MA, JIA. Ph.D. A Reconceptualization of IRT Calibration with DIF Items in a PROMIS Fatigue Measure. (2022)
Directed by Dr. Richard M. Luecht. 90 pp.

Differential item functioning (DIF) is a statistical procedure intended for examining and evaluating test fairness. After DIF items are detected, there are three methods to deal with DIF items, which are to ignore DIF items, remove DIF items, and create two new items from the original DIF items the related demographic variable, named demographic-specific items. In PRO research, current research and practice only focus on the first two methods. The present study evaluated and compared the performance of the three methods by applying IRT calibration.

This study used real word data from MY-Health database with a subset of 1808 cancer patients to provide concrete evidence of the evaluation of the three calibration approaches. Wald test and Welch test were applied for DIF detection, then followed by using GRM and PCM for conducting IRT calibration.

The comparison among the three calibration approaches suggested that demographic-specific group approach had the best performance in item fit and person fit; it demonstrates great advantage with improving measurement precision, and at the same time, content validity of the test is still promising, which had a positive impact on clinical studies. The removed DIF item approach was less favorable; it caused new misfit items and made larger standard errors than the other two approaches. The challenge of this study was to deal with the measurement equivalence issue in an existing instrument and patient sample, and it was not aimed at modifying the existing instrument.

A RECONCEPTUALIZATION OF IRT CALIBRATION WITH DIF ITEMS IN A PROMIS FATIGUE MEASURE

by

Jia Ma

A Dissertation
Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro

2022

Approved by

Dr. Richard M. Luecht
Committee Chair

APPROVAL PAGE

This dissertation written by Jia Ma has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair ________________________
Dr. Richard M. Luecht

Committee Members ________________________
Dr. Terry A. Ackerman

____
Dr. Robert A. Henson

____
Dr. Lori D. McLeod

____
Dr. Xia Zhao

March 17, 2022

Date of Acceptance by Committee

March 14, 2022

Date of Final Oral Examination

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

In recent years, the concept of patient-centered outcomes in the healthcare industry has gained increasing focus and recognition on the demand of patient-reported outcome (PRO) measures. One of its important roles is to support drug development that PROs have been considered as key primary or secondary end points in clinical studies to evaluate the benefit and risks, and support medical product labeling claims (FDA, 2009; McLeod et al., 2011; Fehnel et al., 2013). Traditionally, the development and validation of these measures have been guided by classical test theory. With the adoption of a modern test theory framework, item response theory (IRT) can provide better measurement precision (Hambleton & Jones, 1993; Thissen & Orlando, 2001) and address practical measurement problems in health outcome research. Among the many applications of IRT, one important application is to provide a sensitive means for the detection of subtle measurement inequivalence across different subgroups.

The conceptual and measurement equivalence of self-report questionnaires are important to the evaluation of psychometric validity using demographic and cross-cultural subgroup comparisons. This is often referred to as test fairness, an important consideration in the test development of the traditional educational and psychological assessments (Camilli, 2006). Testing organizations have published standards and guidance for industry both in the United States (US) and other countries, some well-know standard, such as the Standards for Educational and Psychological Testing by American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014), the Code of Professional Responsibilities in Educational Measurement (Schmeiser et al., 1995), the Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 2008), and the Educational Testing Service (ETS) Standards for Quality and Fairness (Educational Testing

Service, 2015). A test must not discriminate unfairly between examinees of equal ability but who are different in terms of sex, race, and other demographic and culture-related factors.

Statistical procedures intended to evaluate test fairness can be referred to as differential item functioning (DIF). The methodology applies to tests having one or more items that perform differently in various subgroups, which can be varies demographic variables, such as sex, age, or social status. For applying IRT, the probability of answering an item correctly depends on the examinee's ability level on that test construct being measured by that single item, along with other relevant factors, if applicable, such as item difficulty and discrimination. If the probability is different across subgroup factors, such as race, ethnicity, sex, or socioeconomic status, then DIF may be present. The performance of a test item identified as DIF is not only related to the examinees' ability but also influenced by these subgroup factors. It may bring irrelevant variance into test scores, thereby impacting score validity. Theoretically, a test is developed to have the DIF-free construct; in practice, it requires that test developers analyze items to detect DIF and avoid administering these items in the test "unless they are judged to be fair and are required for valid measurements" (Zieky, 2013).

Several DIF detection techniques based on statistical methods have been described in the literature (Camilli, 2006; Clauser & Mazor, 1998; Millsap & Everson, 1993; Ackerman, 1992; Penfield & Camilli, 2007, Holland & Wainer, 2012), some of which were developed for dichotomously scored items: the Mantel-Haenszel test (Holland & Thayer, 1988), the standardization statistic (Dorans, 1989), the SIBTEST statistic (Shealy & Stout, 1993), logistic regression (Swaminathan & Rogers, 1990), Lord's chi-square test (Lord, 1980), Raju's area measures (Raju, 1988; Raju, 1990), and the likelihood ratio test (Thissen, Steinberg, & Wainer, 1988; Thissen, Steinberg, & Wainer, 1993). Several of these methods have been adapted for polytomous items, including Mantel test (Zwick, Donoghue, & Grima, 1993), the standardization statistic (Zwick & Thayer, 1996), the SIBTEST statistic (Chang, Mazzeo, & Roussos, 1996), logistic regression (French & Miller, 1996), and the likelihood ratio test (Kim & Cohen, 1998).

Different from traditional PRO measures that commonly use a summated score, newer PRO measures are often developed using modern psychometric theory within an IRT framework, such that patients have a different probability of endorsing a symptom or impact positively or satisfactorily performing a queried function (Fayers & Machin, 2016). The PROMIS measures (PROMIS Cooperative Group, 2008) represent this group which aimed to revolutionize the way PRO tools are selected and employed in clinical research and practice. A PROMIS Fatigue measure with polytomous items were investigated in this study and DIF is extended to the psychometric evaluation of patients with a specific disease, detecting subgroup differences associated with demographic characteristics and disease severity. Typical examples of DIF analyses in the context of PRO measures include ePRO versus paper-based PRO, female versus male responses, treatment arms in clinical trials, and levels of fatigue by older and younger age groups.

In practice using item response theory, an item with DIF indicates measurement invariance across groups, helping test developers to understand the differential item performance. A sensitivity review or expert review is then conducted to qualitatively evaluate whether the construct is measured by the test as intended; if not, the “biased” item should be considered for removal (Penfield & Camilli, 2007). For the validation of the PROMIS measures, items that exhibited DIF are considered “flawed” and possible ways to deal with DIF include removing DIF items from the measure, revising items to be free of DIF, or creating a demographic-specific item (PROMIS analysis plan, 2008). Eliminating an item exhibiting evidence of DIF may improve the fit of the data to the model, but it may reduce the reliability of the measure. In addition, because the items of a PRO measure are developed from extensive qualitative research with patient input regarding the symptoms and impacts of the disorder, the content validity can be compromised if the items and construct are changed. The PROMIS cooperative team suggested that only items with small DIF statistics be removed. With respect to creating demographic-specific items, Crane (2006, 2007a, 2007b) recoded the levels of the demographic variable that produced DIF into two variables for IRT calibration and scoring, yielding results that showed that the effect on scoring was negligible.

There is a need to evaluate and compare the performance of different methods for handling PRO measures consisting of polytomous items when they contain DIF items. This study design is limited to the methods of removing DIF items or creating demographic-group for DIF items and compare them with the original test. IRT and logistic regression method was used for evaluating the measurement equivalence of PROMIS measures for DIF to detect the "biased" items. When DIF is identified, three datasets were created: the PROMIS Fatigue measure with all items included to ignore DIF items, the PROMIS Fatigue measure with DIF items removed, and the PROMIS Fatigue measure with DIF items recoded and divided into two demographic-specific grouping variables. The data used in this study was from the Measuring Your Health (MY-HEALTH) database, a large diverse cancer population with a variety of patient characteristics and demographic and disease-specific variables for providing real world evidence. The MY-HEALTH database, when merged with the SEEK patient registry, includes further demographic information as well. Both the graded response model and the partial credit model were applied on calibrating the PROMIS Fatigue measure in this study consist of polytomous items, and the performance of the three calibration strategies were evaluated. Ideally by creating demographic-specific variables for DIF items, the measurement construct remains unchanged, and DIF ceases to be an issue both quantitatively and qualitatively. This study that evaluated the reconceptualization approach to DIF items using IRT methods and logistic regression will provide evidence to support future study using simulated data.

Research Question:

1. What is the finding from DIF analysis of the PROMIS Fatigue measure in the MY-Health study across the selected anchors?
2. Compare with the three calibration approaches, how good is the quality of item fit?

   2.1 Using Grade Response Model

   2.2 Using Partial Credit Model
3. Compare with the three calibration approaches, how good is the quality for person fit?

   3.1 Using Grade Response Model

   3.2 Using Partial Credit Model
4. Compare with the three calibration approaches, how good is the quality of overall model fit?

   4.1 Using Grade Response Model

CHAPTER II: REVIEW OF THE LITERATURE

The purpose of this chapter is to provide a brief introduction to the measurement of patient-reported outcomes and summarize the psychometric methods to be applied in this study to the PROMIS tool, including item response theory (IRT) and differential item functioning (DIF) for polytomous items. The first section provides definitions and describes the development of PRO measures. In section two, the psychometric methods are summarized. In the final section, a review of select IRT models and methods of DIF detection for polytomous items are described and discussed.

Patient-Reported Outcomes

In order to place patients at the center of health outcomes and healthcare and to provide high-quality care, PRO assessments are developed to have patients involved for evaluating the benefit and limitation of clinical studies and health services. A PRO is defined as "any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else" (US FDA, 2009). The term PRO covers a wide range of measures, such as patient's health status, quality of life, and functional status associated with disease, but all specifically refer to "self-reporting" from the patient's perspective, collected via self-report, individual interviews, or focus group interviews. Different from clinical outcomes that relate to biological treatments and responses, PROs measure a patient's "quality of life"(QoL). QoL is short for quality of life, but health-related quality of life (HRQoL) includes domains related to health (Fayers & David, 2016): general health, physical functioning, physical symptoms, emotional functioning, cognitive function, social well-being and functioning, sexual functioning, among these areas, physical, emotional, and social functioning . PRO measures are employed to identity patient status and change in a patient’s HRQoL over time or with treatment. Benefits of PRO measures include (Fayers & David, 2016): identifying side effect during treatment for curative disease, improving well-being for patients with uncurable diseases, relief of symptoms, helping with patient communications, learning patient’s preference, and supporting medical decision-making.

A large number of PRO measures have been developed and may be classified in three types: generic instruments, disease-specific instrument, and area-specific instruments. The RAND SF-36 (Hays, Sherbourne & Mazel) and EQ-5D (Herdman, et al., 2011) are examples of generic instruments that measure patients’ general health status. Examples of disease-specific instruments designed to assess HRQoL within cancer populations include the Functional Assessment of Cancer Therapy – General (FACT-G, Cella, 1993) and European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-30 (EORTC QLQ-C30; Sprangers & Bonnetain, 2014); Area-specific instruments are intended to measure a particular area comprehensively and precisely, with the Hospital Anxiety and Depression Scale (HADS, Zigmond & Snaith, 1983) and Multidimensional re Inventory (MFI, Smets et al., 1996) as examples of this type.

Instrument Development

The development of a PRO assessment is different from the development of an educational assessment. For the development of a PRO measure, qualitative research is first conducted to understand the therapeutic area and patient-related issues and typically includes a search and review of the related literature, in-depth interviews with patients, which can be individual one-on-one interviews or focus groups. Often, interviews with clinicians are also conducted. Next, items are written to cover all the issues and also construct the conceptual framework using the set of items to form the scale(s) with consideration of the expected scoring.

Finally, quantitative methods are applied to assess the structure and scoring of the new instrument, as well as its reliability and validity.

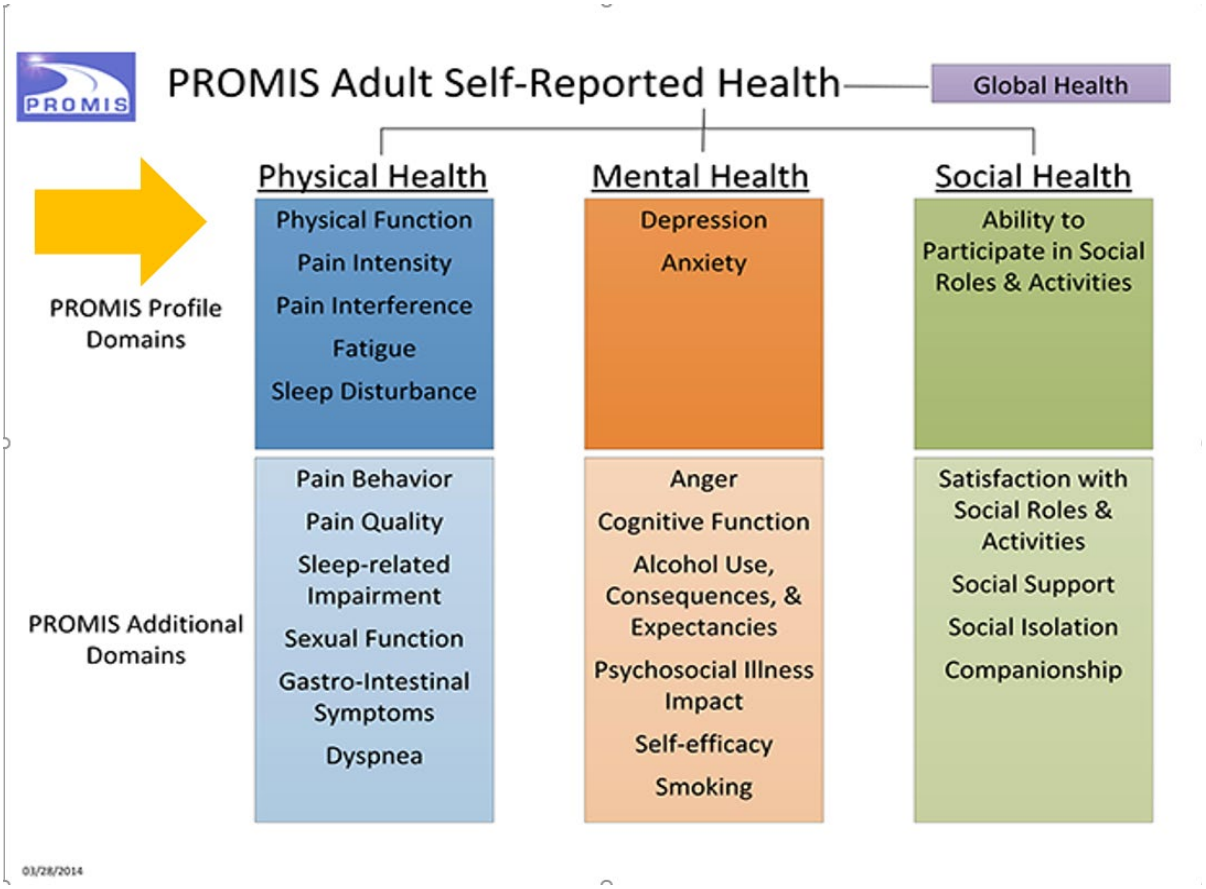
PROMIS

The Patient-Reported Outcomes Measurement Information System (PROMIS) is an initiative funded by the National Institutes of Health with the goal of developing valid and reliable PRO measures that are applicable to a wide range of chronic diseases and patient populations (PROMIS Cooperative Group, 2008; Cella et al., 2007). It is a state-of-art self-report health assessment system available in the public domain and focused on evaluating HRQOL for monitoring physical health, mental health, and social well-being by assessing symptoms and health outcomes relevant to a variety of chronic diseases including cancer. As a result, approximately 70 measures of pain, fatigue, depression, anxiety, sleep disturbance, physical function, social function, and sexual function are available for use today. PROMIS measures (Bevans et al., 2014) were developed for adults (Figure 1) as well as children and adolescents (Figure 2) and have been translated and cross-culturally adapted into more than 40 languages. PROMIS questionnaires are available in fixed-length short forms as well as for computer adaptive testing.

PROMIS measures have greater precision than most conventional paper-based questionnaires. Greater precision (less error) enhances statistical power in a less costly way than increasing sample size; providing a larger range of measurement than conventional measures decreases floor and ceiling effects as a result. PROMIS measures also require fewer items than conventional measures, thereby reducing patient burden. In the context of computer adaptive testing, PROMIS measures deliver a precise measure of health-related constructs using only four to six items at a time. For scoring, PROMIS measures provide a common metric on the T-score scale (mean = 50, standard deviation = 10). In most cases 50 equals the mean of the general U.S. population. This metric has also been formally linked to other conventional PRO measures, and when other measures are used, it may be possible to report results in the PROMIS T-score metric, which is a considerable advantage for ensuring comparability across studies.

PROMIS also offers great flexibility with multiple published short forms of the same concept and custom selection of specific items for use in computer adaptive tests. Scores from each measure can be compared to other measures derived from the same item bank.

Figure 1. PROMIS Adult Assessment

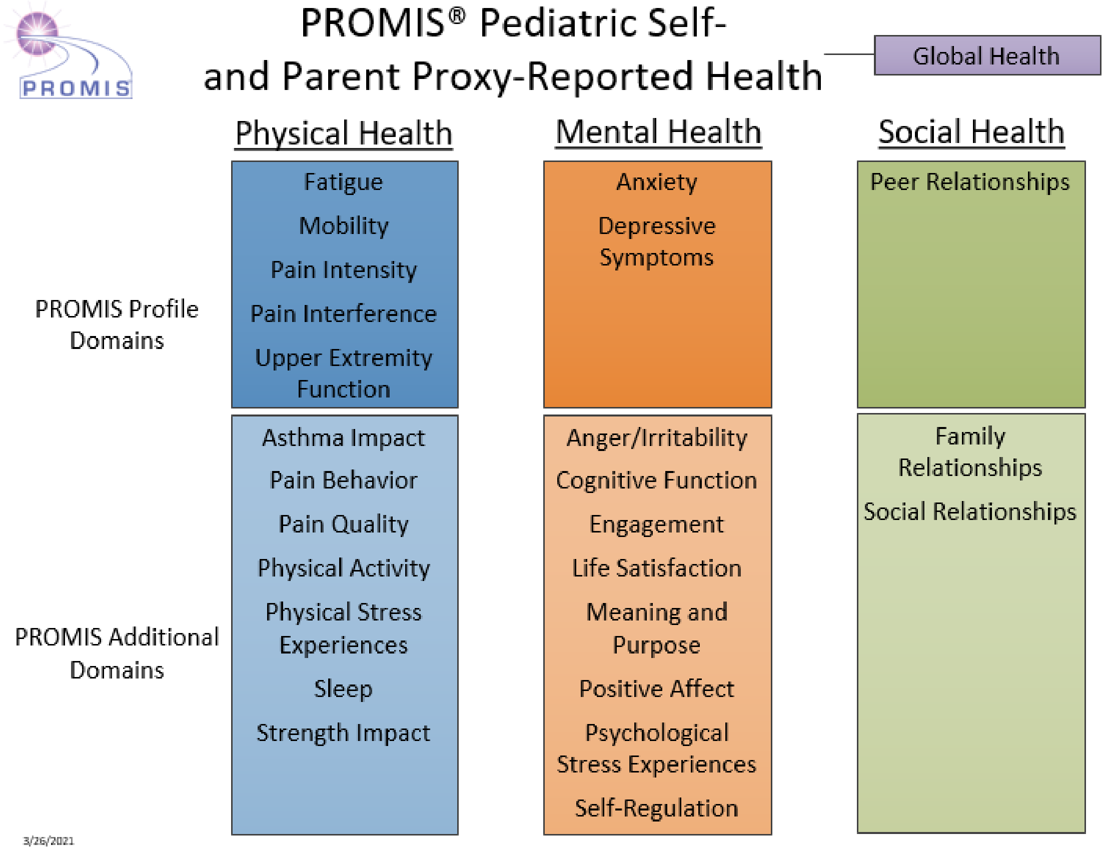


PROMIS Adult Self-Reported Health — Global Health

	Physical Health	Mental Health	Social Health
PROMIS Profile Domains	Physical Function Pain Intensity Pain Interference Fatigue Sleep Disturbance	Depression Anxiety	Ability to Participate in Social Roles & Activities
PROMIS Additional Domains	Pain Behavior Pain Quality Sleep-related Impairment Sexual Function Gastro-Intestinal Symptoms Dyspnea	Anger Cognitive Function Alcohol Use, Consequences, & Expectancies Psychosocial Illness Impact Self-efficacy Smoking	Satisfaction with Social Roles & Activities Social Support Social Isolation Companionship

03/28/2014

Figure 2. PROMIS Pediatric Assessment



PROMIS® Pediatric Self- and Parent Proxy-Reported Health

Global Health

	Physical Health	Mental Health	Social Health
PROMIS Profile Domains	Fatigue Mobility Pain Intensity Pain Interference Upper Extremity Function	Anxiety Depressive Symptoms	Peer Relationships
PROMIS Additional Domains	Asthma Impact Pain Behavior Pain Quality Physical Activity Physical Stress Experiences Sleep Strength Impact	Anger/Irritability Cognitive Function Engagement Life Satisfaction Meaning and Purpose Positive Affect Psychological Stress Experiences Self-Regulation	Family Relationships Social Relationships

3/26/2021

Psychometric evaluation

PROMIS includes an IRT-calibrated set of item banks, with each measure calibrated as an individual test construct (National Institutes of Health, 2013). Most of the PROMIS items use response scales with five ordinal categories, e.g., 1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much. These response options were selected after extensive discussion of previous work and analyses of available large data sets, in which five response options produced data with ample variability for IRT analysis with sufficient discrimination in terms of item characteristic curves, without producing failures of monotonicity, scalability, or item misfit; and performed well in cognitive testing (Reeve et al., 2007)

One interest of this study is the calibration of selected PROMIS domains. As a patient-centered standardized test, PROMIS provides the opportunity to explore and understand the important measurement properties using advanced psychometric methods. The briefing package (version 2.0) of the PROMIS instrument development and validation standards was submitted to the FDA (National Institutes of Health, 2017) and focused on the following standards:

1. Definition of Target Concept and Conceptual Model
2. Composition of Individual Items
3. Item Pool Construction
4. Determination of Item Bank Properties
5. Testing and Instrument Formats
6. Validity
7. Reliability
8. lnterpretability
9. Language Translation and Cultural Adaptation

The PROMIS cooperative group (Reeve et al., 2007) published the PROMIS psychometric evaluation and analysis plan (National Institutes of Health, 2017) for conducting the psychometric evaluation and calibration. The study used nine PROMIS domains including physical functioning, fatigue, pain, emotional distress, and social role participation etc., and data collected from a large sample (n = 7523) representative of the US general population by demographic and patient characteristics (e.g., gender, age, ethnicity, education, and disease types). The PROMIS item bank and PROMIS domains have been analyzed using both classical test theory and IRT. Factor analysis was used to confirm the underlying structures of the constructs as well as test the IRT assumptions (i.e., unidimensionality, local independence, and monotonicity). The graded response model was used for item calibration.

Data analyses were driven by a statistical analysis plan (Reeve et al., 2007) for checking IRT modeling assumptions, evaluating IRT item and model fit, and detecting potential DIF items. This procedure provided support in item bank composition, statistical and psychometric analysis to the domain teams. Results were discussed and decisions were made regarding each PROMIS item (Cella, 2010).

Item Response Theory

Overview

Item response theory (IRT), also known as latent trait theory, is a psychometric theory that is based on mathematical models to present the probability of individuals responding to individual items on psychological and educational tests as a mathematical function (Lord, 1980). The function is referred to as an item response function, which relates the characteristics of items and individuals to the probability of endorsing a given response to a given item. The item characteristic curve visually displays the item response function. There are two sets of parameters — person parameters representing the property test to be measured, such as knowledge, attitude, or ability on the unobservable trait, and item parameters that describe the items, including difficulty, discrimination, and guessing. Consequently, parameter estimation consists of item parameter estimation and ability estimation. The values for item parameters are estimated through statistical methods, such as maximum-likelihood and Bayesian methods, and the goodness-of-fit tests evaluate the appropriateness of the IRT model with respect to the data; the estimation of ability provides the position of a person's ability on the latent trait continuum.

In IRT, the reliability is conceptualized as information that reflects the precision of the measurement across the level of underlying trait, the relationship between information and standard error:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

Where SE is standard error of estimated, I is information, θ is the score estimated on the trait level. From this formula, Information is inversely related to the standard error of estimate. Thus, the greater information, the better measurement precision, which means the smaller standard error on the person score. The test information is additive from the information of each item to create the test information curve for comparing measurement precision. A test with more items provides greater information. Depending on the position on the trait continuum and the discrimination of the items, different measures will realize different test information functions. As PROMIS measures are used in this study, this review focuses on unidimensional polytomous IRT models. Detailed discussion of IRT models and their application are in the literature (Hambleton & Swaminathanm 1985; Van der Linden & Hambleton, 1997; Kolen & Brennan, 2004; Yen & Fizpatrick, 2006; Thissen, 1995; Ackerman, 1994, 1996[a], 1996[b], 2005; Luecht, 1998, 2006; Luecht & Hambleton, 2021).

Assumptions

Before applying IRT, the core assumptions of the model — unidimensionality, local independence, and monotonicity — must be evaluated. Unidimensionality assumes that a group of items appearing on the same scale measure only one latent trait. Factor analysis can be used to explore and verify unidimensionality, in which case, a 1-factor solution should be confirmed. Local independence requires that an individual's response to an item is statistically independent of other items on the same measure. This can be examined by computing the residual correlation; Q3 statistics (Yen, 1984) and LD index statistics (Chen & Thissen, 1997) can also be used to identify local independence. Monotonicity is displayed on the graph of the item character curve depicting the relationship between the trait and the responses to the item — when the trait level is increasing, the probability of a correct response also increases. In addition to these assumptions, the trait level should not depend on the items administered or the examinees sampled — this is referred to as parameter invariance (Hambleton & Jones, 1993).

Polytomous IRT Model

A polytomous item is an item with more than two response categories, as commonly used in psychosocial assessment as opposed to educational measurement which typically uses "correct" and "incorrect." In the larger family of polytomous models, based on the procedure for determining each respondent's conditional probability in a particular response category, polytomous models can be considered as direct models and indirect models (Embretson, 2009). For direct model, by its name, only one equation is needed for describing the relationship between the respondent's trait level and the probability for the response category. This is also known as an adjacent model and includes the partial credit model (PCM, Masters, 1982), rating scale model (RSM, Andrich, 1978). Indirect polytomous models, such as the graded response model (GRM, Samejima, 1969) and generalized partial credit model (Muraki, 1992), require two steps to obtain the respondent's conditional probability. The nominal response model (NRM, Bock, 1972) can be applied if the responses are not in ordered categories.

In the context of PRO assessments, items are typically developed using ordered categorical responses. For example, PROMIS physical function item PF1, "Does your health now limit you in doing vigorous activities, such as running, lifting heavy objects, participating in strenuous sports?" uses five response options: 5 = Not at all, 4 = Very little, 3 = Somewhat, 2 = Quite a lot, 1 = Cannot do. Patients with a higher numerical score report better physical well-being. The polytomous IRT models most appropriate for this type of item are the GRM and PCM.

Graded Response Model

The GRM (Samejima, 1969) is an extension of the 2-parameter logistic (2PL) model with a common discrimination parameter a (slope) as:

$$\mathrm{P}_{\mathrm{ix}}^{*}(\theta) = \frac{\exp\left[\mathrm{a_i}\left(\theta - \mathrm{b_{ij}}\right)\right]}{1 + \exp\left[\mathrm{a_i}\left(\theta - \mathrm{b_{ij}}\right)\right]}$$

Where θ represents the latent ability or trait, and its level in the subject; $\mathrm{b_{ij}}$ is a constant specific to the item, the location parameter, or category boundary for score x; $P_{ix}^{*}(\theta)$ is estimated for each between-category threshold. PROMIS measures typically use five response categories so four b parameters and one a parameter are estimated for each item using GRM. The point on the ability scale where P = 0.5 is called the discrimination parameter, $\mathrm{a_{jx}}$, another constant over response categories for a given item.

Because the GRM is an "indirect" model, the probability of responding to each category is captured by obtaining the item response functions (IRFs) from the difference between adjacent step functions. The category response probability for each x is obtained by subtraction:

$$\mathrm{P_{ix}}(\theta) = \mathrm{P^*_{ix}}(\theta) - \mathrm{P^*_{i(x+1)}}(\theta)$$

Partial Credit Model

The PCM (Masters, 1982, 1987, 1988) is an application of the unidimensional Rasch model (Rasch, 1980) applied to responses with two or more ordered categories. As a "direct model," the probability of responding in a particular category can be directly expressed as the exponential divided by the sum of exponentials. At category j, the category response function can be expressed as:

$$\mathrm{P_{ix}}(\theta) = \frac{\exp\left[\sum_{j=0}^{x} \theta - \mathrm{b_{ij}}\right]}{\sum_{r=0}^{m_i} \exp\left[\sum_{j=0}^{x} \theta - \mathrm{b_{ij}}\right]}$$

Where b_{ij} (j = 1, 2, … , k) is the location or severity of a category, such that higher values of b_{ij} reflects the greater severity of a category compared to all the response categories, P represents the probability of success on category j given a subject's trait level θ.

The GRM and the PCM can both be applied to items with responses with more than two ordered categories; each category can be described using binary trace lines, such that for items with k response categories, k –1 trace lines are presented; therefore, in both models, item responses are divided into k –1 subitems (Thissen & Steinberg, 1986). The difference in the computational procedures lies in the fact that the GRM requires a cumulative approach and the PCM involves an adjacent approach.

The GRM is the most flexible polytomous IRT model, and with separate discrimination parameters and separate category response parameters estimated for each item, the GRM provides flexibility and better model fit for PRO measures. The PCM belongs to the Rasch family, so it estimates fewer parameters than the GRM as it assumes the discrimination parameter is equal across all items. With more parameters to be estimated, the GRM requires a larger sample to obtain stable parameter estimates compared with the PCM.

DIF

A formula may be used to mathematically describe the null hypothesis of DIF in a given item (Penfield & Camilli, 2007) by the expression of the probability distribution of Y conditional on θ by $f(Y|\theta)$:

$$\mathrm{f}(\mathrm{Y}|\theta, \mathrm{G} = \mathrm{R}) = \mathrm{f}(\mathrm{Y}|\theta, \mathrm{G} = \mathrm{F})$$

Where Y denotes the response category of a given item; G corresponds to the grouping variable, with two groups of subjects denoted as the reference group (R) and the focal group (F). Specifically, the conditional probability of Y is dependent only on θ, and if that error distributions for the two groups are different, then the item contains DIF. There are two types of DIF, uniform DIF and non-uniform DIF. Uniform DIF exists when the DIF is consistent across the entire range of the construct θ; in this case the item location parameters differ between the subgroups. Non-uniform DIF exists when the magnitude or direction of DIF differs across the range of the construct. DIF is defined in Chapter 1, as well as numerous DIF detection methods. This review focuses on IRT-based methods and methods that specifically apply to PROMIS measures.

DIF methods for PROMIS

IRT method

Lord (1980) defined DIF using IRT such that an item should have exactly the same IRF in both groups, and subjects at the same θ level should have exactly the same probability of endorsing a given response option regardless of how they are grouped. Lord's original Wald test was developed to compare the vectors of item discrimination and location parameters between groups. $\chi 2$ statistics can be used for testing the statistical significance.

$$\mathrm{d_i} = \frac{\hat{\mathrm{b}}_{\mathrm{F_i}} - \hat{\mathrm{b}}_{\mathrm{R_i}}}{\sqrt{\mathrm{Var}\left(\hat{\mathrm{b}}_{\mathrm{F_i}}\right) - \mathrm{Var}\left(\hat{\mathrm{b}}_{\mathrm{R_i}}\right)}}$$

Where $\hat{b}_{F_i}$ and $\hat{b}_{R_i}$ are the difficulty parameters estimated for the focal group and the reference group via maximum likelihood; the variances of the difficulty parameter estimates are $Var\left(\hat{b}_{F_i}\right)$ and $Var(\hat{b}_{R_i})$. This method has been shown to inflate the Type I error rates in several simulation studies (Donoghue & Isham, 1998; Kim & Cohen, 1994; McLaughlin & Drasgow, 1987). The Wald-2 (Langer, 2008) and Wald-1 (Cai et al., 2011) tests were developed to better control the standard error and allow for more accurate item parameter estimation.

The IRT-Likelihood Ratio Test (IRT-LR, Thissen, 1988) is popular and flexible method among IRT-based DIF detection methods. The IRT-LR method exams for the presence of DIF in the 2PL model and compares the difference between the two models with goodness of fit tests using -2 log likelihood values that are distributed as a $\chi 2$ statistic with df = 2. If the test statistic is statistically significant, subsequent tests compare the fit of the models to the two groups. This allows for model fit comparison assuming parameter estimate equality for the item in question across reference and focal groups (compact model), and when this equality constraint is relaxed and differences in the item parameters across groups are allowed (free or full model). The test statistic for comparing the two groups on all item parameters is:

$$\mathrm{LR} = \ -2\ \log \mathrm{L_c} - (-2 \log \mathrm{L_A})$$

where logLc is the log likelihood of the compact model (i.e., equality condition with no differences present), and $\log L_A$ is the log likelihood of the augmented model (i.e., more free conditions). By using the IRT-LR test, differences in test response functions can be constructed by summing the expected item scores to obtain an expected scale score and these can be graphically displayed.

Logistic regression method

Logistic regression DIF method has a standard expression as:

$$\log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 S + \beta_2 G$$

Where x is the item being examined for DIF and G is grouping variable, p is the probability of selecting x=1 for given S and G. Rasch-Welch t-test (Linacre, 2022) that is in the form of logistic regression was used in this study. The difficulty is estimated from the model by the item type in Rasch family, dichotomous, rating scale, partial credit, etc. During the estimation procedure, it re-estimates the overall item difficulty using logistic regression model. The log-odds value for the classification group was computed by the previously estimated Rasch effects with adding the DIF effect.

DIF treatment approaches

In the PROMIS analysis plan (Reeve et al., 2007), four different approaches for treating DIF items are described:

- Delete
- Ignore
- Multigroup
- Multidimensional modeling

The first approach is an extreme option that deletes or eliminates the DIF items from the PROMIS measure. Randall and Camilli (2006) suggested the treatment as removing the biased items. As stated in Chapter I, the deletion of one or more items from a measure is likely to violate the content validity of the measure, especially for symptoms and impact items in PRO measures.

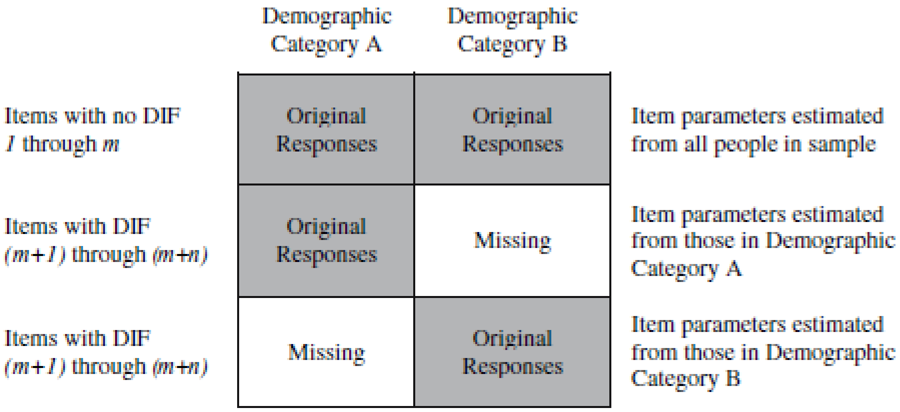
The second approach is not to remove the DIF item but to retain the item and ignore the detected DIF. In the PROMIS analysis plan (Reeve et al., 2007), it is suggesting that:

> if especially in key areas of the trait continuum that are sparsely populated by items, or if content experts determine that the items with DIF are central to the meaning of the construct, other options are to ignore DIF if it is small" (p. S29).

A study by Groenvold and colleagues (1995) found DIF between Caucasian and Japanese cancer patients' groups on the EORTC QLQ-30. Several studies (Johnson et al., 1998; Bjorner et at., 1998; Teresi et al., 2000) have shown that ethnicity commonly leads to DIF in HRQoL, but the conclusion from qualitative considerations was that DIF items should not be removed. Bjorner at al.(2004) examined the effect of removing language-related DIF items from the scoring algorithm and found that the resulting scores did not perform as well as those that ignored DIF. Pagano and Gotay (2005) examined the impact of removing items with DIF from the scales and similarly concluded that DIF items should not be removed.

The third approach is to control DIF by using demographic-specific item parameters. Crane (2006, 2007a, 2007b) used a demographic-specific design to treat DIF items detected in their studies (Figure 1) using logistic regression modeling methods. These studies concluded that a demographic-specific design was favored compared with removing DIF items.

Figure 3. Demographic-specific design (Crane, 2007)



	Demographic Category A	Demographic Category B	
Items with no DIF *1* through *m*	Original Responses	Original Responses	Item parameters estimated from all people in sample
Items with DIF *(m+1)* through *(m+n)*	Original Responses	Missing	Item parameters estimated from those in Demographic Category A
Items with DIF *(m+1)* through *(m+n)*	Missing	Original Responses	Item parameters estimated from those in Demographic Category B

Cho, Suh & Lee (2016) conducted a simulation study to compare the four approaches (Delete, Ignore, Multigroup, Multidimensional modeling) using the IRT-LR method and dichotomous items and found the multigroup and multidimensional modeling DIF treatments outperformed the deleting and ignoring treatments. Liu & Rogers (2021) also conducted a simulation study to compare the same four approaches using the 3PL model for dichotomous items with conclusion confirming Cho et al.'s results.

CHAPTER III: METHOD

Context of Study

The research questions in this study are addressed using real world data (FDA, 2017; EMA, 2016) collected in an observational, non-interventional study that contributes to our understanding of patients' health status and health care routine. By analyzing real world data from a diverse, large, and unrestricted patient population, evidence, and the interpretation of patterns of a wide range of health outcomes are described (Blonde, Khunti, et al., 2018). The dataset proposed for use in this study is from MY-Health (PROMIS 2 MY HEALTH, Potosky & Moinpour, 2016), a cross sectional study in a diverse US population-based sample of more than 5500 adult cancer patients from 4 registries in three states (California, Louisiana, and New Jersey). The aim of the data collection, was to create a large diverse community-based database that would provide the opportunity for evaluating the measurement properties of several PROMIS domains using item response theory, including the detection of bias across various demographic characteristics (e.g., race and/or ethnic group).

The patients were recruited from SEER cancer registries (US Department of Health and Human Services, 2013). Demographic variables, including gender, ethnicity, race, date of cancer diagnosis, cancer type, and stage, are found in the SEER coding manual (Adamo et al., 2013). To be eligible to participate in the MY-Health study, patients had to meet the following inclusion criteria:

1. Age between 21–84 years old at diagnosis.
2. Diagnosed with one of these seven cancers: female breast cancer, uterine and cervical cancers, prostate cancer, colorectal cancer, non-small cell lung cancer (NSCLC) and non-Hodgkin's Lymphoma.

3. Within 6–13 months of diagnosis at the time of recruitment.
4. Able to read in English, Spanish, or Mandarin.

Additional demographic variables collected in the MY-Health survey include various treatment experiences (e.g., chemotherapy, radiation therapy, or hormonal therapy; surgery); comorbid conditions (numbers and types); socioeconomic characteristics (education level; employment status; annual income; marital status; insurance coverage; and whether the patient was born in the US). These demographic variables provide a wide range of selection to sufficiently support the psychometric evaluation.

The MY-Health survey data were collected by a paper-based survey sent by US mail; non-responders were telephoned to follow up. The study included two timepoints, baseline (n = 5506) and 6-month follow-up (n = 2877).

The original study and data collection were sponsored by the National Institutes of Health, specifically, the National Institute of Arthritis and Musculoskeletal and Skin Diseases (Grant# 1U01AR057971).

PRO Measures

In this study, PROMIS item banks were the primary PRO measures. All patients completed 9 PROMIS domains: physical function, pain interference, fatigue, depression, anxiety, sleep Distribution, applied Cognition – general concerns, social support, and ability to participate in social roles and activities that are important to patient's quality of life and well-being. All the items use ordered response scales (Table 1). The Physical Function domain (16 items) represents the key to understand patient's overall well-being for this cancer population. Pain is the most common symptom experienced by cancer patients and Fatigue is also commonly reported by cancer patients, and in addition, cancer-related fatigue occurs often after treatment such as surgery or chemotherapy and has been reported as a major factor that influences a patient's quality of life. The Pain Interference domain consists of 10 items and the Fatigue domain consists of 14 items that assess the severity, frequency, and impact of fatigue. Cancer patients often have psychological disorders such as depression or anxiety—the Anxiety domain has 11 items measuring feelings such as worry, and fear and the Depression domain has 10 items measuring the patient's mood. The Social Roles and Activities domain includes 9 items and the Social Support domains includes 4 items and the Applied Cognition domain includes 8 items that measure cancer-related cognitive changes and impairment.

Table 1. PROMIS Domain and Scale Summary

PROMIS Domain	Number of items	Scale				
Physical Function	16	5 = Not at all		5 = Without any difficulty		
		4 = Very little		4 = With a little difficulty		
		3 = Somewhat	or	3 = With some difficulty		
		2 = Quite a lot		2 = With much difficulty		
		1 = Cannot do		1 = Unable to do		
Pain Interference	10	1= Never				
		2 = Rarely				
		3= Sometimes				
		4 = Often				
		5 = Always				
Fatigue	14	1 = Not at all		1 = Never		
		2 = A little bit		2 = Rarely		
		3 = Somewhat		3 = Sometimes		
		4 = Quite a bit		4 = Often		
		5 = Very much		5 = Always		
Sleep Disturbance	10	1 = Never		5 = Very poor		1 = Not at all
		2 = Rarely		4 = Poor		2 = A little bit
		3 = Sometimes	or	3 = Fair	or	3 = Somewhat
		4 = Often		2 = Good		4 = Quite a bit
		5 = Always		1 = Very good		5 = Very much
Anxiety	11	1 = Never				
		2 = Rarely				
		3 = Sometimes				
		4 = Often				
		5 = Always				
Depression	10	1 = Never				
		2 = Rarely				
		3 = Sometimes				
		4 = Often				
		5 = Always				
Social Roles and Activities	9	5 = Never				
		4 = Rarely				
		3 = Sometimes				
		2 = Often				
		1 = Always				
Social Support	4	1 = Never				
		2 = Rarely				
		3 = Sometimes				
		4 = Often				
		5 = Always				

PROMIS Domain	Number of items	Scale
Applied Cognition	8	0 = Never 1 = Rarely (Once) 2 = Sometimes (Two or three times) 3 = Often (About once a day) 4 = Very often (Several times a day)

PROMIS = Patient-Reported Outcomes Measurement Information System.

PROMIS Fatigue Measure

The PROMIS Fatigue measure was used in this study for illustrate the methodology. MY-HEALTH study (Jensen, et al., 2016) administered 14 items that selected from the 95-item PROMIS Fatigue item bank to build a custom short form. It emphasized to include as many items as possible for validation analysis, with selecting items based on their inclusion on short forms (as of 2010) or their high frequency of selection administered online in computerized adaptive testing (CAT) through assessment center. The summary of the item source is listed in Table 2.

Among the 14 PROMIS Fatigue items, item 1 to 8 measured the frequency of fatigue using 5-point ordinal response categories (1 = Not at all, 2 = A little bit ,3 = Somewhat, 4 = Quite a bit, 5 = Very much), item 9 to 14 measured the severity of fatigue using 5-point ordinal response categories (1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Always).

Table 2. PROMIS Fatigue Items Source

		PROMIS Fatigue Short Forms						
PROMIS Item Code	Question	PROMIS Fatigue item bank*	4a	6a	7a	8a	10a	13a
FATEXP20	1. How often did you feel tired?	X			X			
FATEXP5	2. How often did you experience extreme exhaustion?	X			X			
FATEXP18	3. How often did you run out of energy?	X			X			
FATIMP33	4. How often did your fatigue limit you at work (include work at home)?	X			X			
FATIMP30	5. How often were you too tired to think clearly?	X			X			
FATIMP21	6. How often were you too tired to take a bath or shower?	X			X			
FATIMP40	7. How often did you have enough energy to exercise strenuously?	X			X			
FATIMP3	8. How often did you have to push yourself to get things done because of your fatigue?	X				X		
FATEXP41	9. How run-down did you feel on average?	X	X	X		X		
HI7	10. I feel fatigued	X	X	X		X	X	X
FATEXP40	11. How fatigued were you on average?	X	X	X		X		
An3	12. I have trouble starting things because I am tired.	X	X	X		X	X	X
FATEXP35	13. How much were you bothered by your fatigue on average?	X		X		X		
FATIMP49	14. To what degree did your fatigue interfere with your physical functioning?	X		X				

PROMIS = Patient-Reported Outcomes Measurement Information System.
* PROMIS Fatigue item bank has 95 items in total.

Software for analysis

The IRT analyses were conducted using two R packages, Mirt (Chalmers, 2012) and PerFit (Tendeiro, 2016), which provided comprehensive facilities for the application of IRT to obtain item and person parameter estimates as well as DIF analyses for GRM. Winsteps V5.2.2 (Linacre, 2022) was used for analysis using PCM. Mplus (Muthén & Muthén, 2021) were used to conduct the confirmatory factor analysis. Data manipulation, descriptive statistics, and final outputs were programmed in SAS 9.4 University Edition (SAS institute Inc., 2015).

Data Description

As real world data were analyzed in this study, it is important to comprehensively describe the large diverse population and the health status by computing descriptive statistics for PROMIS measures and patient characteristics. Descriptive statistics for the PROMIS measures used in this study were computed and reported.

Patient Characteristics and Clinical Variables

Descriptive statistics describing demographic and patient characteristics, as well as select clinical variables, were tabulated for the overall population and by selected anchors at baseline to describe the patient sample, for example, age, sex, time since cancer diagnosis, cancer type, and stage.

PROMIS Fatigue Measure

Descriptive statistics of the selected PROMIS measures were computed and reported, including the mean, standard deviation (SD), median, percentiles (25th, 75th), and number of missing responses. The percentage of patients with the lowest and highest possible scores were reported to evaluate floor and ceiling effects, defined as when the percentage responding in an extreme response category is more than twice the expected probability given a uniform distribution. For a 5-point ordinal response scale, 20% is the expected percentage. Item-level response distributions (frequencies and percentages) were presented for each response category, including the frequency of missing responses.

Study Design

DIF Analyses and Anchor Selection

Before conducting DIF analysis using IRT and logistic regression method, it is important to first identify anchor items to serve as ‘bridge’ or equate the underlying scale in the two groups. With using the chosen anchors, the items on a measure can be tested for DIF.

Although expert review should be conducted prior to anchor selection, however, this option is not available in this study. The aim of this section in this study is limited to detect a DIF items, not for exploring the potential source of the bias that are meaningful for future study and beneficial for the patients. The validation study for the PROMIS physical function measure by Jensen (2015) provided insight of this dataset and expert review:

Over half of cancer survivors are likely to experience significant physical limitations. Decline in physical function is often associated with a cancer diagnosis and the ensuing initial treatment , and such decline can have long-lasting effects extending past treatment and is associated with lower quality of life and increased risk of mortality. (P. 2334)

In realistic, the anchor selection, in addition to the large amount of qualitative work, it is essential to explore the data before picking and using the anchor item in blind, for instance, items with high ceiling or floor effect that is often seen in PRO data will not provide accurate estimation for item and person parameter, and the sample size is also a factor to be considered.

DIF source has been lightly discussed in chapter I, Teresa (2008) conducted a comprehensive review on DIF source in PRO. Based on the variables in this study, Table 2 listed the planned analysis and anchors. For the subgroup indicated as “To be decided” (TBD), an example illustrated by Jensen (2015) used the same dataset, the PROMIS PF measure reported in

T-score was 40.2 for Stage I and II cancer patients, compared to stage III and IV cancer patient at T-score of 37.5. So, an example of the selection of anchor could be:

Focus group: Stage III/IV Lung cancer

Reference group: stage I/IV Lung cancer

IRT Wald test and Welch test were used for DIF detection. The interest for this step is only for detection to flag the DIF items in the study dataset to create the condition for evaluating calibration approaches. For each PROMIS measures, patients missed all the items will be to the excluded from analyses. The administered PROMIS measures varies in scales through all used 5-point ordinal scales, for analyses purpose, they will be rescaled as 1 to 5 with higher score indicated better status.

Table 3. Anchor selection example

Conditional grouping Variable	Focus and reference group
Sex	Female and Male
Age	21-65 years old and >66 years
Race	White and all others
Ethnicity	Hispanic and Not Hispanic
Cancer stage	Stage I & II and stage III & IV
Language	English only vs Others
Married	Yes and No
Work	Yes and No
Education	College and Graduate School and Lower than College
Income	$60,000 or higher and $59,999 and below
Sex	Male and Female

Calibration and Ability Estimation

The PROMIS Fatigue measure used in the MY-Health item bank were calibrated as unidimensional structure to obtain the item parameter estimates and person parameter estimates using the GRM and PCM using Maximum-likelihood estimation.

The calibration was performed under three conditions: a) using the original dataset without any modification; b) dropping the DIF item that results from the previous analysis, such that DIF items was shown as "missing"; c) creating a new dataset with new variables constructed such that a DIF item was represented as two new items based on the demographic variable used as the anchor in the previous step (Table 3). The data under the three conditions were calibrated using two polytomous models, assuming and requiring less than three PROMIS Fatigue items detected with DIF in the previous section. The quality of item parameter estimation and theta estimation was compared and evaluated using the criteria described in the next section.

Evaluation Criteria

Item level

Item and person parameters were estimated under the chosen model. It is a common strategy to evaluate the item fit and person (ability) fit. When a parametric model is fit to data, the ability estimation is more accurate and so item-level fit is usually assessed to guide item revision/deletion. Person (ability) fit is assessed to "detect item-score patterns that are improbable given an IRT model or given the other patterns in a sample" (Meijer & Sitsma, 2001). The standard errors (SEs) associated with item and person parameters are estimated, with lower values indicating better fit. Because the value of the SE is sensitive to sample size, Tay et al., 2014 offer the following criteria based on experience:

- Larger than 0.50 is poor

- Between 0.50 and 0.35 is moderate
- Between 0.20 and 0.35 is good
- Less than 0.20 is excellent

This χ^2 statistics (Bock, 1972) and RMSEA calculated from χ^2 were used to examine item-level goodness of fit. With polytomous response items of the null hypothesis is the observed responses across categories for a single item are not significantly different (p<0.0001) from the modeled responses. The model is based on number-correct scores rather than trait scores, and the standard errors in item parameter estimates are account.

$$x^2 = \sum_{k=1} \mathrm{N_k} \frac{(\mathrm{O_{ik}} - \mathrm{E_{ik}})^2}{\mathrm{E_{ik}}(1 - \mathrm{E_{ik}})}$$

Where n is the number of patients answering item *i* with *k* options, O_{ik} and E_{ik} are the proportion of observed and expected proportion. Under the hypothesis of perfect model fit, the χ^2 statistics are approximately distributed as S-χ^2 (Orlando and Thissen's, 2000, 2003; Kang and Chen, 2008) values with the tabulated degrees of freedom; significant values indicate lack of fit.

Person level

Person fit is evaluated by the lz statistic. The expression of lz is generalized to categorical data (Drasgow et al., 1985):

$$lz = p(Y_i|\theta_i) = \sum_{i=1}^{n} \sum_{j=1}^{A+1} \delta_j(v_i) \log Pg(\theta)$$

where Y is the response to item i, θ is the person score estimates, n is the number of examinees, δj(vi), is the random vector of item choices, forcing only to sum the probabilities of the endorsed responses. In practice, a standardized version of lz statistics is used for person fit.

Test level

M_2 statistics (Maydeu-Olivares & Joe, 2005)) was used for assessing the model-data fit. It is based on two-way marginal tables to test the observed response from the modeled response , with good fit expected to be close to zero and not significant is defined as less than 0.05 (Tay et al., 2014). The M_2 statistic is in the form (Xu et al., 2017) of:

$$\mathrm{M}_2 = \mathrm{N}\hat{\mathrm{e}}_2'\hat{\mathrm{C}}_2\hat{\mathrm{e}}_2$$

$$\mathrm{RMSEA} = \sqrt{\mathrm{MAX}\left(\frac{\mathrm{M}_2 - \mathrm{df}}{\mathrm{N} \times \mathrm{df}}, 0\right)}$$

Where N is sample size, e_2 is the first- and second-order residual proportion estimated, $\hat{C}$ is an asymptotic covariance matrix for the first- and second-order residual proportions estimated. More details are presented by Maydeu-Olivares & Joe (2005, 2006).

The Akaike information criterion (AIC; Akaike, 1974), and Bayesian information criterion (BIC; Schwartz, 1978) were used to compare the three calibration approaches. Both the AIC and BIC are two widely used model comparison statistics and are found to be accurate for model comparisons based on polytomous items (Kang, et al., 2009). The absolute AIC or BIC value for the compared models are not used alone but are used together with the interpretation that lower values are preferred and indicate better fit. Using the Expectation-Maximization (EM) algorithm (Bock & Aitken, 1981) to compute the AIC and BIC:

$$AIC = 2k - 2\ln(L)$$

$$BIC = k\ln(N) - 2\ln(L)$$

Where k is equal to the number of parameters in the model, ln is the natural log and L is the maximum likelihood estimate. The last portion of both the AIC and BIC formulas are the same; the first element of the BIC multiplies the number of parameters by the natural log of the sample size (N). Thus, the BIC is more sensitive to sample size; compared with the AIC, and the penalty for additional parameters is greater for the BIC.

CHAPTER IV: RESULT

Patient Characteristic and Demographics

Table 4 presents patient characteristic and demographics for the subsample. Of the 1808 patients from the subgroup with lung, cervical, or colorectal cancer over the overall sample of 5506 patients, 1005 (55.6%) were female and 803 (44.4%) were male; by age group, 1004 (55.5%) were from 21 to 65 years old, 804 (44.5%) were older than 66 years old; the majority of race of this sample were White (1064, 58.8%), Black (366, 20.2%), Asian (294, 13.8%); 1514 (83.7%) of the patients were not Hispanic. From patients' medical record, 843 (46.6%) of the patients were diagnosed as Stage I and II, 882 (48.8%) of the patients were diagnosed as Stage III and IV. 1667 (92.2%) of the patients used English to answer this survey, 576 (32.6%) of the patients were married, 1190 (67.4%) of the patients had employment status as "Not working". For patients' education background, only 379 (21.2%) patients had college degree or graduate degree. 1041 (60.2%) patients earned less than $59,999 and 448 (25.8%) earned more than $60,000.

Table 4. Patient Characteristic and Demographics

Characteristic	Overall (N = 1808)	Cervix Cancer (N = 149)	Colorectal Cancer (N = 937)	Lung Cancer (N = 722)
Sex				
Male	803 (44.4)	0 (0.0)	455 (48.6)	348 (48.2)
Female	1005 (55.6)	149 (100.0)	482 (51.4)	374 (51.8)
Age group				
21 - 65 years old	1004 (55.5)	132 (88.6)	534 (57.0)	338 (46.8)
> 66 years old	804 (44.5)	17 (11.4)	403 (43.0)	384 (53.2)
Race				
Multiple	56 (3.1)	8 (5.4)	34 (3.6)	14 (1.9)
White	1064 (58.8)	85 (57.0)	475 (50.7)	504 (69.8)
Black	366 (20.2)	23 (15.4)	212 (22.6)	131 (18.1)
Asian	249 (13.8)	21 (14.1)	174 (18.6)	54 (7.5)
American Indian or Alaska Native	15 (0.8)	4 (2.7)	8 (0.9)	3 (0.4)
Asian Hawaiian or Pacific Islander	9 (0.5)	2 (1.3)	4 (0.4)	3 (0.4)
Other	49 (2.7)	6 (4.0)	30 (3.2)	13 (1.8)
Ethnicity				
Hispanic	294 (16.3)	51 (34.2)	181 (19.3)	62 (8.6)
Not Hispanic	1514 (83.7)	98 (65.8)	756 (80.7)	660 (91.4)
Cancer type				
Breast	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Cervix	149 (8.2)	149 (100.0)	0 (0.0)	0 (0.0)
Colorectal	937 (51.8)	0 (0.0)	937 (100.0)	0 (0.0)
Lung	722 (39.9)	0 (0.0)	0 (0.0)	722 (100.0)
Non Hodgkin Lymphoma	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Prostate	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Uterus	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Cancer stage				
In situ	22 (1.2)	0 (0.0)	22 (2.3)	0 (0.0)
Stage I	494 (27.3)	80 (53.7)	206 (22.0)	208 (28.8)
Stage II	349 (19.3)	13 (8.7)	236 (25.2)	100 (13.9)
Stage III	511 (28.3)	36 (24.2)	292 (31.2)	183 (25.3)

Characteristic	Overall (N = 1808)	Cervix Cancer (N = 149)	Colorectal Cancer (N = 937)	Lung Cancer (N = 722)
Stage IV	371 (20.5)	12 (8.1)	156 (16.6)	203 (28.1)
Not applicable	4 (0.2)	3 (2.0)	0 (0.0)	1 (0.1)
Stage Occult	7 (0.4)	0 (0.0)	0 (0.0)	7 (1.0)
Stage Unknown	50 (2.8)	5 (3.4)	25 (2.7)	20 (2.8)
Survey language				
English	1667 (92.2)	125 (83.9)	837 (89.3)	705 (97.6)
Spanish	92 (5.1)	23 (15.4)	56 (6.0)	13 (1.8)
Chinese	49 (2.7)	1 (0.7)	44 (4.7)	4 (0.6)
Married				
No	811 (45.3)	81 (55.1)	418 (45.2)	312 (43.5)
Yes	978 (54.7)	66 (44.9)	506 (54.8)	406 (56.5)
Employment				
Working	576 (32.6)	79 (53.7)	338 (37.0)	159 (22.5)
Not working	1190 (67.4)	68 (46.3)	575 (63.0)	547 (77.5)
Education				
Less than High School	401 (22.4)	39 (26.4)	209 (22.6)	153 (21.4)
High School	417 (23.3)	26 (17.6)	199 (21.5)	192 (26.9)
Some College	583 (32.6)	50 (33.8)	291 (31.5)	242 (33.9)
College Degree	233 (13.0)	20 (13.5)	130 (14.1)	83 (11.6)
Graduate Degree	146 (8.2)	11 (7.4)	92 (9.9)	43 (6.0)
Don’t know	7 (0.4)	2 (1.4)	4 (0.4)	1 (0.1)
Insurance				
Private insurance	582 (32.6)	64 (43.8)	320 (34.7)	198 (27.7)
Government insurance	628 (35.2)	57 (39.0)	318 (34.5)	253 (35.4)
Private + Government insurance	492 (27.6)	9 (6.2)	236 (25.6)	247 (34.5)
No Insurance	46 (2.6)	5 (3.4)	30 (3.3)	11 (1.5)
Don't know or Unsure	36 (2.0)	11 (7.5)	19 (2.1)	6 (0.8)
Income				
Less than $10,000	234 (13.5)	27 (19.0)	134 (15.0)	73 (10.5)
$10,000 to $59,999	809 (46.7)	61 (43.0)	394 (44.1)	354 (50.7)
$60,000 to $99,999	264 (15.2)	18 (12.7)	137 (15.3)	109 (15.6)
$100,000 to $199,999	146 (8.4)	6 (4.2)	89 (10.0)	51 (7.3)
$200,000 or more	38 (2.2)	7 (4.9)	17 (1.9)	14 (2.0)

Characteristic	Overall (N = 1808)	Cervix Cancer (N = 149)	Colorectal Cancer (N = 937)	Lung Cancer (N = 722)
Don’t know/Unsure	134 (7.7)	18 (12.7)	68 (7.6)	48 (6.9)
Refuse to answer	109 (6.3)	5 (3.5)	55 (6.2)	49 (7.0)
Did you ever have any surgery as part of your cancer treatment?				
Yes	1240 (68.6)	96 (64.4)	800 (85.4)	344 (47.6)
No	537 (29.7)	50 (33.6)	117 (12.5)	370 (51.2)
Don't Know	9 (0.5)	1 (0.7)	7 (0.7)	1 (0.1)

PROMIS = Patient-Reported Outcomes Measurement Information System.

Item Level Descriptive Statistics and Response Distribution

Standard descriptive statistics were computed at baseline for each PROMIS fatigue item to characterize the extent to which patients experience in frequency and severity associated with cancer fatigue (Table 5).

The item means, or difficulties, ranged on the 1 to 5 response scale from mean of 2.1 (SD = 1.16) for the least difficult item, Item 6 (How often were you too tired to take a bath or shower?), to mean of 3.3 (SD = 1.04) for the most difficult item, Item 1 (How often did you feel tired?).

The full range of the response categories were endorsed. For frequency items (Item 1 to Item 8) “Sometimes” was more frequently endorsed, except for Item 5 (How often were you too tired to think clearly?) and Item 6 (How often were you too tired to take a bath or shower?) with 35.2% and 43.3% responses in “Never”, respectively. For severity questions (Item 9 to 14), the frequently endorses were “A little bit”, “Somewhat”, “ Quite a bit”.

Table 5. Response Frequency Table for PROIS Fatigue Items

PROMIS Fatigue Item	Overall	Cervix Cancer	Colorectal Cancer	Lung Cancer
1. How often did you feel tired?				
Mean (SD), n	3.3 (1.04), 1802	3.3 (1.21), 149	3.2 (1.04), 933	3.5 (0.98), 720
Median (Min, Max)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	4.0 (1.0, 5.0)
Floor/Ceiling (%)	6.3/12.5	8.7/19.5	8.1/9.3	3.3/15.3
Missing (%)	6 (0.3)	0 (0.0)	4 (0.4)	2 (0.3)
Frequency (%)				
1 = Never	113 (6.3)	13 (8.7)	76 (8.1)	24 (3.3)
2 = Rarely	222 (12.3)	25 (16.8)	124 (13.3)	73 (10.1)
3 = Sometimes	683 (37.9)	42 (28.2)	384 (41.2)	257 (35.7)
4 = Often	558 (31.0)	40 (26.8)	262 (28.1)	256 (35.6)
5 = Always	226 (12.5)	29 (19.5)	87 (9.3)	110 (15.3)
2. How often did you experience extreme exhaustion?				
Mean (SD), n	2.5 (1.23), 1791	2.7 (1.42), 146	2.3 (1.18), 929	2.7 (1.24), 716
Median (Min, Max)	2.0 (1.0, 5.0)	3.0 (1.0, 5.0)	2.0 (1.0, 5.0)	3.0 (1.0, 5.0)
Floor/Ceiling (%)	28.1/5.9	30.8/13.0	31.9/4.1	22.6/6.7
Missing (%)	17 (0.9)	3 (2.0)	8 (0.9)	6 (0.8)
Frequency (%)				
1 = Never	503 (28.1)	45 (30.8)	296 (31.9)	162 (22.6)
2 = Rarely	405 (22.6)	19 (13.0)	228 (24.5)	158 (22.1)
3 = Sometimes	455 (25.4)	35 (24.0)	236 (25.4)	184 (25.7)
4 = Often	323 (18.0)	28 (19.2)	131 (14.1)	164 (22.9)
5 = Always	105 (5.9)	19 (13.0)	38 (4.1)	48 (6.7)
3. How often did you run out of energy?				
Mean (SD), n	2.9 (1.16), 1800	2.9 (1.34), 149	2.7 (1.12), 931	3.1 (1.12), 720
Median (Min, Max)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)
Floor/Ceiling (%)	14.8/7.2	20.8/12.8	17.4/4.4	10.1/9.6
Missing (%)	8 (0.4)	0 (0.0)	6 (0.6)	2 (0.3)
Frequency (%)				
1 = Never	266 (14.8)	31 (20.8)	162 (17.4)	73 (10.1)
2 = Rarely	397 (22.1)	30 (20.1)	242 (26.0)	125 (17.4)

PROMIS Fatigue Item	Overall	Cervix Cancer	Colorectal Cancer	Lung Cancer
3 = Sometimes	550 (30.6)	31 (20.8)	290 (31.1)	229 (31.8)
4 = Often	458 (25.4)	38 (25.5)	196 (21.1)	224 (31.1)
5 = Always	129 (7.2)	19 (12.8)	41 (4.4)	69 (9.6)
4. How often did your fatigue limit you at work (include work at home)?				
Mean (SD), n	2.9 (1.28), 1774	2.8 (1.47), 149	2.7 (1.24), 923	3.1 (1.26), 702
Median (Min, Max)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)
Floor/Ceiling (%)	20.5/11.0	29.5/16.8	22.9/8.2	15.5/13.4
Missing (%)	34 (1.9)	0 (0.0)	14 (1.5)	20 (2.8)
Frequency (%)				
1 = Never	364 (20.5)	44 (29.5)	211 (22.9)	109 (15.5)
2 = Rarely	314 (17.7)	18 (12.1)	191 (20.7)	105 (15.0)
3 = Sometimes	496 (28.0)	31 (20.8)	268 (29.0)	197 (28.1)
4 = Often	405 (22.8)	31 (20.8)	177 (19.2)	197 (28.1)
5 = Always	195 (11.0)	25 (16.8)	76 (8.2)	94 (13.4)
5. How often were you too tired to think clearly?				
Mean (SD), n	2.2 (1.16), 1802	2.4 (1.29), 149	2.1 (1.12), 933	2.3 (1.16), 720
Median (Min, Max)	2.0 (1.0, 5.0)	2.0 (1.0, 5.0)	2.0 (1.0, 5.0)	2.0 (1.0, 5.0)
Floor/Ceiling (%)	35.2/3.6	34.9/5.4	38.5/2.9	31.0/4.0
Missing (%)	6 (0.3)	0 (0.0)	4 (0.4)	2 (0.3)
Frequency (%)				
1 = Never	634 (35.2)	52 (34.9)	359 (38.5)	223 (31.0)
2 = Rarely	466 (25.9)	23 (15.4)	238 (25.5)	205 (28.5)
3 = Sometimes	415 (23.0)	37 (24.8)	214 (22.9)	164 (22.8)
4 = Often	223 (12.4)	29 (19.5)	95 (10.2)	99 (13.8)
5 = Always	64 (3.6)	8 (5.4)	27 (2.9)	29 (4.0)
6. How often were you too tired to take a bath or shower?				
Mean (SD), n	2.1 (1.16), 1799	2.2 (1.34), 148	1.9 (1.08), 933	2.2 (1.19), 718
Median (Min, Max)	2.0 (1.0, 5.0)	2.0 (1.0, 5.0)	2.0 (1.0, 5.0)	2.0 (1.0, 5.0)
Floor/Ceiling (%)	43.3/3.4	44.6/8.1	47.8/2.3	37.2/4.0
Missing (%)	9 (0.5)	1 (0.7)	4 (0.4)	4 (0.6)
Frequency (%)				
1 = Never	779 (43.3)	66 (44.6)	446 (47.8)	267 (37.2)
2 = Rarely	379 (21.1)	22 (14.9)	200 (21.4)	157 (21.9)

PROMIS Fatigue Item	Overall	Cervix Cancer	Colorectal Cancer	Lung Cancer
3 = Sometimes	406 (22.6)	31 (20.9)	200 (21.4)	175 (24.4)
4 = Often	173 (9.6)	17 (11.5)	66 (7.1)	90 (12.5)
5 = Always	62 (3.4)	12 (8.1)	21 (2.3)	29 (4.0)
7. How often did you have enough energy to exercise strenuously?				
Mean (SD), n	3.5 (1.34), 1778	3.6 (1.38), 149	3.5 (1.28), 917	3.5 (1.41), 712
Median (Min, Max)	4.0 (1.0, 5.0)	4.0 (1.0, 5.0)	4.0 (1.0, 5.0)	4.0 (1.0, 5.0)
Floor/Ceiling (%)	10.2/31.4	10.1/36.9	8.0/28.9	13.1/33.6
Missing (%)	30 (1.7)	0 (0.0)	20 (2.1)	10 (1.4)
Frequency (%)				
1 = Never	181 (10.2)	15 (10.1)	73 (8.0)	93 (13.1)
2 = Rarely	269 (15.1)	23 (15.4)	149 (16.2)	97 (13.6)
3 = Sometimes	385 (21.7)	28 (18.8)	221 (24.1)	136 (19.1)
4 = Often	384 (21.6)	28 (18.8)	209 (22.8)	147 (20.6)
5 = Always	559 (31.4)	55 (36.9)	265 (28.9)	239 (33.6)
8. How often did you have to push yourself to get things done because of your fatigue?				
Mean (SD), n	2.9 (1.26), 1794	2.8 (1.44), 149	2.7 (1.21), 927	3.1 (1.25), 718
Median (Min, Max)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)
Floor/Ceiling (%)	19.1/10.5	25.5/17.4	21.9/6.9	14.1/13.8
Missing (%)	14 (0.8)	0 (0.0)	10 (1.1)	4 (0.6)
Frequency (%)				
1 = Never	342 (19.1)	38 (25.5)	203 (21.9)	101 (14.1)
2 = Rarely	351 (19.6)	29 (19.5)	202 (21.8)	120 (16.7)
3 = Sometimes	507 (28.3)	29 (19.5)	280 (30.2)	198 (27.6)
4 = Often	405 (22.6)	27 (18.1)	178 (19.2)	200 (27.9)
5 = Always	189 (10.5)	26 (17.4)	64 (6.9)	99 (13.8)
9. How run-down did you feel on average?				
Mean (SD), n	2.8 (1.20), 1794	2.8 (1.38), 149	2.6 (1.14), 928	2.9 (1.22), 717
Median (Min, Max)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	2.0 (1.0, 5.0)	3.0 (1.0, 5.0)
Floor/Ceiling (%)	17.2/7.4	22.1/12.8	18.4/4.4	14.6/10.0
Missing (%)	14 (0.8)	0 (0.0)	9 (1.0)	5 (0.7)
Frequency (%)				
1 = Not at all	309 (17.2)	33 (22.1)	171 (18.4)	105 (14.6)

PROMIS Fatigue Item	Overall	Cervix Cancer	Colorectal Cancer	Lung Cancer
2 = A little bit	504 (28.1)	39 (26.2)	300 (32.3)	165 (23.0)
3 = Somewhat	425 (23.7)	19 (12.8)	221 (23.8)	185 (25.8)
4 = Quite a bit	424 (23.6)	39 (26.2)	195 (21.0)	190 (26.5)
5 = Very much	132 (7.4)	19 (12.8)	41 (4.4)	72 (10.0)
10. I feel fatigued				
Mean (SD), n	2.8 (1.23), 1796	2.8 (1.36), 149	2.6 (1.17), 929	3.0 (1.25), 718
Median (Min, Max)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	2.0 (1.0, 5.0)	3.0 (1.0, 5.0)
Floor/Ceiling (%)	17.7/8.5	21.5/11.4	19.7/5.6	14.3/11.7
Missing (%)	12 (0.7)	0 (0.0)	8 (0.9)	4 (0.6)
Frequency (%)				
1 = Not at all	318 (17.7)	32 (21.5)	183 (19.7)	103 (14.3)
2 = A little bit	504 (28.1)	38 (25.5)	287 (30.9)	179 (24.9)
3 = Somewhat	385 (21.4)	17 (11.4)	216 (23.3)	152 (21.2)
4 = Quite a bit	436 (24.3)	45 (30.2)	191 (20.6)	200 (27.9)
5 = Very much	153 (8.5)	17 (11.4)	52 (5.6)	84 (11.7)
11. How fatigued were you on average?				
Mean (SD), n	2.8 (1.20), 1793	2.9 (1.39), 148	2.7 (1.17), 930	3.0 (1.19), 715
Median (Min, Max)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)
Floor/Ceiling (%)	16.3/7.9	21.6/14.9	18.0/5.8	13.0/9.1
Missing (%)	15 (0.8)	1 (0.7)	7 (0.7)	7 (1.0)
Frequency (%)				
1 = Not at all	292 (16.3)	32 (21.6)	167 (18.0)	93 (13.0)
2 = A little bit	498 (27.8)	37 (25.0)	289 (31.1)	172 (24.1)
3 = Somewhat	427 (23.8)	22 (14.9)	220 (23.7)	185 (25.9)
4 = Quite a bit	435 (24.3)	35 (23.6)	200 (21.5)	200 (28.0)
5 = Very much	141 (7.9)	22 (14.9)	54 (5.8)	65 (9.1)
12. I have trouble starting things because I am tired.				
Mean (SD), n	2.6 (1.30), 1797	2.6 (1.49), 149	2.4 (1.23), 931	2.8 (1.32), 717
Median (Min, Max)	2.0 (1.0, 5.0)	2.0 (1.0, 5.0)	2.0 (1.0, 5.0)	3.0 (1.0, 5.0)
Floor/Ceiling (%)	27.9/8.2	34.2/15.4	30.5/5.4	23.2/10.3
Missing (%)	11 (0.6)	0 (0.0)	6 (0.6)	5 (0.7)
Frequency (%)				
1 = Not at all	501 (27.9)	51 (34.2)	284 (30.5)	166 (23.2)
2 = A little bit	432 (24.0)	26 (17.4)	257 (27.6)	149 (20.8)

PROMIS Fatigue Item	Overall	Cervix Cancer	Colorectal Cancer	Lung Cancer
3 = Somewhat	360 (20.0)	21 (14.1)	184 (19.8)	155 (21.6)
4 = Quite a bit	357 (19.9)	28 (18.8)	156 (16.8)	173 (24.1)
5 = Very much	147 (8.2)	23 (15.4)	50 (5.4)	74 (10.3)
13. How much were you bothered by your fatigue on average?				
Mean (SD), n	2.7 (1.28), 1783	2.7 (1.45), 148	2.6 (1.23), 925	2.9 (1.27), 710
Median (Min, Max)	3.0 (1.0, 5.0)	3.0 (1.0, 5.0)	2.0 (1.0, 5.0)	3.0 (1.0, 5.0)
Floor/Ceiling (%)	21.0/9.0	28.4/14.2	23.5/6.4	16.2/11.4
Missing (%)	25 (1.4)	1 (0.7)	12 (1.3)	12 (1.7)
Frequency (%)				
1 = Not at all	374 (21.0)	42 (28.4)	217 (23.5)	115 (16.2)
2 = A little bit	481 (27.0)	31 (20.9)	281 (30.4)	169 (23.8)
3 = Somewhat	351 (19.7)	19 (12.8)	179 (19.4)	153 (21.5)
4 = Quite a bit	416 (23.3)	35 (23.6)	189 (20.4)	192 (27.0)
5 = Very much	161 (9.0)	21 (14.2)	59 (6.4)	81 (11.4)
14. To what degree did your fatigue interfere with your physical functioning?				
Mean (SD), n	2.7 (1.30), 1778	2.6 (1.42), 148	2.5 (1.24), 925	2.9 (1.32), 705
Median (Min, Max)	3.0 (1.0, 5.0)	2.0 (1.0, 5.0)	2.0 (1.0, 5.0)	3.0 (1.0, 5.0)
Floor/Ceiling (%)	23.7/9.5	30.4/12.8	25.8/6.7	19.6/12.5
Missing (%)	30 (1.7)	1 (0.7)	12 (1.3)	17 (2.4)
Frequency (%)				
1 = Not at all	422 (23.7)	45 (30.4)	239 (25.8)	138 (19.6)
2 = A little bit	434 (24.4)	30 (20.3)	247 (26.7)	157 (22.3)
3 = Somewhat	375 (21.1)	24 (16.2)	208 (22.5)	143 (20.3)
4 = Quite a bit	378 (21.3)	30 (20.3)	169 (18.3)	179 (25.4)
5 = Very much	169 (9.5)	19 (12.8)	62 (6.7)	88 (12.5)

PROMIS = Patient-Reported Outcomes Measurement Information System.

Confirmatory Factor Analysis

To further evaluate the customized PROMIS Fatigue measure in this study, two 1-factor and one 2-factor confirmatory factor analyses (CFAs) were conducted using the baseline responses. Results for the CFA modeling are provided in Table 6. The first CFA proposed 1-

factor solution with all the PROMIS Fatigue items, the standard loadings were very strong in size, ranging from 0.814 for Item 6 (How often were you too tired to take a bath or shower?) to 0.945 for Item 9 (How run-down did you feel on average?) except for item 7 (0.060), and resulted in fit statistics that were not optimal (i.e., RMSEA = 0.102, lightly greater than 0.1), while both comparative fit index [CFI] (0.994) and Tucker-Lewis Index [TLI] (0.993) were larger than 0.95, and standardized root mean residual (SRMR) was 0.019. The second 1-factor CFA was conducted with all the PROMIS Fatigue items except for Item 7, the results showed that the standard factor loading were exactly the same and the fit statistics were similar, CFI was 0.994 and TLI was 0.993 for TLI, RMSEA was 0.106, and SRMR was 0.018. Due to the less than ideal fit indicated from the RMSEA, a 2-fatcor CFA model were conducted (without item 7). The proposed structure had Item 1 to 6 plus Item 8 as frequency factor, and Item 9 to 14 as severity factor. The standard loadings were strong in size on both factors, ranging from 0.823 for Item 6 (How often were you too tired to take a bath or shower?) to 0.922 for Item 4 (How often did your fatigue limit you at work (include work at home)?), and ranging from 0.939 for Item 11 (How fatigued were you on average?) and Item 14 (To what degree did your fatigue interfere with your physical functioning?) to 0.951 for Item 10 (I feel fatigued). The fit statistics, 0.996 for both CFI and TLI, RMSEA was 0.084 and SRMR was 0.014, indicated a good fit. However, the inter-factor correlation were extremely high (r = 0.965). The residual covariance for 2-factor model without item 7 was less than 0.02, except for item 6 with item 9 (-0.034), item 10 (-0.048), item 11 (-0.028) and item 13 (-0.040). As a summary, this customized form of PROMIS Fatigue measure was supported in unidimensional structure for the proposed IRT calibration and scoring as well as DIF analysis. Item 7 was then removed and excluded from the analysis.

Table 6. Confirmatory Factor Analysis for PROMIS Fatigue Measure

	1-Factor	1-Factor	2-Factor	
PROMIS Fatigue Item		(No Item 7)	(No Item 7)	
1. How often did you feel tired?	0.896*(0.003)	0.896*(0.003)	0.909* (0.003)	—
2. How often did you experience extreme exhaustion?	0.889*(0.003)	0.889*(0.003)	0.898*(0.003)	—
3. How often did you run out of energy?	0.902*(0.003)	0.902*(0.003)	0.911* (0.003)	—
4. How often did your fatigue limit you at work (include work at home)?	0.911*(0.003)	0.911*(0.003)	0.922* (0.003)	—
5. How often were you too tired to think clearly?	0.848*(0.005)	0.848*(0.005)	0.857* (0.004)	—
6. How often were you too tired to take a bath or shower?	0.814*(0.006)	0.814*(0.006)	0.823*(0.006)	—
7. How often did you have enough energy to exercise strenuously?	-0.060*(0.013)	—	—	—
8. How often did you have to push yourself to get things done because of your fatigue?	0.887*(0.003)	0.887*(0.003)	0.899*(0.003)	—
9. How run-down did you feel on average?	0.945*(0.002)	0.945*(0.002)	—	0.949* (0.002)
10. I feel fatigued	0.947*(0.002)	0.947*(0.002)	—	0.951*(0.002)
11. How fatigued were you on average?	0.935*(0.002)	0.935*(0.002)	—	0.939*(0.002)
12. I have trouble starting things because I am tired.	0.930*(0.002)	0.930*(0.002)	—	0.934*(0.002)
13. How much were you bothered by your fatigue on average?	0.942*(0.002)	0.942*(0.002)	—	0.945*(0.002))
14. To what degree did your fatigue interfere with your physical functioning?	0.935*(0.002)	0.935*(0.002)	—	0.939*(0.002)
Model Fit			(Factor 1 with Factor 2, r = 0.965)	
χ^2 (df)	4479.810* (77)	4067.220* (65)	2561.811* (64)	

	1-Factor	1-Factor	2-Factor
PROMIS Fatigue Item		(No Item 7)	(No Item 7)
P value	< 0.0001	< 0.0001	< 0.0001
CFI	0.994	0.994	0.996
TLI	0.993	0.993	0.996
RMSEA (90% CI)	0.102 (0.099, 0.104)	0.106 (0.103, 0.109)	0.084 (0.081, 0.087)
SRMR	0.019	0.018	0.014

CFI = comparative fit index; PROMIS = Patient-Reported Outcomes Measurement Information System; RMSEA =root mean square error of approximation; SRMR = standardized root mean squared residual; TLI = Tucker–Lewis index.

Graded Response Model

Item fit

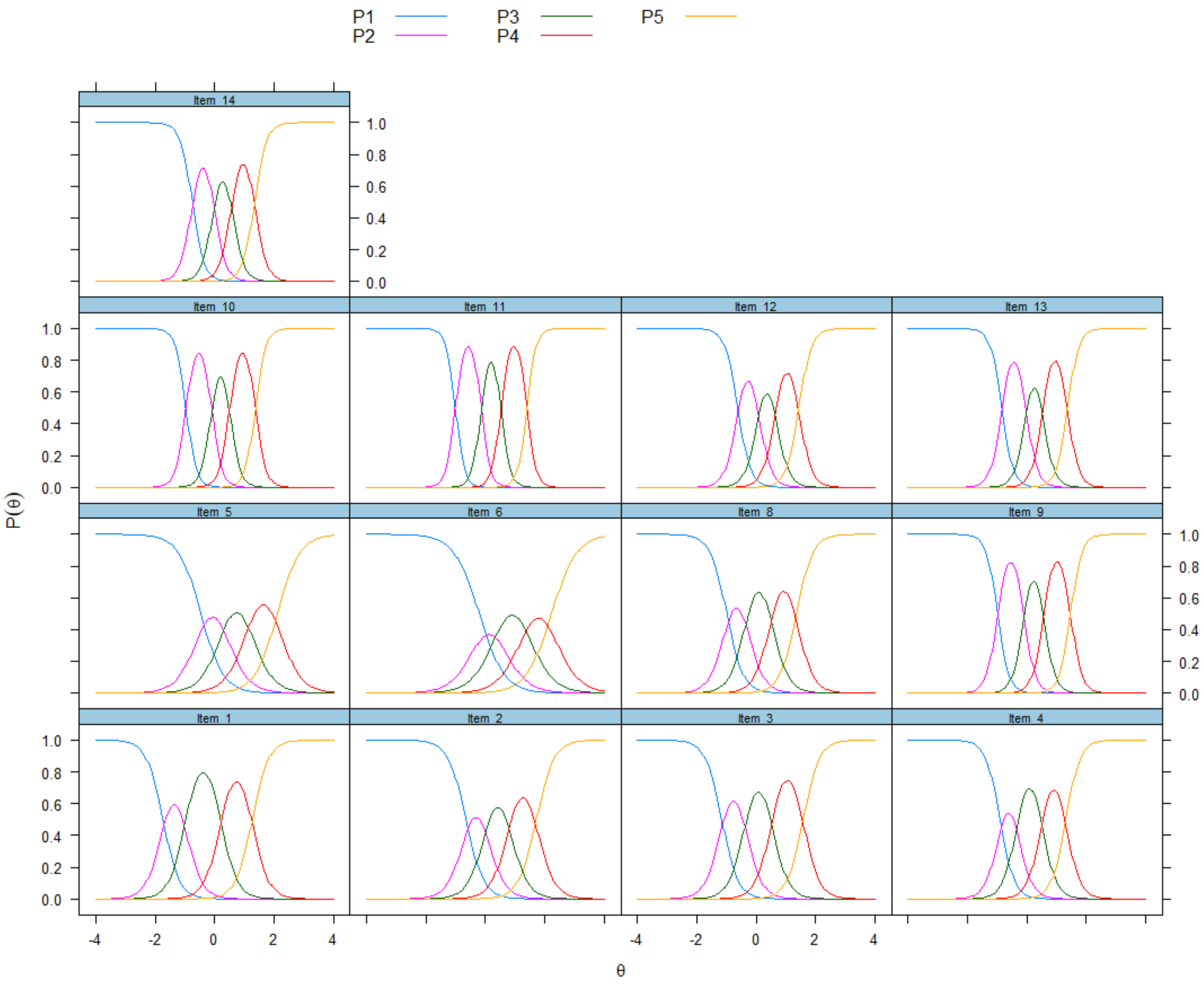
Table 7 lists item fit statistics using χ^2 and RMSEA for all the PROMIS Fatigue items (without item 7). Of the 13 items, item 1 [How often did you feel tired?] was the only item exhibited potential misfit (p<0.0001).The values of root mean-square error of approximation (RMSEA) were also reported to help gauge the magnitude of item misfit. With RMSEA for all the items less than 0.025, it demonstrated data-model fit. The result is presented in Table 7 and Figure 4.

Table 7. Item fit statistics for Graded Response Model

PROMIS Fatigue Item	χ^2	df	RMSEA	p-value
Item_1	184.47	87	0.025	<0.0001
Item 2	101.78	112	<0.001	0.7454
Item 3	128.66	101	0.012	0.0330
Item 4	155.23	97	0.018	0.0002
Item 5	160.14	124	0.013	0.0160
Item 6	161.90	128	0.012	0.0229
Item 8	131.26	108	0.011	0.0635
Item 9	75.31	71	0.006	0.3408
Item 10	56.80	69	<0.001	0.8530
Item 11	56.14	59	<0.001	0.5815
Item 12	83.32	89	<0.001	0.6497
Item 13	103.47	79	0.013	0.0338
Item 14	100.45	86	0.010	0.1365

df = degree of freedom; GRM = graded response model; PROMIS = Patient-Reported Outcomes Measurement Information System.

Figure 4. Item Characteristics Curve



GRM = graded response model.

DIF Analysis

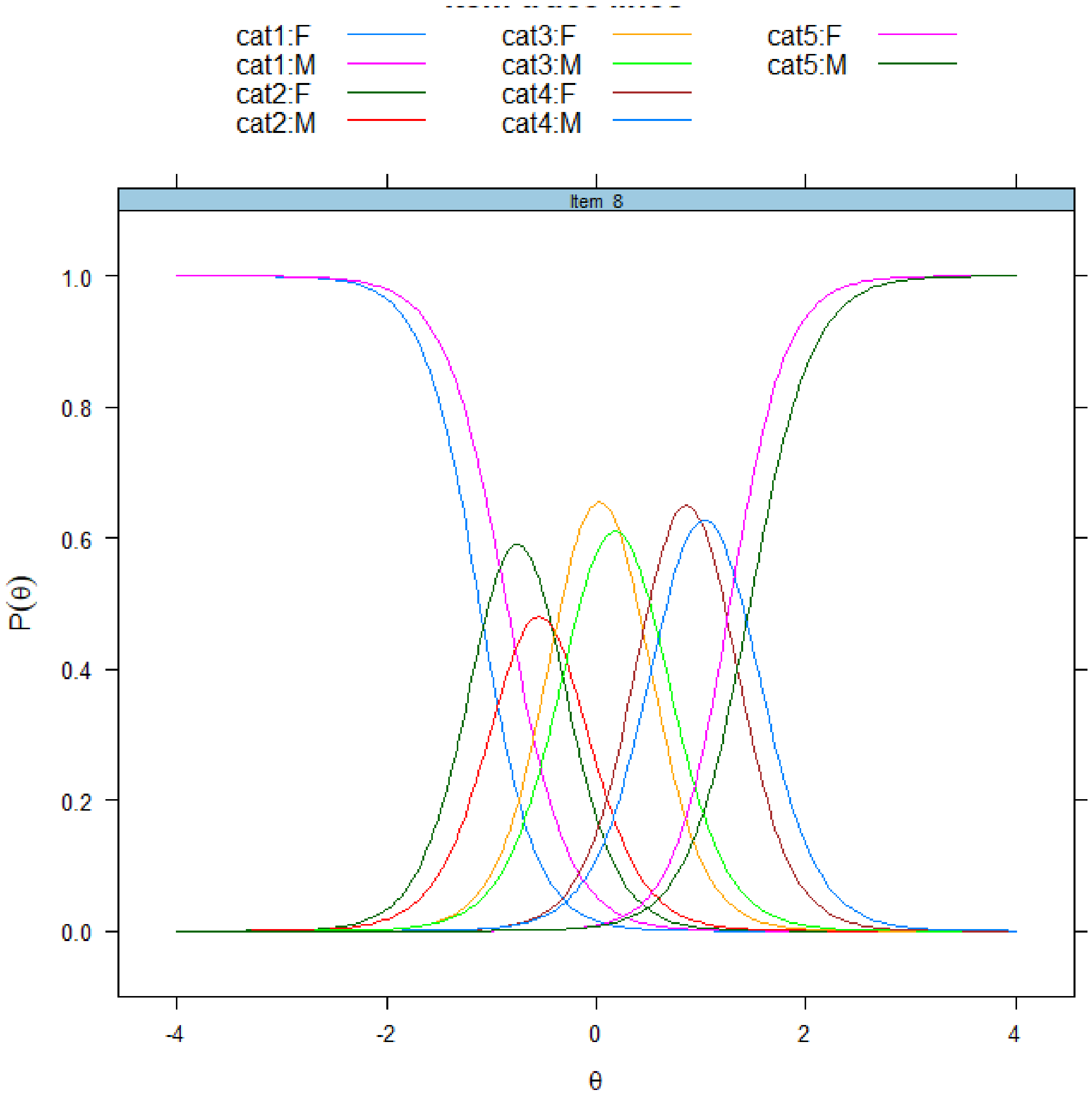
Result from the DIF analyses is shown in Table 8. From Wald χ^2 statistics for DIF analysis, three selected demographic groups had found one DIF items (p < 0.001) in each of them. Item 8 [How often did you have to push yourself to get things done because of your fatigue?] was found on gender DIF for male vs female (p = 0.0007), item 6 [How often did you have to push yourself to get things done because of your fatigue?] was found on marital status DIF for married vs other status (p < 0.0001), and item 2 [How often did you experience extreme exhaustion?] was found on race DIF for white vs others (p = 0.0003). The result was presented in Table 8 and Figure 5 to Figure 7.

Table 8. DIF Analysis

	Gender		Marital Status		Race	
PROMIS Fatigue Item	χ^2 (df)	p-value	χ^2 (df)	p-value	χ^2 (df)	p-value
Item 1	10.82 (4)	0.0287	5.52 (4)	0.2381	11.33 (4)	0.0231
Item 2	8.78 (4)	0.0668	10.8 (4)	0.0289	21.13 (4)	**0.0003**
Item 3	10.59 (4)	0.0316	8.84 (4)	0.0652	11.69 (4)	0.0198
Item 4	10.41 (4)	0.0341	6.96 (4)	0.1380	6.75 (4)	0.1499
Item 5	15.74 (4)	0.0034	9.79 (4)	0.0441	5.38 (4)	0.2501
Item 6	11.58 (4)	0.0208	33.4 (4)	**<0.0001**	5.61 (4)	0.2300
Item 8	19.25 (4)	**0.0007**	11.41 (4)	0.0223	11.62 (4)	0.0204
Item 9	6.12 (4)	0.1905	9.75 (4)	0.0448	9 (4)	0.0610
Item 10	12.99 (4)	0.0113	13.07 (4)	0.0110	0.41 (4)	0.9813
Item 11	5.61 (4)	0.2302	13.65 (4)	0.0085	1.73 (4)	0.7845
Item 12	10.3 (4)	0.0357	7.92 (4)	0.0947	5.15 (4)	0.2718
Item 13	6.46 (4)	0.1674	10.12 (4)	0.0385	3.22 (4)	0.5215
Item 14	5.85 (4)	0.2103	15.15 (4)	0.0044	5.6 (4)	0.2314

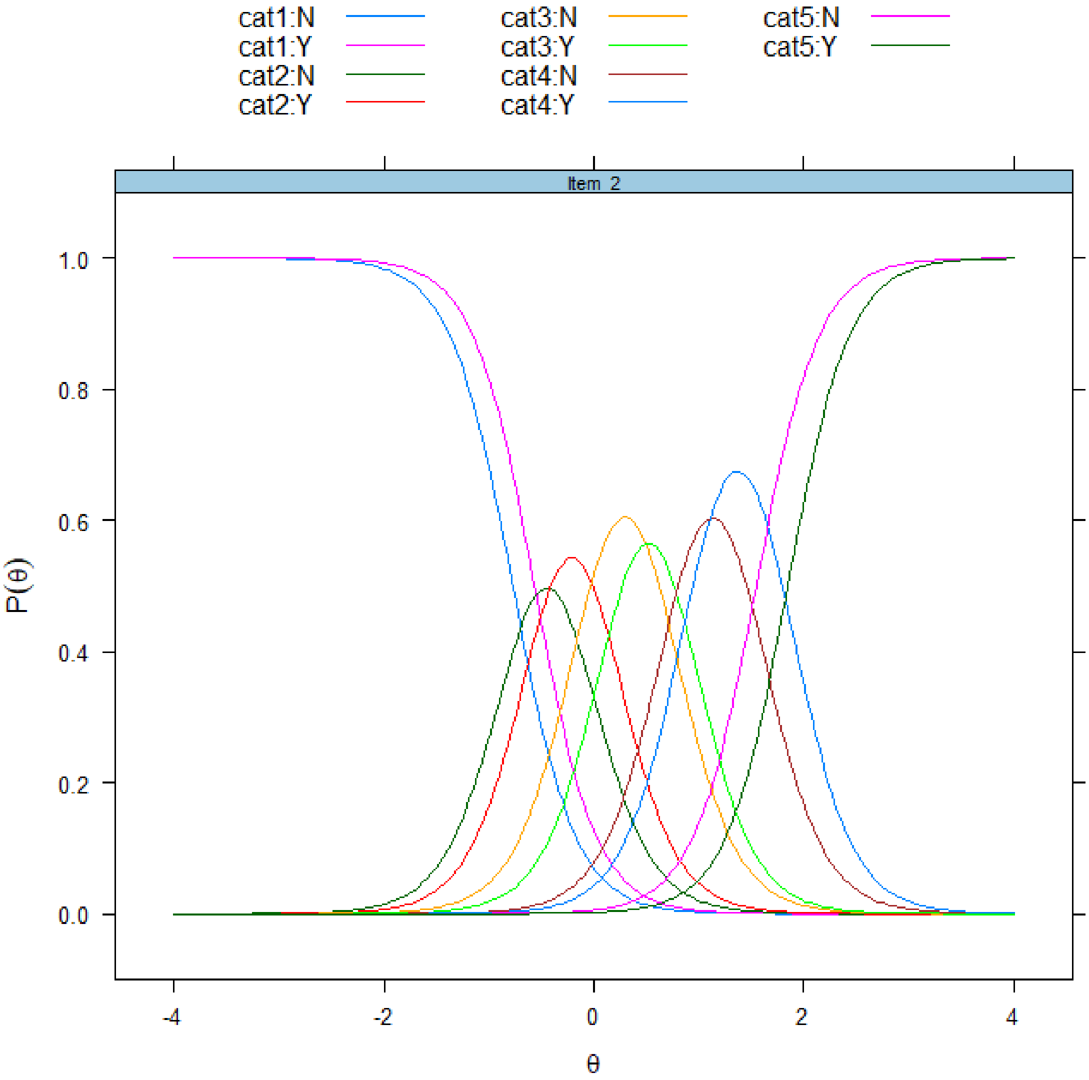
DIF = differential item functioning; PROMIS = Patient-Reported Outcomes Measurement Information System.

Figure 5. Item Character Curve for Item 8 by Gender DIF Group



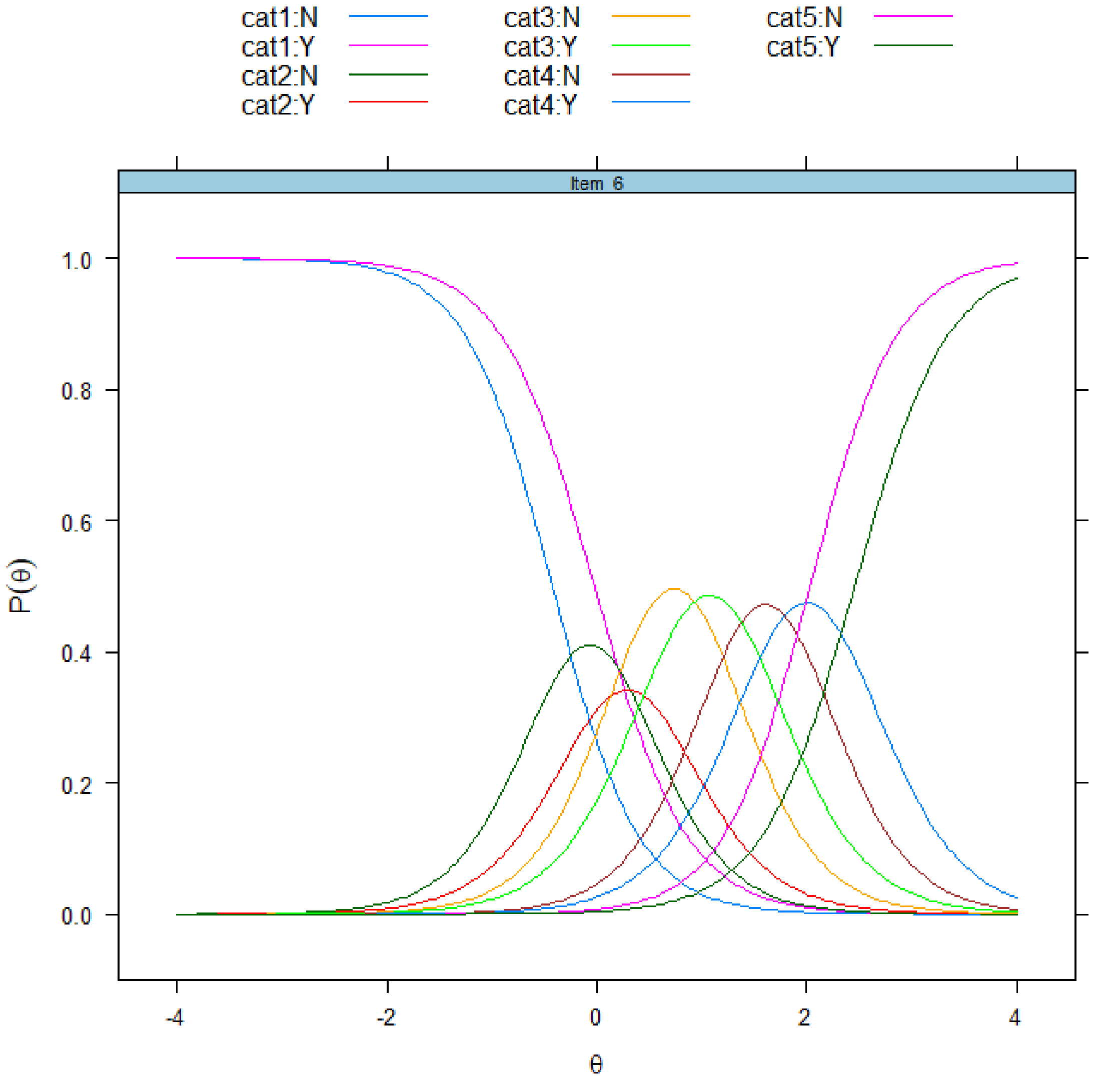
cat = category; DIF = differential item functioning; F = female; M = male.

Figure 6. Item Character Curve for Item 2 by Race DIF Group



cat = category; DIF = differential item functioning; N = no; Y = yes.

Figure 7. Item Character Curve for Item 6 by Martial Status DIF Group



cat = category; DIF = differential item functioning; F = female; M = male.

Calibration using three different calibration approaches

Item fit statistics from three different calibration approaches for the three DIF groups are presented in Table 9 to Table 11. In table 9 for gender DIF group, compared to all items (ignore DIF item) approach, item 1 from removed DIF item approach was significant (p<0.0001) and RMSEA was 0.030, which was larger in magnitude than all items (ignore DIF items) approach; the same item in demographic-specific item approach showed a fit in GRM (p = 0.0066 > 0.0001) and the value of RMSEA was 0.015. Similar results have been found in race and marital status DIF groups. In table 10 for race DIF group, item 1 and item 3 from removed DIF item approach showed potential misfit (p<0.0001), RMSEA values were 0.0265 and 0.019, respectively. In table 11 for marital status DIF group, item 1 from removed DIF item approach had p<0.0001 and RMSEA 0.029. The other items, including demographic-specific items, for example, Item 2 (White) and Item 2 (Other) from race DIF group showed a great item fit ($\text{RMSEA}_{\text{item 2 (White)}} = 0.007$ and $\text{RMSEA}_{\text{item 2 (Others)}} < 0.001$).

IRT Parameter Estimation of three different calibration approaches for the three DIF groups are summarized in Table 13 to Table 14. For all items approach, the a-parameters ranged from 2.65 (item 5 [How often did you feel tired?]) to 6.38 (item 11 [How fatigued were you on average?]), b1 parameter ranged from -1.74 (item 1 [How often did you feel tired?]) to -0.20 (item 5 [How often did you feel tired?]), b2 parameter ranged from -0.98 (item 1 [How often did you feel tired?]) to 0.45 (item 5 [How often did you feel tired?]), b3 parameter ranged from 0.22 (item 1 [How often did you feel tired?]) to 1.36 (item 6 [How often were you too tired to take a bath or shower?]), b4 parameter ranged from 1.27 (item 1 [How often did you feel tired?]) to 2.23 (item 6 [How often were you too tired to take a bath or shower?]). The parameter estimation and standard error were very similar among the three calibration approaches for all the DIF groups.

Table 9. Item fit statistics with Gender DIF Group

	Ignore DIF				Removed DIF				Demographic-specific			
Item	χ^2	df	RMSEA	p-value	χ^2	df	RMSEA	p-value	χ^2	df	RMSEA	p-value
Item 1	184.47	87	0.025	<0.0001	222.76	86	0.030	<0.0001	118.33	83	0.015	0.0066
Item 2	101.78	112	<0.001	0.7454	115.67	111	0.005	0.3617	106.12	112	<0.001	0.6387
Item 3	128.66	101	0.012	0.0330	144.42	102	0.015	0.0037	123.11	100	0.011	0.0583
Item 4	155.23	97	0.018	0.0002	140.80	99	0.015	0.0037	154.20	97	0.018	0.0002
Item 5	160.14	124	0.013	0.0160	157.37	124	0.012	0.0230	163.28	123	0.013	0.0088
Item 6	161.90	128	0.012	0.0229	171.06	129	0.013	0.0078	164.44	128	0.013	0.0165
Item 8	131.26	108	0.011	0.0635	—	—	—	—	—	—	—	—
Item 8 (Male)		—	—	—	—	—	—	—	48.38	60	<0.001	0.8592
Item 8 (Female)		—	—	—	—	—	—	—	92.75	70	0.013	0.0358
Item 9	75.31	71	0.006	0.3408	98.84	72	0.014	0.0197	80.04	71	0.008	0.2165
Item 10	56.80	69	<0.001	0.8530	80.84	66	0.011	0.1034	61.20	69	<0.001	0.7368
Item 11	56.14	59	<0.001	0.5815	55.23	54	0.004	0.4279	60.75	59	0.004	0.4127
Item 12	83.32	89	<0.001	0.6497	89.13	88	0.003	0.4462	90.86	89	0.003	0.4253
Item 13	103.47	79	0.013	0.0338	94.41	79	0.010	0.1138	110.10	80	0.014	0.0145
Item 14	100.45	86	0.010	0.1365	117.82	86	0.014	0.0129	101.76	87	0.010	0.1332

DIF = differential item functioning.

Table 10. Item level fit statistics with Race DIF Group

PROMIS Fatigue Item	Ignore DIF				Removed DIF				Demographic-specific			
	χ^2	df	RMSEA	p-value	χ^2	df	RMSEA	p-value	χ^2	df	RMSEA	p-value
Item 1	184.47	87	0.025	**<0.0001**	199.31	88	0.026	**<0.0001**	116.56	83	0.015	0.0089
Item 2	101.78	112	<0.001	0.7454	–	–	–	–	–	–	–	–
Item 2 (White)	–	–	–	–	–	–	–	–	76.43	70	0.007	0.2798
Item 2 (Others)	–	–	–	–	–	–	–	–	55.32	60	<0.001	0.6473
Item 3	128.66	101	0.012	0.0330	167.35	102	0.019	**<0.0001**	119.68	100	0.010	0.0876
Item 4	155.23	97	0.018	0.0002	145.23	99	0.016	0.0017	138.68	98	0.015	0.0043
Item 5	160.14	124	0.013	0.0160	181.37	126	0.016	0.0009	170.15	123	0.015	0.0032
Item 6	161.90	128	0.012	0.0229	187.37	128	0.016	0.0005	166.99	128	0.013	0.0117
Item 8	131.26	108	0.011	0.0635	160.21	110	0.016	0.0013	122.51	108	0.009	0.1609
Item 9	75.31	71	0.006	0.3408	72.39	72	0.002	0.4649	79.19	71	0.008	0.2363
Item 10	56.80	69	<0.001	0.8530	69.80	68	0.004	0.4167	54.68	69	<0.001	0.8957
Item 11	56.14	59	<0.001	0.5815	76.97	58	0.013	0.0485	65.41	58	0.008	0.2352
Item 12	83.32	89	<0.001	0.6497	103.32	88	0.010	0.1265	83.37	90	<0.001	0.6761
Item 13	103.47	79	0.013	0.0338	110.46	76	0.016	0.0060	99.26	79	0.012	0.0614
Item 14	100.45	86	0.010	0.1365	91.06	84	0.007	0.2806	106.80	88	0.011	0.0843

DIF = differential item functioning.

Table 11. Item level fit statistics with Martial Status DIF Group

	Ignore DIF				Removed DIF				Demographic-specific			
PROMIS Fatigue Item	χ^2	df	RMSEA	p-value	χ^2	df	RMSEA	p-value	χ^2	df	RMSEA	p-value
Item 1	180.60	87	0.025	**<0.0001**	212.92	86	0.029	**<0.0001**	121.76	84	0.016	0.0045
Item 2	108.42	112	<0.001	0.5782	130.53	111	0.010	0.0994	109.85	112	<0.001	0.5398
Item 3	132.81	101	0.013	0.0186	152.27	100	0.017	0.0006	140.69	99	0.015	0.0038
Item 4	139.14	97	0.016	0.0033	131.45	97	0.014	0.0114	141.59	96	0.016	0.0017
Item 5	168.42	122	0.015	0.0034	173.24	123	0.015	0.0019	176.80	122	0.016	0.0009
Item 6	180.29	127	0.015	0.0013	–	–	–	–	–	–	–	–
Item 6 (Married)	–	–	–	–	–	–	–	–	103.46	72	0.016	0.0089
Item 6 (Other)	–	–	–	–	–	–	–	–	82.38	69	0.010	0.1296
Item 8	120.87	108	0.008	0.1872	140.87	108	0.013	0.0184	130.41	107	0.011	0.0616
Item 9	61.98	71	<0.001	0.7688	72.61	70	0.005	0.3921	62.95	72	<0.001	0.7680
Item 10	54.12	68	<0.001	0.8896	63.98	67	<0.001	0.5820	54.34	67	<0.001	0.8670
Item 11	54.72	58	<0.001	0.5982	67.78	56	0.011	0.1344	57.59	58	<0.001	0.4906
Item 12	96.78	89	0.007	0.2688	101.49	90	0.008	0.1917	103.89	88	0.010	0.1186
Item 13	104.07	77	0.014	0.0217	121.81	77	0.018	0.0009	99.52	77	0.013	0.0431
Item 14	97.06	84	0.009	0.1562	104.15	87	0.011	0.1014	98.71	84	0.010	0.1303

DIF = differential item functioning.

Table 12. IRT Parameter Estimation with Gender DIF Group

PROMIS Fatigue Item	Ignore DIF					Removed DIF					Demographic-specific				
	a	b1	b2	b3	b4	a	b1	b2	b3	b4	a	b1	b2	b3	b4
Item 1	3.61 (0.13)	-1.74 (0.06)	-0.98 (0.04)	0.22 (0.03)	1.27 (0.04)	3.61 (0.13)	-1.74 (0.06)	-0.98 (0.04)	0.21 (0.03)	1.27 (0.04)	3.60 (0.13)	-1.74 (0.06)	-0.99 (0.04)	0.21 (0.03)	1.27 (0.04)
Item 2	3.33 (0.12)	-0.65 (0.04)	0.03 (0.03)	0.82 (0.04)	1.72 (0.06)	3.32 (0.12)	-0.65 (0.04)	0.03 (0.03)	0.82 (0.04)	1.72 (0.06)	3.33 (0.12)	-0.65 (0.04)	0.03 (0.03)	0.81 (0.04)	1.71 (0.06)
Item 3	3.64 (0.13)	-1.16 (0.04)	-0.37 (0.03)	0.53 (0.03)	1.58 (0.05)	3.58 (0.13)	-1.16 (0.04)	-0.37 (0.03)	0.53 (0.03)	1.59 (0.05)	3.64 (0.13)	-1.16 (0.04)	-0.37 (0.03)	0.52 (0.03)	1.58 (0.05)
Item 4	4.14 (0.15)	-0.91 (0.04)	-0.33 (0.03)	0.50 (0.03)	1.31 (0.04)	4.05 (0.15)	-0.91 (0.04)	-0.33 (0.03)	0.50 (0.03)	1.32 (0.04)	4.14 (0.15)	-0.91 (0.04)	-0.33 (0.03)	0.5 (0.03)	1.31 (0.04)
Item 5	2.65 (0.1)	-0.46 (0.04)	0.33 (0.03)	1.17 (0.05)	2.11 (0.08)	2.63 (0.1)	-0.46 (0.04)	0.33 (0.03)	1.17 (0.05)	2.13 (0.08)	2.65 (0.1)	-0.46 (0.04)	0.32 (0.03)	1.17 (0.05)	2.11 (0.08)
Item 6	2.36 (0.09)	-0.20 (0.04)	0.45 (0.04)	1.36 (0.05)	2.23 (0.09)	2.33 (0.09)	-0.21 (0.04)	0.45 (0.04)	1.37 (0.05)	2.24 (0.09)	2.36 (0.09)	-0.21 (0.04)	0.45 (0.04)	1.36 (0.05)	2.22 (0.09)
Item 8	3.58 (0.13)	-0.99 (0.04)	-0.32 (0.03)	0.51 (0.03)	1.36 (0.05)	–	–	–	–	–	–	–	–	–	–
Item 8 (Male)	–	–	–	–	–	–	–	–	–	–	3.42 (0.17)	-0.92 (0.05)	-0.31 (0.04)	0.52 (0.04)	1.38 (0.06)
Item 8 (Female)	–	–	–	–	–	–	–	–	–	–	3.73 (0.17)	-1.06 (0.05)	-0.34 (0.04)	0.50 (0.04)	1.33 (0.05)
Item 9	5.21 (0.2)	-0.99 (0.04)	-0.10 (0.03)	0.57 (0.03)	1.47 (0.04)	5.27 (0.2)	-0.99 (0.04)	-0.11 (0.03)	0.56 (0.03)	1.47 (0.04)	5.21 (0.2)	-0.99 (0.04)	-0.11 (0.03)	0.56 (0.03)	1.47 (0.04)
Item 10	5.68 (0.23)	-0.96 (0.04)	-0.1 (0.03)	0.51 (0.03)	1.38 (0.04)	5.82 (0.24)	-0.96 (0.04)	-0.1 (0.03)	0.5 (0.03)	1.38 (0.04)	5.67 (0.23)	-0.97 (0.04)	-0.1 (0.03)	0.51 (0.03)	1.38 (0.04)
Item 11	6.38 (0.27)	-1.01 (0.04)	-0.14 (0.03)	0.53 (0.03)	1.40 (0.04)	6.56 (0.28)	-1.01 (0.04)	-0.14 (0.03)	0.52 (0.03)	1.40 (0.04)	6.38 (0.27)	-1.02 (0.04)	-0.14 (0.03)	0.52 (0.03)	1.4 (0.04)
Item 12	4.64 (0.17)	-0.62 (0.03)	0.08 (0.03)	0.66 (0.03)	1.44 (0.04)	4.64 (0.17)	-0.62 (0.03)	0.07 (0.03)	0.65 (0.03)	1.44 (0.05)	4.64 (0.17)	-0.62 (0.03)	0.07 (0.03)	0.65 (0.03)	1.43 (0.04)
Item 13	5.19 (0.20)	-0.85 (0.04)	-0.03 (0.03)	0.53 (0.03)	1.37 (0.04)	5.19 (0.20)	-0.84 (0.04)	-0.03 (0.03)	0.53 (0.03)	1.37 (0.04)	5.19 (0.20)	-0.85 (0.04)	-0.03 (0.03)	0.53 (0.03)	1.36 (0.04)
Item 14	4.92 (0.19)	-0.75 (0.03)	-0.02 (0.03)	0.58 (0.03)	1.35 (0.04)	4.94 (0.19)	-0.75 (0.03)	-0.03 (0.03)	0.58 (0.03)	1.35 (0.04)	4.92 (0.19)	-0.75 (0.03)	-0.03 (0.03)	0.58 (0.03)	1.34 (0.04)

DIF = differential item functioning.

Table 13. IRT Parameter Estimation with Race DIF Group

	Ignore DIF					Removed DIF					Demographic-specific				
PROMIS Fatigue Item	A	b1	b2	b3	b4	a	b1	b2	b3	b4	a	b1	b2	b3	b4
Item 1	3.61 (0.13)	-1.74 (0.06)	-0.98 (0.04)	0.20 (0.03)	1.20 (0.04)	3.50 (0.13)	-1.70 (0.06)	-0.90 (0.04)	0.20 (0.03)	1.20 (0.04)	3.60 (0.13)	-1.70 (0.06)	-0.90 (0.04)	0.20 (0.03)	1.20 (0.04)
Item 2	3.33 (0.12)	-0.65 (0.04)	0.03 (0.03)	0.82 (0.04)	1.72 (0.06)	—	—	—	—	—	—	—	—	—	—
Item 2 (White)	—	—	—	—	—	—	—	—	—	—	3.52 (0.16)	-0.56 (0.04)	0.13 (0.04)	0.86 (0.04)	1.79 (0.07)
Item 2 (Others)	—	—	—	—	—	—	—	—	—	—	3.25 (0.17)	-0.79 (0.05)	-0.12 (0.04)	0.74 (0.05)	1.61 (0.07)
Item 3	3.64 (0.13)	-1.16 (0.04)	-0.37 (0.03)	0.53 (0.03)	1.58 (0.05)	3.55 (0.13)	-1.16 (0.04)	-0.37 (0.03)	0.53 (0.03)	1.59 (0.05)	3.65 (0.13)	-1.16 (0.04)	-0.37 (0.03)	0.52 (0.03)	1.58 (0.05)
Item 4	4.14 (0.15)	-0.91 (0.04)	-0.33 (0.03)	0.5 (0.03)	1.31 (0.04)	4.08 (0.15)	-0.91 (0.04)	-0.33 (0.03)	0.50 (0.03)	1.31 (0.04)	4.14 (0.15)	-0.91 (0.04)	-0.33 (0.03)	0.50 (0.03)	1.31 (0.04)
Item 5	2.65 (0.1)	-0.46 (0.04)	0.33 (0.03)	1.17 (0.05)	2.11 (0.08)	2.60 (0.1)	-0.46 (0.04)	0.33 (0.03)	1.18 (0.05)	2.12 (0.08)	2.65 (0.1)	-0.46 (0.04)	0.32 (0.03)	1.17 (0.05)	2.11 (0.08)
Item 6	2.36 (0.09)	-0.20 (0.04)	0.45 (0.04)	1.36 (0.05)	2.23 (0.09)	2.34 (0.09)	-0.20 (0.04)	0.45 (0.04)	1.36 (0.05)	2.23 (0.09)	2.36 (0.09)	-0.21 (0.04)	0.45 (0.04)	1.36 (0.05)	2.22 (0.09)
Item 8	3.58 (0.13)	-0.99 (0.04)	-0.32 (0.03)	0.51 (0.03)	1.36 (0.05)	3.57 (0.13)	-0.99 (0.04)	-0.32 (0.03)	0.51 (0.03)	1.36 (0.05)	3.59 (0.13)	-0.99 (0.04)	-0.33 (0.03)	0.51 (0.03)	1.35 (0.05)
Item 9	5.21 (0.2)	-0.99 (0.04)	-0.10 (0.03)	0.57 (0.03)	1.47 (0.04)	5.25 (0.20)	-0.99 (0.04)	-0.10 (0.03)	0.57 (0.03)	1.47 (0.04)	5.21 (0.2)	-0.99 (0.04)	-0.11 (0.03)	0.56 (0.03)	1.47 (0.04)
Item 10	5.68 (0.23)	-0.96 (0.04)	-0.10 (0.03)	0.51 (0.03)	1.38 (0.04)	5.77 (0.23)	-0.96 (0.04)	-0.10 (0.03)	0.51 (0.03)	1.38 (0.04)	5.68 (0.23)	-0.97 (0.04)	-0.1 (0.03)	0.51 (0.03)	1.38 (0.04)
Item 11	6.38 (0.27)	-1.01 (0.04)	-0.14 (0.03)	0.53 (0.03)	1.4 (0.04)	6.44 (0.27)	-1.01 (0.04)	-0.14 (0.03)	0.53 (0.03)	1.40 (0.04)	6.37 (0.27)	-1.02 (0.04)	-0.14 (0.03)	0.52 (0.03)	1.40 (0.04)
Item 12	4.64 (0.17)	-0.62 (0.03)	0.08 (0.03)	0.66 (0.03)	1.44 (0.04)	4.72 (0.18)	-0.62 (0.03)	0.08 (0.03)	0.66 (0.03)	1.43 (0.04)	4.64 (0.17)	-0.62 (0.03)	0.07 (0.03)	0.65 (0.03)	1.43 (0.05)
Item 13	5.19 (0.20)	-0.85 (0.04)	-0.03 (0.03)	0.53 (0.03)	1.37 (0.04)	5.30 (0.20)	-0.84 (0.04)	-0.03 (0.03)	0.54 (0.03)	1.36 (0.04)	5.19 (0.2)	-0.85 (0.04)	-0.03 (0.03)	0.53 (0.03)	1.36 (0.04)

	4.92	-0.75	-0.02	0.58	1.35	5.00	-0.75	-0.02	0.58	1.35	4.91	-0.75	-0.03	0.58	1.35
Item 14	(0.19)	(0.03)	(0.03)	(0.03)	(0.04)	(0.19)	(0.03)	(0.03)	(0.03)	(0.04)	(0.19)	(0.03)	(0.03)	(0.03)	(0.04)

Table 14. IRT Parameter Estimation with Martial Status DIF Group

	Ignore DIF					Removed DIF					Demographic-specific				
PROMIS Fatigue Item	a	b1	b2	b3	b4	a	b1	b2	b3	b4	a	b1	b2	b3	b4
Item 1	3.6 (0.13)	-1.75 (0.06)	-0.98 (0.04)	0.21 (0.03)	1.27 (0.04)	3.61 (0.13)	-1.75 (0.06)	-0.99 (0.04)	0.21 (0.03)	1.27 (0.04)	3.6 (0.13)	-1.75 (0.06)	-0.99 (0.04)	0.21 (0.03)	1.27 (0.04)
Item 2	3.32 (0.12)	-0.65 (0.04)	0.03 (0.03)	0.81 (0.04)	1.72 (0.06)	3.29 (0.12)	-0.66 (0.04)	0.03 (0.03)	0.82 (0.04)	1.72 (0.06)	3.32 (0.12)	-0.66 (0.04)	0.02 (0.03)	0.81 (0.04)	1.71 (0.06)
Item 3	3.62 (0.13)	-1.17 (0.04)	-0.37 (0.03)	0.53 (0.03)	1.59 (0.05)	3.6 (0.13)	-1.17 (0.04)	-0.37 (0.03)	0.53 (0.03)	1.59 (0.05)	3.62 (0.13)	-1.17 (0.04)	-0.37 (0.03)	0.52 (0.03)	1.58 (0.05)
Item 4	4.14 (0.15)	-0.91 (0.04)	-0.33 (0.03)	0.5 (0.03)	1.31 (0.04)	4.1 (0.15)	-0.91 (0.04)	-0.33 (0.03)	0.5 (0.03)	1.32 (0.04)	4.14 (0.15)	-0.91 (0.04)	-0.33 (0.03)	0.49 (0.03)	1.31 (0.04)
Item 5	2.63 (0.1)	-0.46 (0.04)	0.33 (0.03)	1.18 (0.05)	2.12 (0.08)	2.57 (0.1)	-0.46 (0.04)	0.33 (0.04)	1.18 (0.05)	2.13 (0.08)	2.63 (0.1)	-0.46 (0.04)	0.33 (0.03)	1.17 (0.05)	2.12 (0.08)
Item 6	2.34 (0.09)	-0.2 (0.04)	0.46 (0.04)	1.37 (0.05)	2.24 (0.09)	–	–	–	–	–	–	–	–	–	–
Item 6 (Married)	–	–	–	–	–	–	–	–	–	–	2.26 (0.12)	-0.09 (0.04)	0.54 (0.05)	1.49 (0.07)	2.40 (0.13)
Item 6 (Other)	–	–	–	–	–	–	–	–	–	–	2.45 (0.13)	-0.35 (0.05)	0.36 (0.05)	1.25 (0.06)	2.09 (0.1)
Item 8	3.57 (0.13)	-1.00 (0.04)	-0.32 (0.03)	0.51 (0.03)	1.36 (0.05)	3.54 (0.13)	-1.00 (0.04)	-0.33 (0.03)	0.51 (0.03)	1.36 (0.05)	3.57 (0.13)	-1.00 (0.04)	-0.33 (0.03)	0.51 (0.03)	1.35 (0.05)
Item 9	5.2 (0.2)	-0.99 (0.04)	-0.1 (0.03)	0.57 (0.03)	1.47 (0.05)	5.25 (0.2)	-0.99 (0.04)	-0.1 (0.03)	0.57 (0.03)	1.47 (0.05)	5.2 (0.2)	-1 (0.04)	-0.11 (0.03)	0.56 (0.03)	1.47 (0.05)
Item 10	5.64 (0.23)	-0.97 (0.04)	-0.1 (0.03)	0.51 (0.03)	1.38 (0.04)	5.76 (0.23)	-0.97 (0.04)	-0.1 (0.03)	0.51 (0.03)	1.38 (0.04)	5.65 (0.23)	-0.97 (0.04)	-0.1 (0.03)	0.5 (0.03)	1.38 (0.04)
Item 11	6.35 (0.27)	-1.02 (0.04)	-0.14 (0.03)	0.53 (0.03)	1.41 (0.04)	6.46 (0.27)	-1.02 (0.04)	-0.14 (0.03)	0.53 (0.03)	1.41 (0.04)	6.35 (0.27)	-1.02 (0.04)	-0.14 (0.03)	0.52 (0.03)	1.4 (0.04)
Item 12	4.61 (0.17)	-0.62 (0.03)	0.08 (0.03)	0.66 (0.03)	1.44 (0.05)	4.56 (0.17)	-0.63 (0.03)	0.08 (0.03)	0.66 (0.03)	1.44 (0.05)	4.61 (0.17)	-0.63 (0.03)	0.07 (0.03)	0.65 (0.03)	1.44 (0.05)
Item 13	5.18 (0.2)	-0.85 (0.04)	-0.03 (0.03)	0.53 (0.03)	1.37 (0.04)	5.21 (0.2)	-0.85 (0.04)	-0.03 (0.03)	0.53 (0.03)	1.37 (0.04)	5.18 (0.2)	-0.85 (0.04)	-0.03 (0.03)	0.53 (0.03)	1.37 (0.04)
Item 14	4.91 (0.19)	-0.75 (0.03)	-0.02 (0.03)	0.58 (0.03)	1.35 (0.04)	4.87 (0.19)	-0.76 (0.03)	-0.02 (0.03)	0.59 (0.03)	1.35 (0.04)	4.91 (0.19)	-0.76 (0.03)	-0.03 (0.03)	0.58 (0.03)	1.35 (0.04)

Table 15 lists theta estimation from the three calibration approaches. The three DIF groups, gender, race, and marital status had very close values in theta estimated, for example, for all items approach, the theta value on average were 0.0019, 0.0021 and 0.0019, respectively. The standard errors for the three DIF groups had same value of 0.1608. The theta values estimated for removed DIF item approach were slightly less than all item approach, the values were 0.0018, 0.0016 and 0.0017, respectively; and the standard error with values of 0.1647, 0.1629 and 0.1635 for the three DIF group were slightly higher than all item approach at 0.1608. For demographic specific approach, the theta values estimated were -0.0015, -0.0013 and -0.0015, respectively, which were lower than all items approach; the standard error were 0.1607, 0.1608 and 0.1608 for the three DIF groups, and the standard errors were about the same as all item approach.

Table 16 shows misfit calculation for the person fit estimation. The misfit percentage for all items approach ranged from 4.2% to 4.6%, removed DIF item approach ranged from 4.0% to 6.2, and demographic-specific approach was from 3.6% to 5.3%. The range was summarized from the three DIF groups, however, no evidence showed one approach performed better than the others with less person misfit.

Table 17 shows the test model fit. The overall model fit using M2 statistics was available for all items approach and remove DIF approach. The estimation of M2 indicated the model does not fit well ($p < 0.0001$) . However, the RMSEA computed from M2 statistics was ranged from 0.05 to 0.08 that indicated acceptable model fit. The AIC and BIC showed that the ignore DIF item approach and demographic-specific approach were similar. As lower values indicating better fit, the result could be summarized as either remove DIF approach or demographic-specific approach have better relative model fit, however, the large AIC and BIC values were both relatively high for all three DIF group.

Partial Credit Model

DIF Analysis

Table 18 lists DIF analysis that applied Rasch-Welch test on all the PROMIS Fatigue items (without item 7). Of the 13 items, item 5 was detected from gender DIF group (p = 0.021); 7 of the 13 items showed DIF by race DIF group, such as item 1 through item 3 with p < 0.001; two items were found as DIF items by Marital Status DIF group, item 1 (p = 0.015) and item 6 (<0.001). The result is presented in Table 18.

Calibration using three different calibration approaches

By applying partial credit model to calibrate the item parameters, the three approaches were applied. The demographic-specific approach has combined item 5 and 6 as more problematic items from gender and marital status DIF groups. Result from race DIF group was not considered as too many DIF items were detected by Rasch-Welch test. Item fit statistics from three different calibration approaches for the combined DIF groups are presented in Table 20. The cut-score for mean-square fit was 1.3 for good fit, from 1.3 to 3.0 are considered as moderate misfit, and over 3.0 was extreme misfit. Compared to all items (ignore DIF item) approach, the percentage of misfit item were 15.38%, 0%, and 26.67%. By converting DIF items to demographic-specific items, the item fit was not improved.

Table 20 to Table 21 and Figure represents person fit related statistics. The mean theta estimation and mean person fit for demographic-specific approach were -1.0052 and 0.5953, which were closed to all item approach that had mean theta estimation and mean person fit of -1.0025 and 0.5952 (Figure 8). The percentage of the perfect for the three calibration approaches were similar at 78% (Figure 9). The standard error estimated for all item approach and demographic-specific approach were similar, the remove DIF approach had the highest standard error (Figure 10 and Figure 11).

Table 15. Person Score Estimated

	Ignore DIF		Removed DIF		Demographic-specific	
DIF Group	Theta	SE	Theta	SE	Theta	SE
Race	0.0019	0.1608	0.0018	0.1647	-0.0015	0.1607
Marital Status	0.0021	0.1608	0.0016	0.1629	-0.0013	0.1608
Gender	0.0019	0.1608	0.0017	0.1655	-0.0015	0.1608

DIF = differential item function; SE = standard error.

Table 16. Person Score Fit Statistics

		Misfit (%)	
DIF group	Ignore DIF	Removed DIF	Demographic-specific
Gender	76 (4.2%)	107 (5.9%)	65 (3.6%)
Marital Status	82 (4.6%)	72 (4.0%)	71 (4.0%)
Race	76 (4.2%)	112 (6.2%)	96 (5.3%)

DIF = differential item function.
Note: The values reported are unit-normal deviates and 0.05% 2-sided significance corresponds to 1.96.

Table 17. Test Model Fit

	M_2	df	p-value	RMSEA (95% CI)	AIC	BIC
Race						
All items	204.4135	26	<0.0001	0.0635 (0.0556, 0.0717)	43701.71	44059.21
Removed DIF items	103.4204	18	<0.0001	0.0526 (0.0430, 0.0627)	40323.89	40653.88
Demographic-specific grouping	–	–	–	–	43659.92	44044.92
Marital Status						
All items	202.7233	26	<0.0001	0.0634 (0.0554, 0.0717)	43314.20	43671.01
Removed DIF items	154.1735	18	<0.0001	0.0668 (0.0573, 0.0768)	39744.26	40073.62
Demographic-specific grouping	–	–	–	–	43293.57	43677.83
Gender						
All items	204.4135	26	<0.0001	0.0635 (0.0556, 0.0717)	43701.71	44059.21
Removed DIF items	186.4424	18	<0.0001	0.0740 (0.0646, 0.0838)	40333.17	40663.17
Demographic-specific grouping	–	–	–	–	43699.80	44084.80

Table 18. DIF Analysis of PROMIS Fatigue using Partial Credit Model

PROMIS Fatigue	Gender		Race		Marital Status	
Item	X^2 (df)	p-value	X^2 (df)	p-value	X^2 (df)	p-value
Item 1	0.58 (1)	0.445	21.25 (1)	<0.001	8.42 (2)	0.015
Item 2	0.00 (1)	1.000	54.19 (1)	<0.001	2.42 (2)	0.294
Item 3	0.00 (1)	1.000	16.23 (1)	<0.001	0.77 (2)	0.680
Item 4	3.71 (1)	0.054	0.70 (1)	0.403	1.70 (2)	0.424
Item 5	5.34 (1)	0.021	8.94 (1)	0.003	1.09 (2)	0.577
Item 6	1.40 (1)	0.237	1.89 (1)	0.169	30.52 (2)	<0.001
Item 8	1.84 (1)	0.175	11.66 (1)	0.001	1.67 (2)	0.430
Item 9	0.00 (1)	1.000	0.00 (1)	1.000	0.94 (2)	0.623
Item 10	0.00 (1)	1.000	4.14 (1)	0.042	0.33 (2)	0.851
Item 11	1.54 (1)	0.214	0.00 (1)	1.000	0.01 (2)	0.995
Item 12	0.00 (1)	1.000	0.00 (1)	1.000	0.01 (2)	0.997
Item 13	0.59 (1)	0.441	1.75 (1)	0.186	0.28 (2)	0.872
Item 14	0.31 (1)	0.579	4.96 (1)	0.026	0.02 (2)	0.989

DIF = differential item functioning; PROMIS = Patient-Reported Outcomes Measurement Information System.

Table 19. Item level fit statistics with Gender and Race DIF Groups

	Ignore DIF		Remove DIF		Demographic-specific	
PROMIS Fatigue Item	Infit	Z-Standardized	Infit	Z-Standardized	Infit	Z-Standardized
Item 1	1.04	1.23	1.1	2.93	1.04	1.24
Item 2	1.17	4.49	1.36	9.08	1.17	4.49
Item 3	1.01	0.31	1.12	3.42	1.01	0.31
Item 4	0.89	-3.09	1.02	0.67	0.89	-3.1
Item 5	1.45	9.90	—	—	—	—
Item 5 (Female)	—	—	—	—	1.46	9.65
Item 5 (Male)	—	—	—	—	1.39	5.03
Item 6	1.72	9.90	—	—	—	—
Item 6 (Married)	—	—	—	—	2.14	3.88
Item 6 (Others)	—	—	—	—	1.71	9.90
Item 8	1.12	3.42	1.28	7.31	1.12	3.44
Item 9	0.76	-7.34	0.82	-5.49	0.76	-7.33
Item 10	0.73	-8.28	0.77	-6.99	0.74	-8.27
Item 11	0.66	-9.9	0.69	-9.66	0.66	-9.9
Item 12	0.84	-4.76	0.97	-0.74	0.84	-4.74
Item 13	0.78	-6.61	0.84	-4.76	0.78	-6.59
Item 14	0.78	-6.59	0.89	-3.30	0.78	-6.60

DIF = differential item function.

Table 20. Person fit Statistics

	Ignore DIF		Removed DIF		Demographic-specific	
	Theta	Infit	Theta	Infit	Theta	Infit
Combine DIF Group	-1.0025	0.5952	-0.7691	0.6712	-1.0052	0.5953

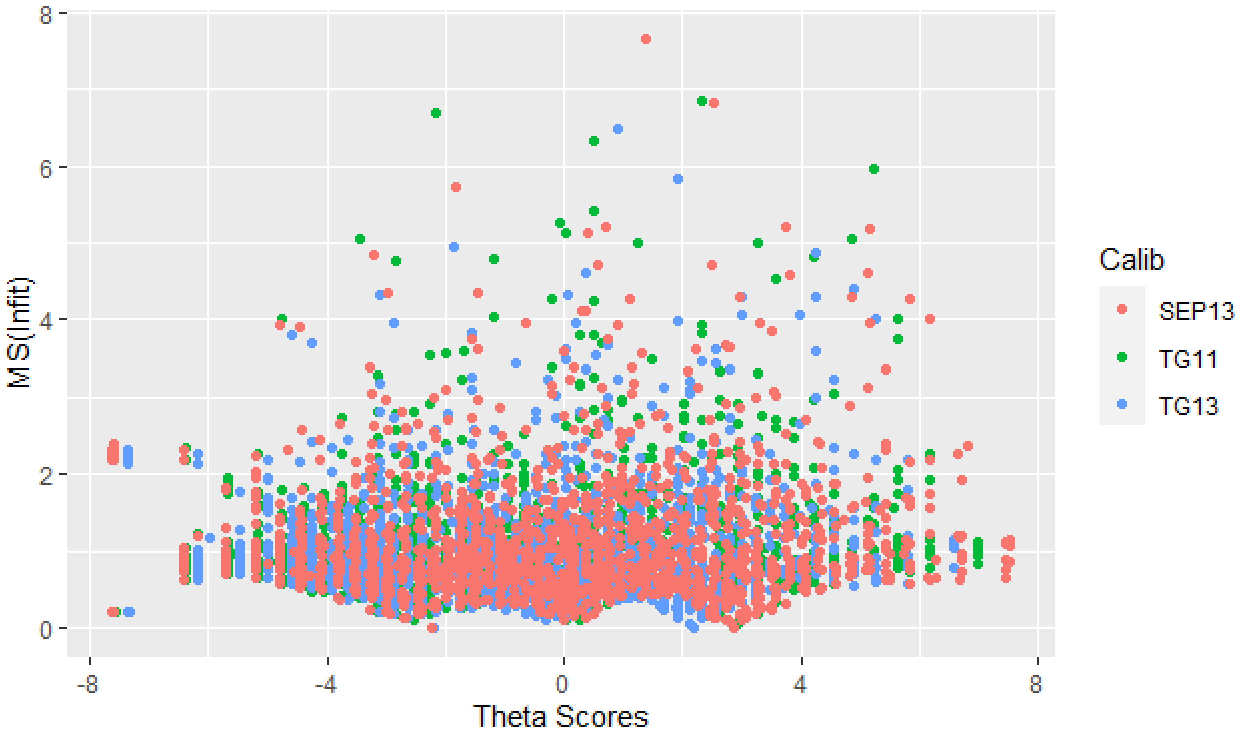
DIF = differential item function.

Table 21. Percentage of Person Fit for the Three Calibration Approach

Fit Category	Ignore DIF	Remove DIF	Demographic-specific
Good Fit	1330 (77.87%)	1327 (78.34%)	1329 (77.81%)
Moderate Misfit	337 (19.73%)	328 (19.36%)	337 (19.73%)
Extreme Misfit	41 (2.40%)	39 (2.30%)	42 (2.46%)

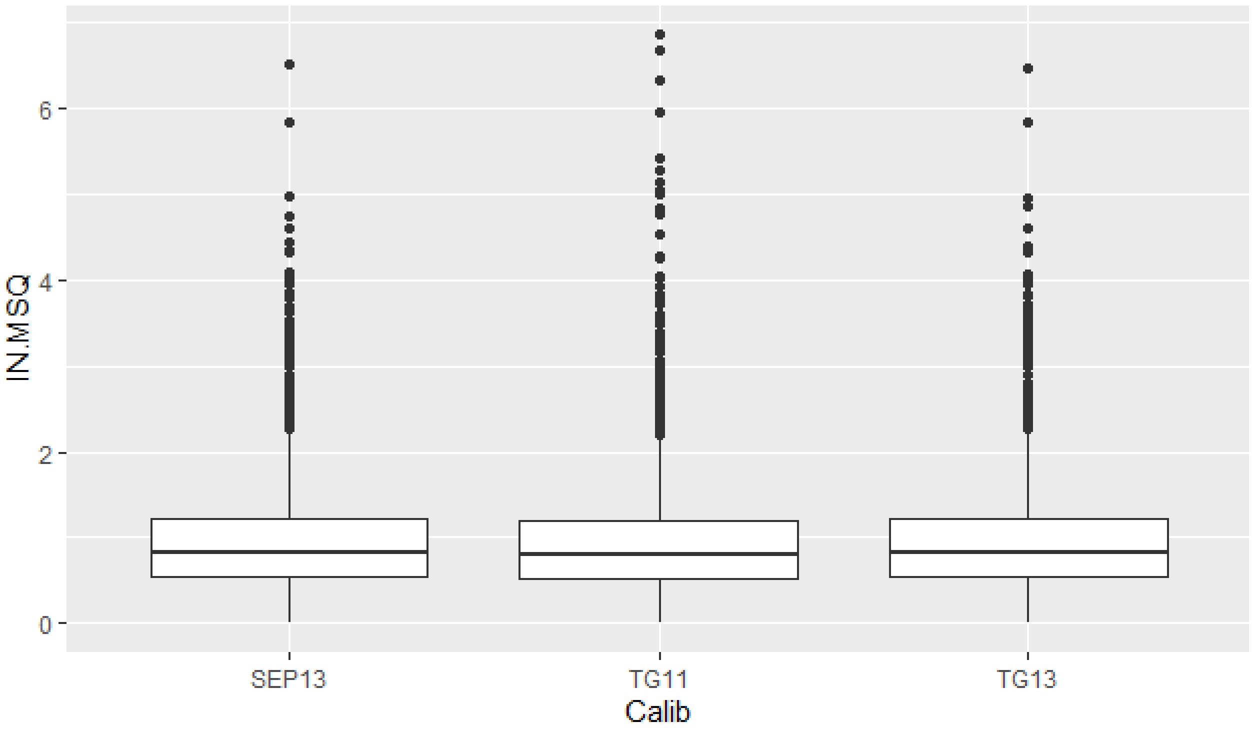
DIF = differential item function.

Figure 8. Scatterplot of Theta Scores and Person Score Misfit



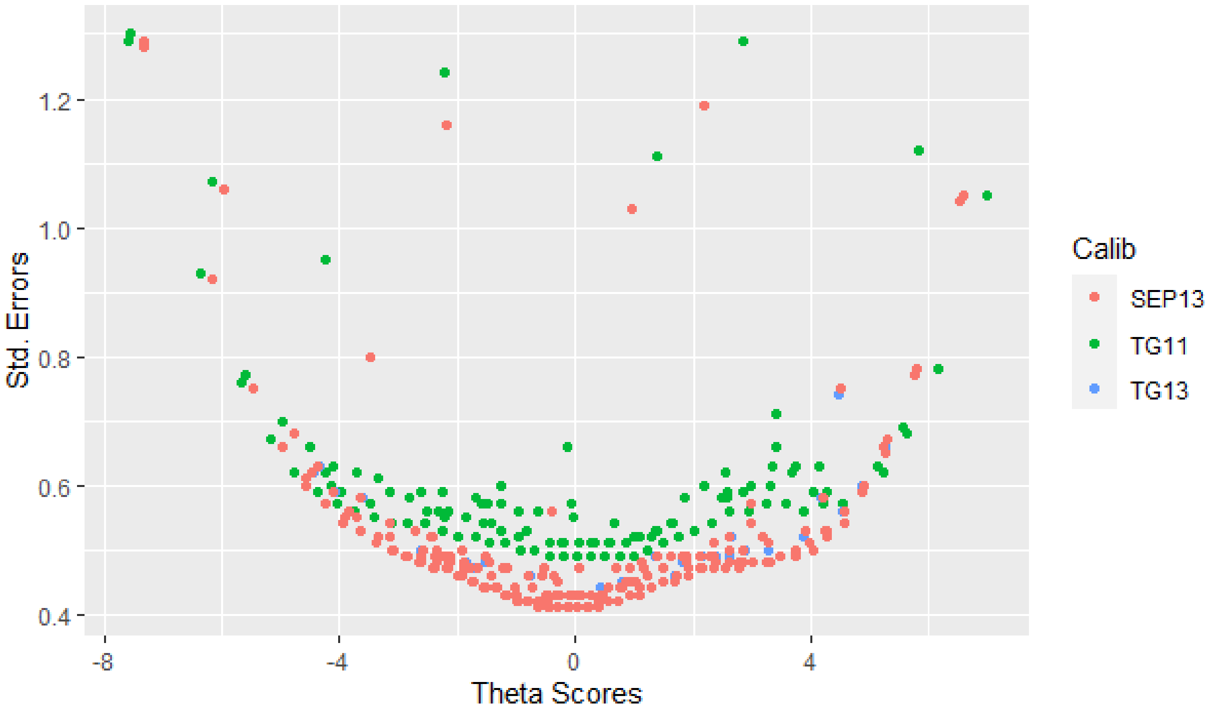
Calib = calibration; DIF = differential item function; MS = mean square; SEP13 = demographic-specific groups; TG11 = remove DIF approach; TG13 = ignore DIF approach.

Figure 9. Boxplot of Infit for the Three Calibration Approaches



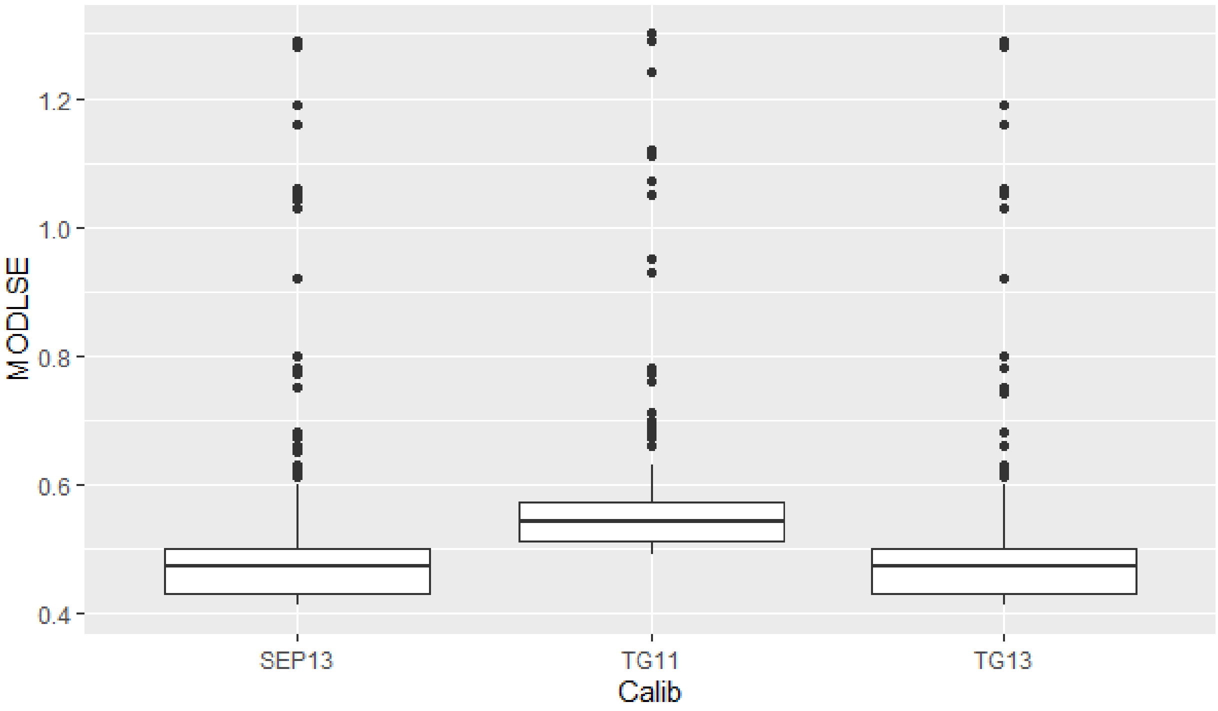
Calib = calibration; DIF = differential item function; IN.MSQ = infit mean square; SEP13 = demographic-specific groups; TG11 = remove DIF approach; TG13 = ignore DIF approach.

Figure 10. Scatterplot of Theta Scores and Standard Error



Calib = calibration; DIF = differential item function; Std = standard; SEP13 = demographic-specific groups; TG11 = remove DIF approach; TG13 = ignore DIF approach.

Figure 11. Boxplot of Model Standard Error for the Three Calibration Approaches



Calib = calibration; DIF = differential item function; MODLSE = model standard error; SEP13 = demographic-specific groups; TG11 = remove DIF approach; TG13 = ignore DIF approach.

CHAPTER V: CONCLUSIONS

Summary

DIF Analysis

DIF items have been detected from the PROMIS Fatigue measures in MY-Health study using both Wald test for GRM and Welch test for PCM. From the result of Wald test, one item was detected from the PROMIS Fatigue measure using each of the three anchors: gender, race, and marriage status. The DIF items by anchor were:

- Gender (Male vs Female) DIF in Item 8 [How often did you have to push yourself to get things done because of your fatigue?]
- Race (White vs Others) DIF in Item 2 [How often did you experience extreme exhaustion?]
- Marriage Status (Married vs Other) DIF in Item 6 [How often were you too tired to take a bath or shower?]

These anchors were selected from patient demographic variables. DIF exists widely in this highly diverse population. For example, five items (Item 2, Item 4, Item 5, Item 9, and Item 11) were detected as DIF items by age group (21 – 64 years old vs 65 years and older). The study was designed to have less than three DIF on each anchor item to enter the game room for conducting IRT calibration and comparison to demonstrate the advantage of demographic-specific group approach. However, the demographic-specific group approach to DIF reflects legitimate socio-cultural or demographic differences in the way groups interpret and respond to an item. Legitimate DIF should be driven by measurement theory, not exploratory DIF detection in practice and operation.

Before applying DIF analysis, it was essential to ensure that the scale was unidimensional in terms of accounting for most of the non-random residual covariance. CFA had been conducted to confirm the dimensionality. Even the fit indices of the 2-factor solution were more favorable than the 1-factor solution, with the inter-factor correlation of 0.965, a unidimensional structure was confirmed. In addition, item 7 [How often did you have enough energy to exercise strenuously?] had extremely low factor loading, by reviewing the item level statistics and response frequency found on the scale (1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Always), the higher scores indicated better outcome, which was found in opposite direction compared with the other items. The inter-item correlations between item 7 and all other items were less than 0.1. Based on this evidence, item 7 may not be appropriate to be administered to this cancer patient population and was dropped from this customized PROMIS Fatigue measure for analysis.

Item fit

Item calibration for the 13-item PROMIS Fatigue measure by GRM indicated misfits in Item 1 [How often did you feel tired?]. The result focuses on RMSEA for items, not the significance of the χ^2 values that were impacted by sample size. For the other two calibration approaches, removing the DIF item resulted in a new misfit item, item 3 [How often did you feel tired?] the race DIF group. However, by creating the demographic-specific groups for calibration, the item fit for this only misfit item was improved. In this case, demographic-specific group approach demonstrated great advantage over either the ignore DIF or the removed DIF approach to item fit.

For item calibration by PCM, the three anchors used in GRM, gender, race, marriage status were directly applied to PCM. The DIF items detected by the Welch test were different.

Only one set of calibrations was conducted on the three approaches. For removed DIF item approach, it also found a new misfit item, which was similar to the finding from GRM. The demographic-specific group approach contained two demographic-specific items, item 5 for male and female, and item 6 for married and other. As two DIF items, item 5 and item 6 showed moderate misfits, and the new demographic-specific variables were in the same situation. Refer to the item statistics, item 5 and item 6 had floor effect or significant skewness. Even though PCM had relaxed a-parameter that item fit was easier than any other models, the condition of not having all the categories being used caused the items so hard to fit into the model. A potential solution is to collapse the inefficiency categories on the poor fit items. Thus, demographic-specific group approach does not solve all the misfit issues, especially for any “bad” item.

Person fit

Among the three calibration approaches for both models, the demographic-specific group approaches had the smallest standard error, and the removed DIF item approach has the largest standard error. The percentage of misfit person score was similar to the three calibration approaches. The reliability under the IRT framework is conceptualized as “information” that related to measurement precision accounts for theta at different levels. The small standard error from the demographic-specific group approach would be increasing the scale of information and getting greater measurement precision, which impacts clinical decisions at individual level. The impact on group level is cumulated, may not be significant if the patients are randomly assigned to a treatment group, but it depends on different conditions. If more items are affected by DIF in the test and have a larger magnitude of DIF in the same direction, the latent score estimation will be impacted. If the sample size is small in clinical trials, it may need to increase power for DIF detection.

For applying demographic-specific group approach in clinical studies, it has to be driven by measure theory and expert evaluation on the benefit, and make sure this approach meets the requirements by policy, appropriate communication and careful decisions should be made if plan to use this method in clinical studies.

Limitation

It is a challenge to deal with measure equivalence issues in existing instruments on existing demographic variables. The My-Health study offers the opportunity of looking at this method and providing real world evidence. This is different from the simulated data with known conditions of the data, such as the type of DIF, magnitude of DIF, the ratio of DIF items to the total test items. Instead, the condition of DIF using a real world dataset has to be explored by research. The result may vary on another instrument and patient population. The intent of this study was to show the methodology of applying a demographic-specific group approach, but not to modify this existing instrument.

REFERENCES

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29, 67.

Ackerman, T. A., & Evans, J. A. (1994). The influence of conditioning scores in performing DIF analyses. Applied Psychological Measurement, 18(4), 329-342.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. Applied Measurement in Education, 7(4), 255-278.

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. Applied Psychological measurement, 20(4), 311-329.

Ackerman, T. (1996). Developments in Multidimensional Item Response Theory. Applied Psychological Measurement, 20(4), 309-10.

Ackerman, T. A. (2005). Multidimensional item response theory modeling. Contemporary psychometrics, 3-26.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). Standards for educational and psychological testing. American Educational Research Association.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43(4), 561-573.

Bevans, M.F., Ross, A., & Cella, D. (2014). Patient-Reported Outcomes Measurement Information System (PROMIS): efficient, standardized tools to measure self-reported health and quality of life. *Nursing outlook, 62 5*, 339-45 .

Bjorner JB, Kreiner S, Ware JE, Damsgaard MT, Bech P: Differential item functioning in the Danish translation of the SF-36. J Clin Epidemiol 1998, 51: 1189–1202. 10.1016/S0895-4356(98)00111-5

Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician. Adv Ther. 2018 Nov;35(11):1763-1774.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37(1), 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), 443–459.

Camilli, G. (2006). Test fairness. Educational measurement, 4, 221-256.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. Journal of Educational and Behavioral Statistics, 24(4), 323-341. 1984

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B.B., ... & Rose, M. (2007).The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Medical Care, 45(5 Suppl 1), S3.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., ... & PROMIS Cooperative Group. (2010). The Patient-Reported Outcomes Measurement Information System

(PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. Journal of Clinical Epidemiology, 63(11), 1179-1194.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48(1), 1-29.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. Journal of Educational Measurement, 33(3), 333-353.

Cho, S., Suh, Y., & Lee, W. (2016). After Differential Item Functioning Is Detected: IRT Item Calibration and Scoring in the Presence of DIF. Applied Psychological Measurement, 40(8), 573–591.

Clauser, B. E., & Mazor, K. M. (1998). Using Statistical Procedures To Identify Differentially Functioning Test Items. An NCME Instructional Module. Educational Measurement: issues and practice, 17(1), 31-44.

Crane PK, Gibbons LE, Jolley L, et al. Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. Med Care. 2006;44(Suppl 3):S115–S123.

Crane, P. K., Cetin, K., Cook, K. F., Johnson, K., Deyo, R., & Amtmann, D. (2007). Differential item functioning impact in a modified version of the Roland–Morris Disability Questionnaire. Quality of Life Research, 16(6), 981-990.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. Applied Psychological Measurement, 22(1), 33-51.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel--Haenszel method. Applied Measurement in Education, 2(3), 217-233.

Drasgow F., Levine M. V., William E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *Br. J. Math. Stat. Psychol.* 38, 67–86. 10.1111/j.2044-8317.1985.tb00817.x

Educational Testing Service. (2015). ETS Standards for Quality and Fairness. Educational Testing Service.

Embretson, S. E., & Reise, S. P. (2013). Item response theory. Psychology Press.

EMA. Update on real world evidence data collection. 10 March 2016. https:/ec.europa.eu/health/sites/health/files/files/committee/stamp/2016-03_stamp4/4_ real_world_evidence_ema_presentation. pdf. Accessed 08 Oct 2021.

FDA. Developing a framework for regulatory use of real-world evidence; Public Workshop. https:// www.gpo.gov/fdsys/pkg/FR-2017-07-31/pdf/2017-16021.pdf. Accessed 08 Oct 2021.

Fehnel, S., DeMuro, C., McLeod, L., Coon, C., & Gnanasakthy, A. (2013). US FDA patient-reported outcome guidance: great expectations and unintended consequences. Expert Review of Pharmacoeconomics & Outcomes Research, 13(4), 441-446.

Hays RD, Sherbourne CD, Mazel RM. The RAND 36-Item Health Survey 1.0. Health Econ. 1993 Oct;2(3):217-27. doi: 10.1002/hec.4730020305. PMID: 8275167.

Fayers, P. M., & Machin, D. (2015). Quality of life: The assessment, analysis and reporting of patient-reported outcomes. John Wiley & Sons.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. Journal of Educational Measurement, 33(3), 315-332.

Hambleton, R. K., & Swaminathan, H. (1985). A Look at Psychometrics in the Netherlands.

Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Instructional Topics in Educational Measurement. 1993:38–47.

Hays RD, Sherbourne CD, Mazel RM. The RAND 36-Item Health Survey 1.0. Health Econ. 1993 Oct;2(3):217-27. doi: 10.1002/hec.4730020305. PMID: 8275167.

Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011 Dec;20(10):1727-36. doi: 10.1007/s11136-011-9903-x. Epub 2011 Apr 9. PMID: 21479777; PMCID: PMC3220807.

Holland, P. W., & Wainer, H. (2012). Differential item functioning. Routledge.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum Associates, Inc.

Jensen R.E., Potosky A.L., Reeve B.B., et al. Validation of the PROMIS Physical Function Measures in a Diverse U.S. Population-Based Cohort of Cancer Patients. Quality of life

Research: an international journal of quality of life aspects of treatment, care and rehabilitation. 2015;24(10):2333-2344. doi:10.1007/s11136-015-0992-9.

Jensen, R.E., Moinpour, C.M., Keegan, T.H., Cress, R.D., Wu, X., Paddock, L.E., Stroup, A.M., & Potosky, A.L. (2016). The Measuring Your Health Study: Leveraging Community-Based Cancer Registry Recruitment to Establish a Large, Diverse Cohort of Cancer Survivors for Analyses of Measurement Equivalence and Validity of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Short Form Items. *Psychological test and assessment modeling, 58*, 99.

Johnson CD, Wicks MN, Milstead J, Hartwig M, Hathaway DK: Racial and gender differences in quality of life following kidney transplantation. Image J Nurs Sch 1998, 30: 125–130.

Joint Committee on Testing Practices. (2008). Code of fair testing practices in education. Encyclopedia of Special Education, 476-479.

Kang T., Cohen A.S., Sung H-J., Model Selection Indices for Polytomous Items. Applied Psychological Measurement. 2009;33(7):499-518. doi:10.1177/0146621608327800

Kim, S. H., Cohen, A. S., & Kim, H. O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. Applied Psychological Measurement, 18(3), 217-228.

Kim, S. H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. Applied Psychological Measurement, 22(4), 345-355.

Kolen, M. J., Brennan, R. L., & Kolen, M. J. (2004). Test equating, scaling, and linking: Methods and practices (pp. 177-180). New York: Springer.

Linacre, J.M. (2022). Winsteps® (Version 5.2.2) [Computer Software]. Portland, Oregon: Winsteps.com. Available from https://www.winsteps.com/

Linacre, J. M. (2022, March) DIF - DPF - bias - interactions concepts, Retrieved on March 15, 2022 from www.winsteps.com/winman/webpage.htm

Liu, X., & Jane Rogers, H. (2022). Treatments of Differential Item Functioning: A Comparison of Four Methods. Educational and Psychological Measurement, 82(2), 225–253.

Lord, F. M. Applications of item response theory to practical testing problems. Routledge; 2012 Nov 12.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. Applied Psychological Measurement, 22(3), 224-236.

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. Applied Measurement in Education, 19(3), 189-202.

Luecht, R. M., & Hambleton, R. K. (2021). ITEM RESPONSE THEORY. The History of Educational Measurement: Key Advancements in Theory, Policy, and Practice.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149-174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In Handbook of modern item response theory (pp. 101-121). Springer, New York, NY.

Masters, G. N. (1988). The analysis of partial credit scoring. Applied Measurement in Education, 1(4), 279-297.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In Handbook of Modern Item Response Theory (pp. 101-121). Springer, New York, NY.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. Applied Psychological Measurement, 11(2), 161-173.

McLeod, L. D., Coon, C. D., Martin, S. A., Fehnel, S. E., & Hays, R. D. (2011). Interpreting patient-reported outcome results: US FDA guidance and emerging methods. Expert Review of Pharmacoeconomics & Outcomes Research, 11(2), 163-169.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17(4), 297-334.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. Journal of Educational Measurement, 30(2), 107-122.

Muthén, L. K., & Muthén, B. O. (1998-2021). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. ETS Research Report Series, 1992(1), i-30.

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: A unified framework. Journal of the American Statistical Association, 100(471), 1009-1020.

Maydeu-Olivares, A. & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. Psychometrika, 71, 713-732.

National Institutes of Health. (2017). PROMIS Instrument Development and Validation Scientific Standards. Bethesda, MD: National Institutes of Health.

Penfield, R. D., & Algina, J. (2003). Applying the Liu‐Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. Journal of Educational Measurement, 40(4), 353-370.

Penfield, R. D., & Camilli, G. (2006). 5 Differential Item Functioning and Item Bias. Handbook of statistics, 26, 125-167.

Potosky, A. L.; Moinpour, C., 2016, "PROMIS 2 MY Health", https://doi.org/10.7910/DVN/XD1A6B, Harvard Dataverse, V1, UNF:6:No/Ha2bxUBEO7nsiGeazsg== [fileUNF]

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53(4), 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. Applied Psychological Measurement, 14(2), 197-207.

Rasch, G. (1980). Probabilistic model for some intelligence and achievement tests. Chicago, IL: University of Chicago Press.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., ... & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks:

plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical care, S22-S31.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika monograph supplement.

SAS Institute Inc. 2015. SAS/IML® 14.1 User's Guide. Cary, NC: SAS Institute Inc.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika, 58(2), 159-194.

Schmeiser, C. B., Geisinger, K. F., Johnson-Lewis, S., Roeber, E. D., & Schafer, W. D. (1995). Code of professional responsibilities in educational measurement. Washington, DC: National Council on Measurement in Education.

Smets, E. M., Garssen, B., Cull, A., & de Haes, J. C. (1996). Application of the multidimensional fatigue inventory (MFI-20) in cancer patients receiving radiotherapy. British journal of cancer, 73(2), 241–245. https://doi.org/10.1038/bjc.1996.42

Sprangers M.A.G., Bonnetain F. (2014) EORTC QLQ-C30. In: Michalos A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_901

Steiger, J. H., & Lind, J. C. (1980). Statistically-based tests for the number of common factors. Paper presented at the annual Spring Meeting of the Psychometric Society in Iowa City. May 30, 1980.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational measurement, 27(4), 361-370.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. Organizational Research Methods, 18(1), 3-46.

Tendeiro JN, Meijer RR, Niessen ASM (2016). "PerFit: An R Package for Person-Fit Analysis in IRT." *Journal of Statistical Software*, **74**(5), 1–27.

Teresi JA, Kleinman M, Ocepek-Welikson K, Ramirez M, Gurland B, Lantigua R, Holmes D: Applications of item response theory to the examination of the psychometric properties and differential item functioning of the Comprehensive Assessment and Referral Evaluation dementia diagnostic scale among samples of Latino, African-American, and White non-Latino elderly. Research on Aging 2000, 22: 738–773.

Teresi JA, Ramirez M, Lai JS, Silver S. Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. Psychol Sci Q. 2008;50(4):538.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika, 51(4), 567-577.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In Test scoring (pp. 85-152). Routledge.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 147–172). Lawrence Erlbaum Associates, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 67–113). Lawrence Erlbaum Associates, Inc.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. (1995). Item response theory for scores on tests including polytomous items with ordered responses. Applied Psychological Measurement, 19(1), 39-49.

U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, & U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health (2006). Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health and quality of life outcomes, 4, 79. https://doi.org/10.1186/1477-7525-4-79

Van Der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In Handbook of modern item response theory (pp. 1-28). Springer, New York, NY.

Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. Applied Measurement in Education, 6(1), 1-19.

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Yen, W. M., Fitzpatrick, A. R., & Brennan, R. L. (2006). Educational measurement.

Zieky, M. J. (2013). Fairness review in assessment. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology (pp. 293–302). American Psychological Association. https://doi.org/10.1037/14047-017

Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand. 1983 Jun;67(6):361-70. doi: 10.1111/j.1600-0447.1983.tb09716.x. PMID: 6880820.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. Journal of Educational Measurement, 30(3), 233-251.

Zwick, R., & Thayer, D. T. (1996). Evaluating the Magnitude of Differential Item Functioning in Polytomous Items. Journal of Educational and Behavioral Statistics, 21(3), 187–201. https://doi.org/10.3102/10769986021003187