in PROMIS and instead provides focused demonstration of PROMIS scoring methods.

*Correspondence: Robert.Chapman@northwestern.edu

Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, 625 North Michigan Avenue, Chicago, IL 60601, USA

Response option probabilities

Building block of IRT scoring

IRT ranks individuals and their responses to survey items across a latent trait, such as fatigue. Just as two different people might have different levels of fatigue, two different sets of responses to survey items relate to two different levels of fatigue. IRT allows us to infer where an individual most likely ranks on a latent trait continuum. The inference of where an individual ranks on a latent trait is made by transforming an individual's response to survey items (e.g., *I feel tired—Never*, *Sometimes*, and *Always*) to a set of probabilities across all levels of the latent trait. Each probability in the set represents the likelihood that an individual and their selected response options has a particular level of latent trait. Expected A Posteriori scoring reduces these probability sets to a single point-estimate of the latent trait (i.e., a score) and provides an estimate of variability and reliability of the point estimate of the latent trait (i.e., standard deviation or standard error).

Two things are required to calculate these probabilities for a PROMIS measure: item calibration parameters, such as those shown in Table 1, and the two parameter logistic IRT model shown in Formula (1). The calibration parameters represent the relationship between a sample of individuals, their responses to a set of survey items and the latent trait. The formula allows a mathematical transformation of an individual's response to an item to a set of probabilities across the spectrum of the latent trait.

In Formula (1) we can see the calibration parameters, annotated as "discrimination" and "threshold." Each item has one discrimination calibration parameter and a number of threshold calibration parameters equal to the number of response options minus one. The subscript "i" in Formula (1) indicates that these parameters vary by item, and the subscript "k" indicates that there are multiple thresholds per item. An example is PROMIS Fatigue item FATEXP42 (*In the past 7 days, how much mental energy did you have on average?*) which has five response options (*Not at all*, *A little bit*, *Somewhat*, *Quite a bit*, and *Very much*). It follows that FATEXP42 has one discrimination calibration parameter (abbreviated "a"), and four threshold calibration parameters (abbreviated and numbered from "cb1" to "cb4"). The item calibrations parameters for FATEXP42 are provided here in Table 1 for reference.

The remaining undefined variable in Formula (1) is "theta," which refers to the latent trait being measured (e.g., fatigue or physical functioning). Theta is actualized as a single number for an individual level of latent trait, ranging from negative infinity to infinity. Theta is constructed based on the population included in the calibration sample and is often scaled to have a mean center of 0 and a standard deviation of 1. For the PROMIS Profile measures (that include anxiety, depression, fatigue, pain, sleep disturbance, physical function, and satisfaction with participation in social roles), Cella and Liu [1, 20] provide a picture of the people representing the PROMIS calibrations and metric. An individual's theta score represents their level of latent trait in the context of the sample that was used to generate the calibration parameters.

Table 1 IRT Calibration Statistics for PROMIS Fatigue item FATEXP42: *How much mental energy did you have on average?*

a	cb1	cb2	cb3	cb4
1.44	−1.26	0.78	1.95	3.51

For the purposes of calculation, the range of theta is limited to ±4, with a higher theta relating to more of what is being measured, e.g., higher PROMIS Fatigue theta values relate to more fatigue or higher PROMIS Physical Function theta values relate to better physical functioning.

$$Probability = \frac{1}{1 + e^{-1 * discrimination_i (theta - threshold_{ik})}} \quad (1)$$

Once we evaluate Formula (1) for all levels of theta (e.g., −4 to 4) and for all item calibrations parameters provided in Table 1, we can create a set of probability curves that represent each item's response options. Figure 1 shows an example of how the response options of FATEXP42 are ordered across level of theta (level of fatigue), with response option *Not at all* having higher probabilities at lower levels of theta (lower fatigue), and response option *Very much* having higher probabilities at higher levels of theta (higher fatigue).

This paper demonstrates how PROMIS measures are scored using graphical representations of probability curves, such as those in Fig. 1. To aid interpretation, these probability curves are plotted with consistent formatting styles. All colors used in figures were selected from the *colorBlindness* package in R [21].

Figure 2 provides a more detailed example of how Formula (1) and FATEXP42's item calibration parameters can be used to generate sets of probabilities and plot what are referred to as item characteristic curves. The black curves in the top graph of Fig. 2 are calculated with Formula (1) and the calibration parameters in Table 1 and are labeled as the "probability associated with a threshold parameter across theta" or $P_{cb1-cb4}$(Theta). These curves represent the probability that a respondent at a given level of theta would endorse any response option above one of the response options, e.g., P_{cb2}(Theta) represents the probability that an individual would endorse the third, fourth or fifth response option (*Somewhat*, *Quite a bit*, and *Very much*), but not the first or second response option (*Not at all* and *A little bit*). The threshold parameters (e.g., cb1 = −1.26) represents value of theta where its corresponding threshold probability

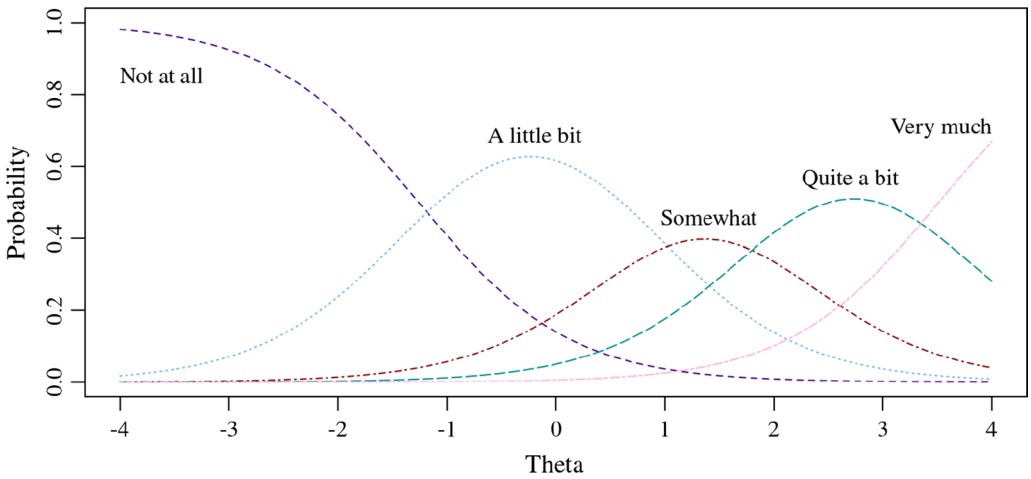

Fig. 1 Response option probabilities across theta for PROMIS Fatigue item FATEXP42

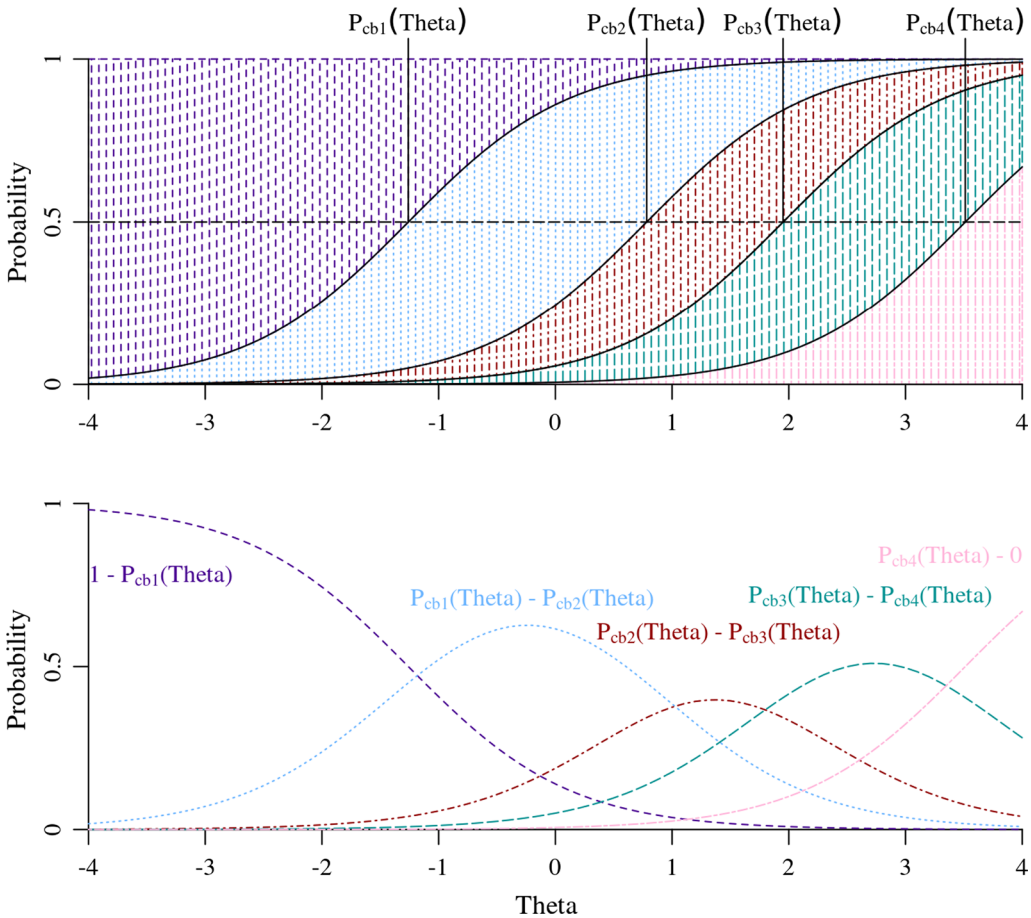

Fig. 2 Calculation of response option probabilities across theta using the graded response model calibrations from PROMIS Fatigue item FATEXP42

curve reaches 0.5, as represented by the intersection of the dotted black horizontal line and the vertical line segments underneath the threshold probability curve labels.

The bottom plot of Fig. 2 presents the same item characteristic curves in Fig. 1, but with the response option probability curves labeled with their calculations. To isolate the probability associated with an individual response option,

we calculate a set of probability differences between the probability curves of adjacent thresholds [e.g., P_{cb2}(Theta)—P_{cb3}(Theta)]. The last threshold probability curve, P_{cb4}(Theta), does not have an adjacent threshold probability because the item FATEXP42 does not have a response option greater than the fifth (*Very Much*). To calculate the probability associated with the fifth and highest response option, we subtract P_{cb4}(Theta) from 0. In other words, the probability associated with a respondent endorsing the fifth and highest response option is equal to the probability that a respondent will endorse any response option above the fourth, P_{cb4}(Theta), minus the impossibility (0 probability) that a participant will endorse a response option higher than fifth and highest. The first threshold probability curve, P_{cb1}(Theta), does not have another threshold probability curve below it. To calculate the probability associated with the lowest response option (*Not at all*) we subtract P_{cb2}(Theta) from 1. In other words, the probability of respondent endorsing the lowest response option is equal the certainty (1 probability) that a participant will endorse any response option minus the probability that a respondent will endorse a response option above the first and lowest, P_{cb1}(Theta).

The procedure of subtracting adjacent threshold probability curves to obtain probabilities curves of individual response options is reflected in the graded response model, Formula (2). To generate probabilities, we find the difference between two equations, one with threshold "k" and the other with threshold "k+1". The graded response model formula is the companion equation for interpreting PROMIS item calibration statistics and calculating probabilities. Although originally published by Samejima, the graded response model is explained in more accessible terms by Reeve, Chang, Fayers and Embretson [13, 19, 22, 23].

$$Probability = \frac{1}{1 + e^{-1*a_i(theta - cb_{ik})}} - \frac{1}{1 + e^{-1*a_i(theta - cb_{ik+1})}} \quad (2)$$

Expected a posteriori scoring

How do we go from IRT probabilities to scores?

IRT provides probability-based modeling to evaluate item- and scale-level characteristics for scale development, but we can also use IRT to find an estimate of where an individual is on the theta spectrum. In other words, we can score individuals on the latent trait. PROMIS scores are reported on the "T-score" metric, which is a linear transformation of the standardized theta scores, as shown in Formula (3). This paper reports scores on either the standardized z-score metric (labeled "theta") or the T-score metric.

$$T - score = (theta*10) + 50 \quad (3)$$

As a score calculation example, we will again use PROMIS Fatigue item FATEXP42 (*In the past 7 days, how much mental energy did you have on average?*). See item response option probability curve for the second response option (*A little bit*) in Fig. 3. A logical IRT score is the most probable level of theta, also known as the maximum likelihood of theta. Using this method, an individual that selected the second response option of FATEXP42 would be assigned a maximum likelihood score of −0.2 theta or T-score of 48, as shown in Fig. 3.

This simple example has two problems, however. The first problem comes from a practical issue in measurement and the second stems from mathematical limitations. The practical measurement issue is that we are unable to differentiate individuals at the extreme ends of our measurement scale, which occurs when respondents select the absolute highest or lowest response option in an item (e.g., *Never* or *Always*). Using another fatigue item as an example, in FATEXP29 (*In the past 7 days, how often did you feel totally drained?*) the extreme response of *Never* is likely selected by people with very different experiences of fatigue: *Never* would be selected a by respondent with low-level fatigue (e.g., feels slightly, but not totally drained over the past week), *Never* would be selected by a respondent who didn't experience fatigue (e.g., didn't feel drained at all over the past week) and *Never* would be selected by a respondent who had an unusually high energy over the past week. While the extreme response option of *Never* is selected by all three respondents for this item, we can be more certain that respondents with even less fatigue (or more energy) are increasingly likely to pick the *Never* response option.

This is also true for the other extreme response option, *Always*. A response of *Always* is likely to be selected by a respondent who just had a totally draining week, by a respondent who had a totally draining month, or by a respondent who had a totally draining year. The inability of an item or scale to distinguish between extreme levels is a measurement property known as the "floor" and "ceiling"

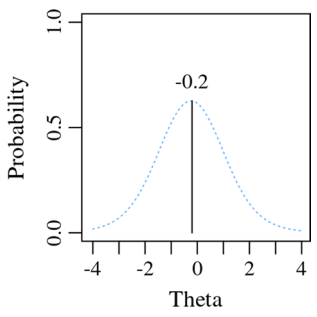

Fig. 3 *A little bit* response option probabilities across theta for PROMIS Fatigue item FATEXP42

effect [24]. The response probability curves of the extreme responses options show these floor and ceiling effects (Fig. 4). Probabilities of these extreme responses categories are assumed to be monotonic, meaning they have a constantly increasing probability of being selected with increasingly extreme levels of theta, and there is no single point of maximum likelihood for us to use as a score.

The second problem related to mathematical limitations is the infinite range that we assume exists for the latent trait (theta). All response option probability curves are asymptotic, meaning the probability curves expand over an infinite range of theta and never reach probability values of either 0 or 1. It is mathematically complex and computationally costly to perform calculations in an infinite range.

To solve these two problems, we use an IRT scoring mechanism called "Expected A Posteriori" (EAP) scoring [15, 25]. This form of scoring works by imposing constraints on how we calculate probabilities. The first constraint comes from limiting the infinite theta space to a "quadrature," which can be visualized in Fig. 5 as a set of evenly spaced points on a number line or x axis between two bounds. Boundaries of −4 theta to 4 theta or (T-scores of 10–90), with 0.1 theta increments (1 T-score point) are used. Theta can be interpreted as standard deviations of the population, a range of −4 to 4 theta encompasses 99.994% of people.

The quadrature stops the constant growth of the extreme response option's probability curve at its limits (−4 to 4), which means that a 'maximum likelihood' theta score for an extreme response option will be the same as the quadrature limit. Expanding or shrinking the limits of the quadrature (e.g., −6 to 6 or −2 to 2) will increase or decrease the scores of extreme response options. An individual who endorses an extreme response option would receive different theta scores only due to the choice of quadrature limits, not any real difference in the latent trait (e.g., fatigue).

EAP scoring uses a "prior" in the calculation of scores to address this problem. Generally, a prior is a bayesian concept that refers to our best guess of an individual's theta score before they've selected a response option [24]. The EAP scoring prior used in PROMIS is a normal distribution which reflects the population mean ($\mu=0$) and standard deviation ($\sigma=1$). It is a reasonable assumption that any individual is a member of the population.

After multiplication of the item characteristic curve by the normal prior probability curve, the extreme response probability curve is reshaped, repositioned and called the "posterior probability." The new posterior probability curve is pulled back from the quadrature limit and is no longer monotonic: instead it looks like the normal curve of the prior. The amount of the lateral repositioning of the posterior (and movement of the maximum likelihood score) away from the quadrature limit is a function of the area under the curve of the original extreme response option probability and the area under the curve of the prior.

Figure 6 shows a graphical example of the new posterior curve. In Fig. 6, the dashed purple line represents the response probabilities from FATEXP42's extreme response option (*Not at all*), the solid green line represents the prior probability curve, and the bold solid orange line represents the new posterior probability curve with a maximum probability of −0.87. The posterior (bold solid orange) can be visualized as 'splitting the difference' between the probability curves of the extreme response option (dashed purple) and the prior (solid green).

The bottom half of Fig. 6 shows the calculation of the posterior probability curve using the theta quadrature. At each increment on the theta quadrature (−4 to 4 by increments of 0.1), the response option probability is multiplied by the prior probability. For example, at a theta of −1, the response probability of 0.407 is multiplied by a prior probability of 0.242, which equals a posterior probability of 0.099. The size of the posterior probabilities are shrunk due to the multiplication of decimals, but we are only concerned with the location of the maximum likelihood point estimate that we'll use as an EAP score. Without the theta quadrature, integral calculus would be required to multiply the prior and the response option probability curves.

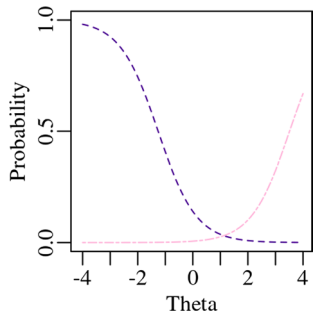

Fig. 4 Extreme response option probabilities across theta for PROMIS Fatigue item FATEXP42

Fig. 5 Theta quadrature as number line

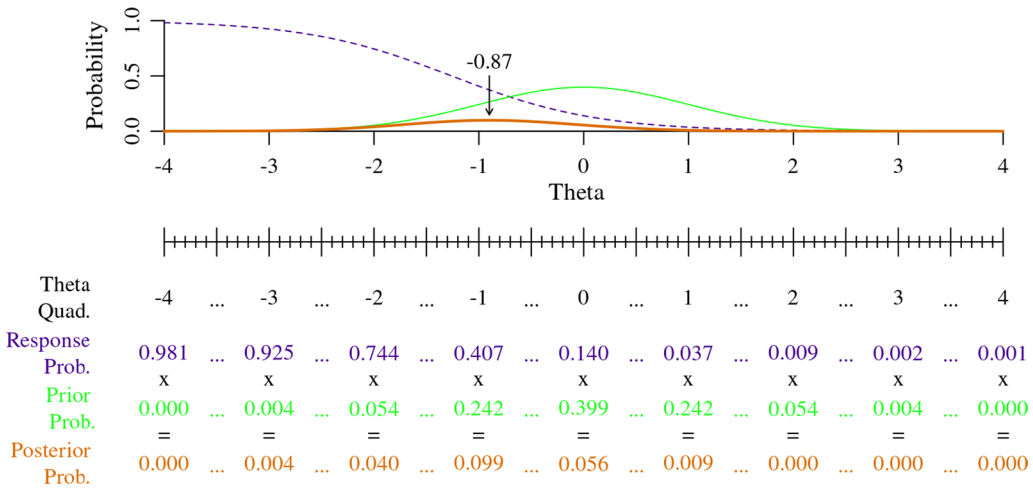

Fig. 6 Calculation of extreme response option probabilities across theta for PROMIS Fatigue item FATEXP42

Figure 7 further demonstrates the method for calculating a single theta score from posterior probabilities across the theta quadrature. The quadrature again allows us to use simple multiplication in lieu of calculus, by multiplying posterior probabilities at each theta increment by their corresponding theta level to create a set of theta weighted posterior probabilities, e.g., theta of −2 multiplied by a posterior probability of 0.04 equals a weighted probability of −0.08. Dividing the sum of the weighted posterior probabilities (−1.82) by the sum of the posterior probabilities (2.08) gives us the final theta estimate (−0.87).

We originally introduced the prior into the scoring calculation in order to circumvent problems with extreme responses. However, in order to make sure that scores from all response options (extreme or not) are comparable, the prior is used in calculating all scores. This is also true for scores calculated from multiple items.

To calculate a single score from an individuals' responses to multiple items, we combine the probability curves through multiplication. This operation is analogous to calculating the joint probability of two independent events, e.g., the probability of obtaining two heads from two coin flips is calculated as $0.5 \times 0.5 = 0.25$. A combined probability can then be multiplied by the prior to obtain a posterior probability.

In calculating a score from multiple items, we multiply all response probabilities together, and then multiply by the prior to generate a set of single set of posterior probabilities, as in Fig. 8 below. Figure 8 uses two response options probabilities from PROMIS Physical Function items PFA56 (*Are you able to get in and out of a car?*) and PFC46 (*Are you able*

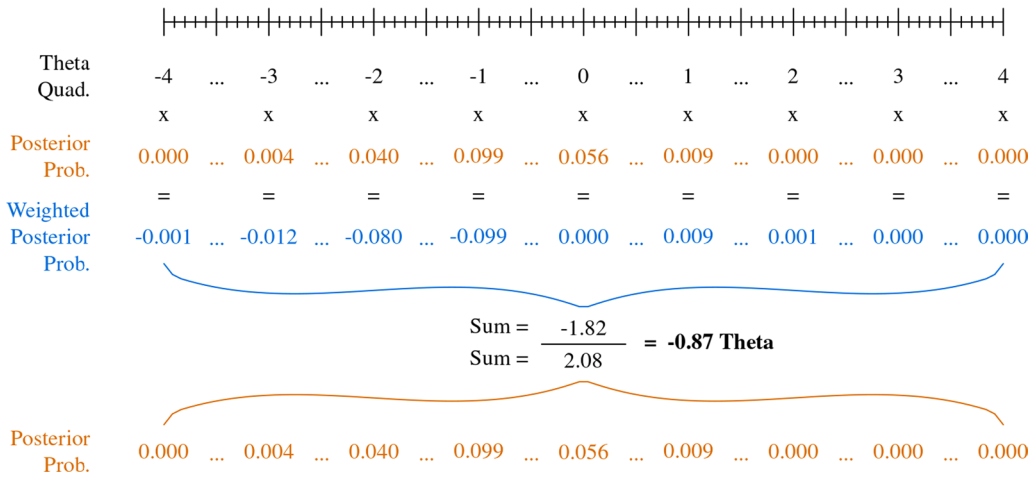

Fig. 7 Calculation of theta for PROMIS Fatigue item FATEXP42, *Not at all* response option only

to transfer from a bed to a chair and back?). The calibration statistics for PFA56 and other PROMIS Physical Function items mentioned in this work can be found in the first Table of Rose et al. [26] without the "PF" item code prefix, e.g., "A56" is the same as "PFA56."

The probabilities in the graph of Fig. 8 are scaled to make the posterior probability curve more visible. The dashed purple line represents the scaled response probabilities for extreme response option of PFC46, *Unable to do* and the dot-dashed brown line represents the scaled response probabilities of PFA56, *With some difficulty*. The solid green line represents the scaled prior probabilities and bold solid orange line represents the scaled probabilities of the posterior. The process for calculating a single theta score from multiple items is the same as in the single item example in Fig. 7.

Practical considerations of EAP scoring

There are three practical considerations of EAP scoring: one consideration related to the ordering of items, one related to score resolution, and another related the bias of prior.

Figure 8 shows that simple multiplication can be used to combine IRT response probabilities of multiple items. A property of multiplication is that any order or arrangement of multiplications has the same result (e.g., $1\times2\times3=3\times2\times1$). Consequently, the order of items doesn't matter in score calculation; item responses combined in any order will result in the same score.

Table 2 IRT to Raw Sum Score Look-up Table

Raw Sum score	IRT Theta score
3	−3.59
4	−3.36
5	−3.15
6	−2.96
7	−2.78
8	−2.60
9	−2.42
10	−2.23
11	−2.03
12	−1.81
13	−1.54
14	−1.18
15	0.21

The insensitivity to item order in IRT scoring also means that the resolution of scores increases exponentially with the number of items answered. One item with five response options has 5 possible IRT scores ($5^1=5$), two items have 25 possible IRT scores ($5^2=25$) and three items have 125 possible IRT scores ($5^3=125$). This is a large increase in score resolution over raw sum scoring methods, in which the same three items have only 13 possible sum scores, ranging from 3 to 15. Greater score resolution allows scores to be

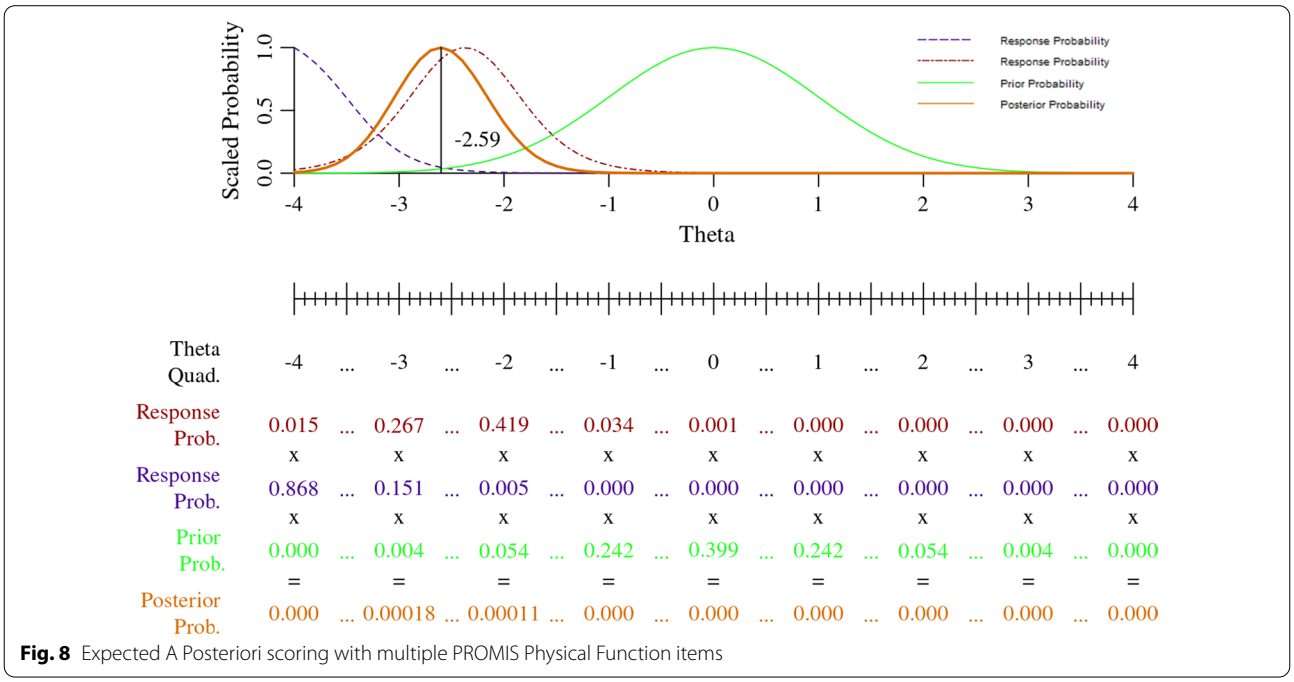

Fig. 8 Expected A Posteriori scoring with multiple PROMIS Physical Function items

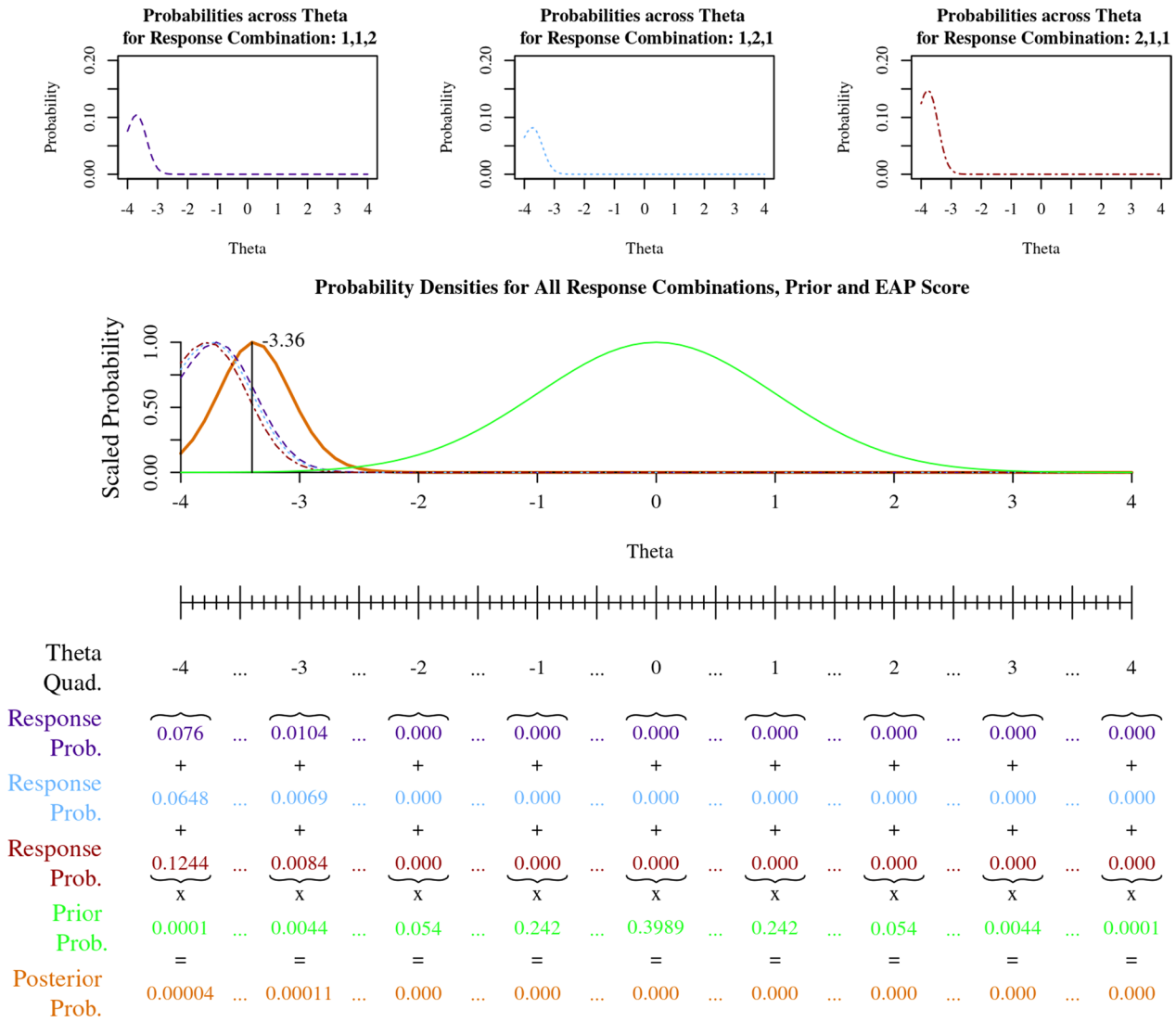

Fig. 9 Calculation of most likely IRT score for raw sum score of 4

more sensitive to an individual's responses to a set of items and is a component of score precision.

As shown in Fig. 6, multiplication by the prior biases an EAP score inward. However, since the prior is only multiplied once in calculation of the posterior, it's influence on the final EAP score will shrink as more items added into the calculation.

For these reasons, this paper doesn't recommend EAP scoring with fewer than 3 items. The shortest PROMIS Profile short form has 4 items and adult PROMIS CAT will administer 4 items as the standard minimum. There are few PROMIS short forms with less than 4 items, including 2 item Global Physical and Mental Health scales [27].

Raw sum score to IRT look-up table scoring

The previous sections demonstrate that Expected A Posteriori scoring is flexible and can be efficiently calculated by computers, but requires both statistical coding and calibration parameters to generate scores from item responses. For PROMIS users who do not have access to statistical code or calibration parameters, the HealthMeasures Scoring Service (https://www.assessmentcenter.net/ac_scoringservice) allows users upload their data to be scored with EAP scoring methods. An alternative to the HealthMeasures Scoring Service is a "look-up" table to convert a raw sum score to an EAP score. The scores in these look-up tables are calculated with EAP methods and represent the most probable theta level across all possible response pattern combinations for a single scale-level sum score [14, 28]. The maximum and

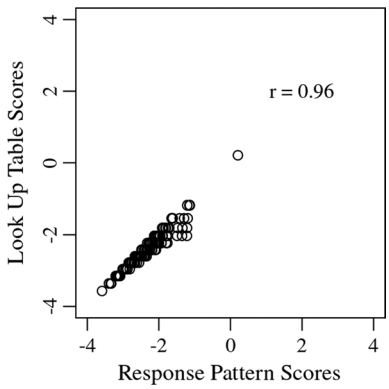

Fig. 10 Comparison of look-up and pattern response Expected A Posteriori scoring methods

minimum scale-level scores in the table relate to the floor and ceiling of the scale. Table 2 shows an example look-up table.

Figure 9 shows an example of how an EAP score for a raw sum score of 4 in Table 2 is calculated. In this example, three Physical Function items (PFA51, PFB25 and PFC46) make up a three-item scale. The minimum possible scale score on the three item scale is 3 (all three items have a raw score of 1) and maximum scale score of 15 (all three items have a raw score of 5), as shown in Table 2.

To calculate the EAP score for a scale-level raw sum score of 4, we first calculate the theta probabilities for each of the three possible combinations that sum to 4. Each response combination includes two 1's and one 2, i.e., 1,1,2; 1,2,1; 2,1,1. Each of the response probability curves are shown in the top three plots of Fig. 9. The total probability of multiple independent events (or in this case, three independent response patterns which each have a sum-score of 4) can be found by summation, shown in bottom of Fig. 9. The center plot in Fig. 9 shows each scaled probability curves, including the three dotted, dashed and dot-dashed response pattern probability curves, and their sum multiplied by the prior. The result is a posterior probability curve (bold solid orange line) with a theta maximum likelihood of −3.36 for all response combinations which sum to 4.

In order to differentiate between the two forms of scoring, one is referred to as "response pattern scoring" or "pattern response scoring" because it uses an individual's pattern of responses and the other is referred to as "look-up table scoring." Scores calculated for a look-up table are typically very highly correlated (e.g., >0.9) with response pattern scoring. Figure 10 shows a plot of look-up and pattern response scoring methods for all response option combinations of the three physical function items used in Table 2 and Fig. 9. The

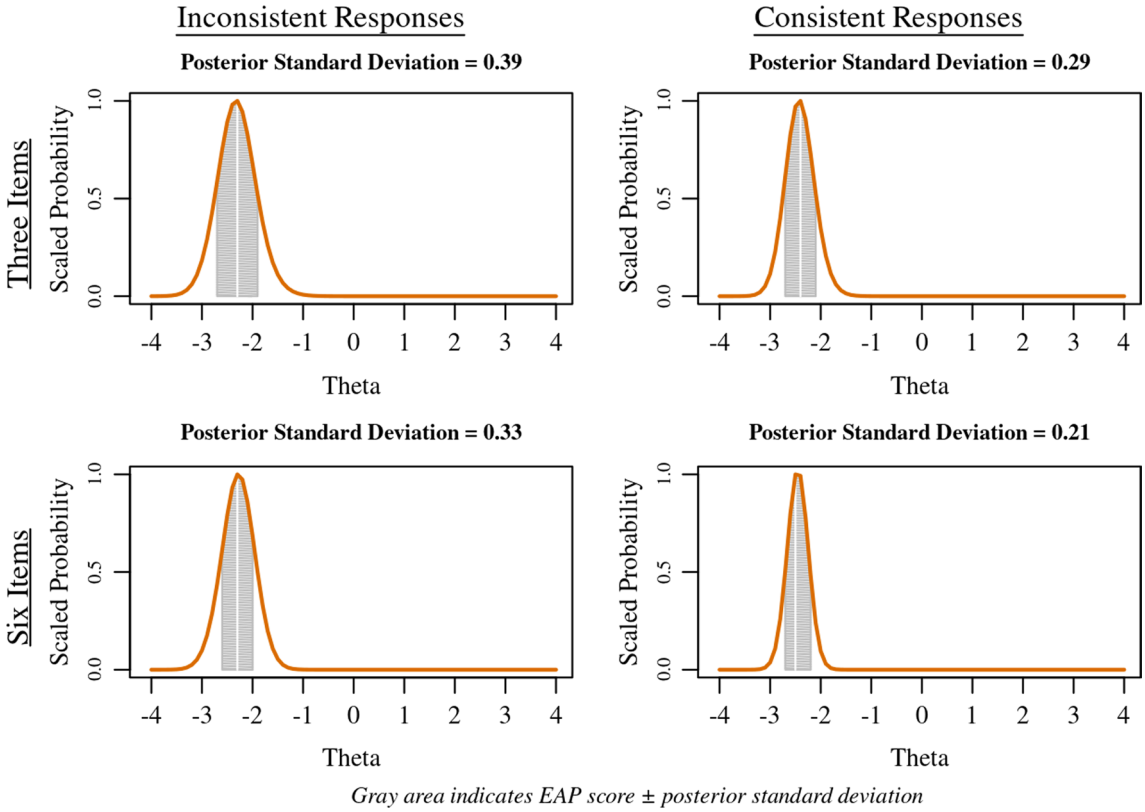

Fig. 11 Posterior standard deviation, number of items and response consistency

two scoring methods have a pearson correlation coefficient of 0.96.

It is important to recognize that relative ease of use of look-up tables is balanced by a loss in resolution in comparison to pattern response scoring. Look-up scoring treats responses of equal raw score values (1, "Unable to do") as equal, even if the responses relate to items of unequal difficulty ("Are you able to go for a walk of at least 15 min?" and "Are you able to run or jog for two miles (3 km)?"). This results in score differences or error between the pattern-response scoring and look-up table methods. The choice of implementing pattern response scoring or look-up scoring should reflect the context of measurement (e.g., regulatory decision making) and the corresponding level of precision needed. Pattern response scoring methods are more sensitive to an individual's pattern of responses and are recommended whenever possible, and where appropriate, look-up table scoring is a good alternative.

Posterior standard deviation and standard error

Because of the inclusion of the prior in estimating the theta score, EAP scores don't have a traditional standard error. Instead, we can calculate the standard deviation of the posterior distribution. The method for calculating the posterior standard deviation is the same for both pattern response and look-up table scoring methods. Formula (4) details the calculation of the posterior standard deviation.

There are parallels between the posterior standard deviation and the common standard deviation formula (5), notably, the size of the numerator of both formulas is driven by the sum of squared deviations from a single point, either the EAP score in Formula (4) or the mean in Formula (5) and both Formulas use a square root. They differ in that the squared deviation at each level of the theta quadrature is multiplied by the posterior probability before summation in Formula (4), and that the sum of the posterior distribution is the denominator in Formula (4) and the sample size is in the denominator of Formula (5).

$$Posterior\ \ SD = \sqrt{\frac{\sum \left(Posterior * (Theta\ \ Quadrature - EAP\ \ Score)^2 \right)}{\sum (Posterior)}} \tag{4}$$

$$SD = \sqrt{\frac{\sum \left(X - \overline{X}\right)^2}{N}} \tag{5}$$

While the posterior standard deviation is not a standard error, it is related in a number of ways. The posterior standard deviation is a function of the shape of the posterior probability curve, which is informed by the consistency of response probabilities and the number of items scored.

Figure 11 shows an example of the relationship between the number of items scored (e.g., 3 or 6 items), consistency of item responses (e.g., raw scores of 3,3,3 or 1,3,5) and the resulting posterior standard deviation. The gray shaded area under the bold solid orange posterior probability curve in Fig. 11 indicates a bandwidth of one standard deviation from the EAP score. Generally, a smaller posterior standard deviation occurs with a larger number of items with consistent responses, which maps onto a smaller standard error. Conversely, a smaller number of inconsistent item responses leads to a larger posterior standard deviation and larger standard error. Bock draws a direct and "near identity" relationship between the posterior standard deviation and standard error as the number of items increases (p. 437) [15].

Similar to how T-scores are a linear transformation of theta (Formula (3)), posterior standard deviations can be put on the T-score metric by multiplication by 10, e.g., a posterior standard deviation of 0.21 on the theta metric is a posterior standard deviation of 2.1 on the T-score metric.

Conclusion

Expected A Posteriori (EAP) scoring is a flexible and efficient scoring method that can be visualized and logically explained. Item response option probabilities distributed across a latent trait spectrum, theta, are the building blocks of EAP scoring and the maximum likelihood of these probabilities can provide a score estimate. An EAP score represents the level of latent trait experienced by the respondent compared to the level of latent trait present in the people who make up the calibration sample. Introduction of a theta quadrature and a Bayesian "prior" simplifies complex mathematical operations and alleviates measurement problems. For users who don't have access to the statistical code and item calibration statistics, a scale-level raw sum score to EAP score look-up table can be calculated for custom short-forms or accessed on HealthMeasures website for existing short-forms. A posterior standard deviation can be calculated for all EAP scoring methods, which reflects the score standard error. A more complete understanding of the operation and options in PROMIS EAP scoring will help ground PROMIS IRT methods with existing users and will support the further adoption and implementation of PROMIS among researchers, clinicians, industry sponsors and regulators.

Appendix

ThetaSEeap.R

The "ThetaSEeap.R" script is an R script for calculating "pattern response" EAP scores, and was originally written by Choi [29].

```
### ThetaSEeap is a commonly used script for scoring PROMIS measures with
Expected A Posteriori scoring methods.
### This scoring script was originally developed by Seung Choi, and is very
flexible. The default parameters, reflected below reflect PROMIS standards
### e.g.,  Theta quadrature of -4 to 4, by increments of 0.1, prior mean of
0, prior sd of 1, using logistic probabilities (D=1)
##### ipar = item parameter file
##### resp.data = data input to be scored
##### maxCat = maximum number of response categories (normally 5)
##### minTheta = -4.0
##### maxTheta = 4.0
##### inc = 0.1
##### prior.mean = 0.0
##### prior.sd = 1.0
##### D = 1 logistic probability estimation

thetaSE.eap<-function(ipar,resp.data,maxCat=5,model=1,minTheta=-
4.0,maxTheta=4.0,inc=0.1,prior.dist=1,prior.mean=0.0,prior.sd=1.0,D=1.0) {

  # ni: number of items
  ni<-nrow(ipar);

  # nExaminees: number of participants, or examinees
  nExaminees<<-nrow(resp.data);

  # NCAT: number of response categories for all items
  NCAT<-ipar$NCAT;

  # Create the theta quadrature using 'sequence' function - 1:81 & -4 : 4 by
increments by 0.1
  theta<-seq(minTheta,maxTheta,inc);

  # Number or length of the theta quadrature, useful in declaring size of
matrices, etc.
  nq<-length(theta);

  # Create prior based on normal density of the theta quadrature
  # using the 'dnorm' function. The prior is
  # centered at (i.e., has a mean of) 0 and
  # has a standard deviation of 0.1
  if (prior.dist==1) {
    prior<-dnorm((theta-prior.mean)/prior.sd); #normal prior
  } else if (prior.dist==2) {
    prior<-exp((theta-prior.mean)/prior.sd)/(1+exp((theta-
prior.mean)/prior.sd))^2; #logistic prior
  } else if (prior.dist==3) {
```

```
    prior<-rep(1,nq); #uniform prior
  }

  # DISC: discrimination parameters from IPAR
  DISC<-ipar[["a"]];

  # CB: Threshold parameters from the IPAR file
  CB<-ipar[paste("cb",1:(maxCat-1),sep="")];


  # Prep.prob is a function to calculate item response probabilities
  # across the theta quadrature using item calibration statistics
  prep.prob<-function(){

    # Create a 3d array of zeros to store EAP probabilities with
    # array dimensions of
    # 1) nq: length of theta quadrature
    # 2) ni: number of items
    # 3) maxCat: number of response options

    pp<-array(0,c(nq,ni,maxCat));
    if (model==1) {

      # for each item, "i" (i in 1:ni)...
      for (i in 1:ni) {

        # create a 2d matrix of zeros with matrix dimensions of
        # 1) nq: length of theta quadrature
        # 2) NCAT[i]+1: number of response options for the select item plus
one
        ps<-matrix(0,nq,NCAT[i]+1);

        # Create probability boundaries of 1 (certainty) and 0
(impossibility)
        # This will be used for later subtraction of threshold probabilities
to create
        # extreme response option probabilities
        ps[,1]<-1; ps[,NCAT[i]+1]<-0;

        # For all item calibration thresholds, "k", (k in 1:(NCAT[i]-1) in
select item "i"...
        for (k in 1:(NCAT[i]-1)) {

          # between probabilities of 1 and 0 (see above),
          # populate the matrix with probabilities
          # from each item calibration threshold, k, for select item, i
          # using the graded response model formula
          ps[,k+1]<-1/(1+exp(-D*DISC[i]*(theta-CB[i,k])));
        }

        # For all item response options, "k" (k in 1:(NCAT[i]) in select item
```

```
"i"...
        for (k in 1:NCAT[i]) {
          # Fill the final "pp" matrix with response option probabilities,
          # calculated by subtracting adjacent threshold probabilities
          # note that extreme response options probabilities
          # are extreme threshold probabilities minus 1 or 0
          pp[,i,k]=ps[,k]-ps[,k+1];
        }
      }
    }
    # Model 2, coded below,  is unused in scoring PROMIS measures
    else if (model==2) {
      for (i in 1:ni) {
        cb<-unlist(CB[i,]);
        cb<-c(0,cb);
        zz<-matrix(0,nq,NCAT[i]);
        sdsum<-0;
        den<-rep(0,nq);

        for (k in 1:NCAT[i]) {
          sdsum<-sdsum+cb[k];
          zz[,k]<-exp(D*DISC[i]*(k*theta-sdsum));
          den<-den+zz[,k];
        }
        for (k in 1:NCAT[i]) {
          pp[,i,k]<-zz[,k]/den;
        }
      }
    }

    # the prep.prob function returns a finalized matrix of response option
probabilities
    return(pp);
  }

  # pp: 3d array object generated from the prep.prob() function above
  pp<-prep.prob();


  # the calcEAP function calculates Expected A Posteriori scores by using the
posterior (made up of the prior and response probabilities) for all
respondent scores
  calcEAP<-function() {

    # posterior: matrix that initially includes only the prior probability
distribution, repeated over number of rows equal to number of participants
    posterior<-matrix(rep(prior,nExaminees),nExaminees,nq,byrow=T);

    # for "i" in 1: number of items
    for (i in 1:ni) {
```

```
      # resp: matrix of all participant's response data for item "i"
      resp<-matrix(resp.data[,i],nExaminees,1);

      # prob: a set of transposed (turned 90 degrees) response probability
matrices for response options of item "i"
      prob<-t(pp[,i,resp]);

      # turn NA probabilities into 1.0
      prob[is.na(prob)]<-1.0

      # actually calculate the posterior by multiplying the prob array by the
posterior matrix
      # initially this object only contains the prior, but will add response
probabilities for each successive item's response probabilities
      posterior<-posterior*prob;
    }
    #EAP: final EAP scores, as calculate by calculating the sum of posterior
distribution by the theta quad
    #divided the sum of the posterior distribution
    EAP<-as.vector(posterior%*%theta/rowSums(posterior));
    #SE: Posterior Standard Deviation (referred to here as the standard
error)
    SE<-
as.vector(sqrt(rowSums(posterior*(matrix(theta,nExaminees,nq,byrow=T)-
matrix(EAP,nExaminees,nq))^2)/rowSums(posterior)));
    #Return a list of both the EAP and SE objects
    return(list(theta=EAP,SE=SE))
  }
  #Return a final data frame object with the EAP scores
  return(data.frame(calcEAP()));
}
```

RSSS.R

The "RSSS.R" script is an R script for calculating EAP to raw sum score "Look Up" tables, and was originally written by Choi [30].

```
### RSSS is a commonly used script for generating look-up table scoring for
PROMIS measures with Expected A Posteriori scoring methods.
### This scoring script was originally developed by Seung Choi, and is very
flexible. The default parameters, below, reflect PROMIS standards
### e.g.,  Theta quadrature of -4 to 4, by increments of 0.1, prior mean of
0, prior sd of 1, using logistic probabilities (D=1), Tscore or theta score
##### ipar = item parameter file
##### resp.data = data input to be scored
##### maxCat = maximum number of response categories (normally 5)
##### minTheta = -4.0
##### maxTheta = 4.0
##### inc = 0.1
##### prior.mean = 0.0
##### prior.sd = 1.0
##### D = 1, logistic probability estimation
##### Tscore = Tscore output (if TRUE) or theta output (if FALSE)

rsss<-function(ipar,model=1,minTheta=-
4.0,maxTheta=4.0,inc=0.01,prior.mean=0.0,prior.sd=1.0,D=1.0,maxCat=5,minScore
=1,Tscore=T){

  # NCAT: number of response categories for all items
  NCAT<-ipar[,"NCAT"]

  # DISC: discrimination parameters from IPAR
  DISC<-ipar[,"a"]

  # CB: Threshold parameters from the IPAR file
  CB<-ipar[paste("cb",1:(maxCat-1),sep="")]

  # ni: number of items
  ni<-dim(ipar)[1]

  # Create the theta quadrature using 'sequence' function - 1:81 & -4 : 4 by
increments by 0.1
  theta<-seq(minTheta,maxTheta,by=inc)

  # Number or length of the theta quadrature, useful in declaring size of
matrices, etc.
  nq<-length(theta)


  # Create a 3d array of zeros to store EAP probabilities with
  # array dimensions of
  # 1) nq: length of theta quadrature
  # 2) ni: number of items
  # 3) maxCat: number of response options

  pp<-array(0,c(nq,ni,maxCat))
```

```
  # Create prior based on normal density of the theta quadrature
  # using the 'dnorm' function. The prior is
  # centered at (i.e., has a mean of) 0 and
  # has a standard deviation of 0.1
  if (model==1) {

    # for each item, "i" (i in 1:ni)...
    for (i in 1:ni) {

      # create a 2d matrix of zeros with matrix dimensions of
      # 1) nq: length of theta quadrature
      # 2) NCAT[i]+1: number of response options for the select item plus one
      ps<-matrix(0,nq,NCAT[i]+1);

      # Create probability boundaries of 1 (certainty) and 0 (impossibility)
      # This will be used for later subtraction of threshold probabilities to
create
      # extreme response option probabilities
      ps[,1]<-1; ps[,NCAT[i]+1]<-0;

      # For all item calibration thresholds, "k", (k in 1:(NCAT[i]-1) in
select item "i"...
      for (k in 1:(NCAT[i]-1)) {

        # between probabilities of 1 and 0 (see above),
        # populate the matrix with probabilities
        # from each item calibration threshold, k, for select item, i
        # using the graded response model formula
        ps[,k+1]<-1/(1+exp(-D*DISC[i]*(theta-CB[i,k])));
      }
      pp[,i,1]<-1-ps[,1];
      pp[,i,NCAT[i]]<-ps[,NCAT[i]];

      # For all item response options, "k" (k in 1:(NCAT[i]) in select item
"i"...
      for (k in 1:NCAT[i]) {
        # Fill the final "pp" matrix with response option probabilities,
        # calculated by subtracting adjacent threshold probabilities
        # note that extreme response options probabilities
        # are extreme threshold probabilities minus 1 or 0
        pp[,i,k]=ps[,k]-ps[,k+1];
      }
    }
  }

  # Model 2, coded below,  is unused in scoring PROMIS measures
  else if (model==2) {
    for (i in 1:ni) {
      cb<-unlist(CB[i,]);
      cb<-c(0,cb);
```

```
      zz<-matrix(0,nq,NCAT[i]);
      sdsum<-0;
      den<-rep(0,nq);



      for (k in 1:NCAT[i]) {
        sdsum<-sdsum+cb[k];
        zz[,k]<-exp(D*DISC[i]*(k*theta-sdsum));
        den<-den+zz[,k];
      }
      for (k in 1:NCAT[i]) {
        pp[,i,k]<-zz[,k]/den;
      }
    }
  }

  # Sets the minimum possible score to 0
  # This isn't a PROMIS Standard, but it is preserved incase a zero anchored
scale is needed
  min.Raw.Score<-0

  # Sets the max possible score to the number of response categories minus
the number of items
  # This isn't a PROMIS Standard, but it is preserved incase a zero anchored
scale is needed
  max.Raw.Score<-sum(ipar[,"NCAT"])-ni #maximum obtainable raw score

  # nScore: variable with the number of raw sum score points (+1 max-min)
  nScore<-max.Raw.Score-min.Raw.Score+1

  # TCCinv: Create TCC scoring table object
  TCCinv<-numeric(nScore)

  # Raw.Score: vector of raw sum scores, ranging from the minimum to maximum
  Raw.Score<-min.Raw.Score:max.Raw.Score #raw scores

  # LH: likelihood matrix, with theta quadrature across the rows and number
of raw sum scores across the columns
  LH<-matrix(0,nq,nScore) #initializing distribution of summed scores

  ncat<-ipar[1,"NCAT"]

  # maxScore: variable that will store the maxScore for each item
  maxScore<-0

  # Initialize the Likelihood matrix with the probabilities from the response
probabilities of the first item, for all response probabilities (1:ncat)
  LH[,1:ncat]<-pp[,1,1:ncat]

  # idx: index variable that starts with NCAT
```

```
  idx<-ncat

  # from the second item to the last item (represented by, ni - number of
items) cycle through items included in the scoring table, "i" ...
  for (i in 2:ni) {

    # reload the 'ncat' variable with the number
    ncat<-ipar[i,"NCAT"]

    # reload 'maxScore' with number of item response categories
    maxScore<-ncat-1

    # reload the score vector with raw sum scores from 0 (minimum) to
maxScore
    score<-0:maxScore

    # prob: matrix with response probabilities for response options (1:ncat)
for item 'i'
    prob<-pp[,i,1:ncat]

    # pLH: empty (all zero's) matrix with rows equal to number of quadrature
points and columns equal to the number of scale-level raw sum scores
    pLH<-matrix(0,nq,nScore) #place holder for LH

    # for each response option, "k", in item "i" (1:ncat)...
    for (k in 1:ncat) {

      # for each level "h" in 1:idx
      for (h in 1:idx) {

        # sco is score placeholder for the raw sum score of the item (h)
existing scale-level score[k]
        sco<-Raw.Score[h]+score[k]

        # position: holder for the which Raw.Score equals sco
        position<-which(Raw.Score==sco)

        # pLH: add in LH and prob[,k]*LH[,h]
        pLH[,position]<-pLH[,position]+LH[,h]*prob[,k]
      }
    }

    # increase the idx variable by the max score for the items
    idx<-idx+maxScore

    # recreate the likelihood with the placeholder likelihood
    LH<-pLH
  }

  # Initialize Raw Sum Scoring vector
  Scale.Score<-numeric(nScore)
```

```
  # Initialize Standard Error vector
  SE<-numeric(nScore)

  # Calculate the prior distribution
  prior<-dnorm((theta-prior.mean)/prior.sd)

  # Posterior Distribution, multiply the likelihood by the prior
  posterior<-LH*prior

  # create a denominator with column sums of the posterior
  den<-colSums(posterior)

  # create denominator matrix of denominator
  den<-matrix(rep(den,rep(nq,nScore)),nq,nScore)

  # create a posterior score by denominator
  posterior<-posterior/den

  # from the each score, "j" in 1:nScore ...
  for (j in 1:nScore) {

    # load the Scale IRT EAP Score for with sum(posterior * theta)/
sum(posterior)
    Scale.Score[j]<-sum(posterior[,j]*theta)/sum(posterior[,j])

    # load the SE for EAP IRT
    SE[j]<-sqrt(sum(posterior[,j]*(theta-
Scale.Score[j])^2)/sum(posterior[,j]))
  }

  # Raw score = raw score + ni
  # This makes the sum scores anchored at 1 instead of 0
  if (minScore==1) Raw.Score<-Raw.Score+ni

  #This transforms theta to Tscore for table output
  if (Tscore) {
    Scale.Score=round(Scale.Score*10+50,1)
    SE=round(SE*10,1)
  }

  # Scoring table data frame with both Raw Scores and IRT Scale Scores
  rsss.table<-data.frame(Raw=Raw.Score,Scale=Scale.Score,SE)

  #Return the calculated look-up scoring table
  return(rsss.table)
}
```

Abbreviations

CAT: Computer adaptive test; EAP: Expected A Posteriori; IRT: Item response theory; PROMIS: Patient reported outcomes measurement information system; SD: Standard deviation.

Acknowledgements

Feedback about the legibility and accessibility the manuscript were obtained from the PROMIS user community through posting initial drafts on a PsyArXiv preprint server. The author would like to thank Dr. Erin Anderson for her feedback and assistance in the preparation of the manuscript.

Author contributions

The author designed, wrote, edited and finalized the manuscript. The author read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials
All materials used in this manuscript are openly available and included in the parent R Markdown document, which will be submitted with the manuscript. Additionally, statistical code for performing Expected A Posteriori scoring is provided in the appendices of the manuscript. No data was used in the creation of this manuscript.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 17 November 2021 Accepted: 11 May 2022
Published online: 03 June 2022

References

1. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S et al (2010) Initial adult health item banks and first wave testing of the patient-reported outcomes measurement information system (PROMIS™) network: 2005–2008. J Clin Epidemiol 63(11):1179–1194
2. Schalet BD, Pilkonis PA, Yu L, Dodds N, Johnston KL, Yount S et al (2016) Clinical validity of PROMIS depression, anxiety, and anger across diverse clinical samples. J Clin Epidemiol 73:119–127
3. Cook KF, Jensen SE, Schalet BD, Beaumont JL, Amtmann D, Czajkowski S et al (2016) PROMIS® measures of pain, fatigue, negative affect, physical function, and social function demonstrate clinical validity across a range of chronic conditions. J Clin Epidemiol 73:89–102
4. Schalet BD, Hays RD, Jensen SE, Beaumont JL, Fries JF, Cella D (2016) Validity of PROMIS physical function measures in diverse clinical samples. J Clin Epidemiol 73:112–118
5. Askew RL, Cook KF, Revicki DA, Cella D, Amtmann D (2016) Clinical validity of PROMIS® pain interference and pain behavior in diverse clinical populations. J Clin Epidemiol 73:103–111
6. Cella D, Lai J-S, Jensen SE, Christodoulou C, Junghaenel DU, Reeve BB et al (2016) PROMIS fatigue item bank had clinical validity across diverse chronic conditions. J Clin Epidemiol 73:128–134
7. Hahn EA, Beaumont JL, Pilkonis PA, Garcia SF, Magasi S, DeWalt DA et al (2016) The PROMIS satisfaction with social participation measures demonstrate responsiveness in diverse clinical populations. J Clin Epidemiol 73:135–141
8. Reeve B, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA et al (2007) Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcomes measurement information system (PROMIS). Med Care 45(5):S22-31
9. Stover AM, McLeod LD, Langer MM, Chen W-H, Reeve BB (2019) State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. J Patient-Rep Outcomes 3(1):1–6
10. Rothrock NE, Amtmann D, Cook KF (2020) Development and validation of an interpretive guide for PROMIS scores. J Patient-Rep Outcomes 4(1):16–26
11. Choi S, Schalet B, Cook KF, Cella D (2014) Establishing a common metric for depressive symptoms: linking the BDI-II, CES-d, and PHQ-9 to PROMIS depression. Psychol Assess 26(2):513–527
12. Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A et al (2021) Rmarkdown: dynamic documents for r [Internet]. Available from: https://github.com/rstudio/rmarkdown
13. Samejima F (1969) Estimation of latent ability using a response pattern of graded scores. Psychometric Society, New York
14. Lord FM, Wingersky MS (1984) Comparison of IRT true-score and equipercentile observed-score "equatings." Appl Psychol Meas 8(4):453–461
15. Bock RD, Mislevy RJ (1982) Adaptive EAP estimation of ability in a microcomputer environment. Appl Psychol Meas 6(4):431–444
16. Thissen D (2016) Bad questions: an essay involving item response theory. J Educ Behav Stat 41(1):81–89. https://doi.org/10.3102/1076998615621300 (**Internet**)
17. Symonds PM (1929) Choice of items for a test on the basis of difficulty. J Educ Psychol 20(7):481–493. https://doi.org/10.1037/h0075650 (**Internet**)
18. Nguyen TH, Han H-R, Kim MT, Chan KS (2014) An introduction to item response theory for patient-reported outcome measurement. Pat Patient-Cent Outcomes Res 7(1):23–35
19. Reeve B, Fayers P (2005) Applying item response theory modelling for evaluating questionnaire item and scale properties. In: Fayers PM, Hays RD (eds) Assessing quality of life in clinical trials: methods and practice, 2nd edn. Oxford University Press, Oxford, pp 55–73
20. Liu H (2010) Representativeness of the patient-reported outcomes measurement information system internet panel. J Clin Epidemiol 63(11):1169–1178. https://doi.org/10.1016/j.jclinepi.2009.11.021 (**Internet**)
21. Ou J (2021) colorBlindness: safe color set for color blindness [Internet]. Available from: https://CRAN.R-project.org/package=colorBlindness
22. Chang C-H, Reeve BB (2005) Item response theory and its applications to patient-reported outcomes measurement. Eval Health Prof 28(3):264–282
23. Embretson SE, Reise SP (2000) Item response theory for psychologists. Lawrence Erlbaum Associates Publishers
24. VandenBos GR (2007) APA dictionary of psychology, 1st edn. American Psychological Association, Washington
25. Bock RD, Aitkin M (1981) Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika 46(4):443–459
26. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE (2014) The PROMIS physical function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol 67(5):516–526
27. Hays R (2017) Two-item PROMIS® global physical and mental health scales. J Pat Rep Outcomes. https://doi.org/10.1186/s41687-017-0003-8
28. Cai L (2015) Lord–wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. Psychometrika 80(2):535–559
29. Choi S (2010) ThetaSEeap.r. Version 1
30. Choi S (2010) RSSS.r. Version 1

Publisher's Note