Differential Item Functioning within the Dutch and Flemish PROMIS Pain Behavior and Interference Item Banks: An Empirical Comparison of IRT and CFA Approaches

Master's Thesis

Tihana Okicki

Master's Thesis Methodology and Statistics Master
Methodology and Statistics Unit, Institute of Psychology,
Faculty of Social and Behavioral Sciences, Leiden University
Date: September 14th, 2018
Student number: s1903853
Supervisor: Dr. Fokkema (internal), Dr. Terwee (external)

Table of Contents

Abstract

This study investigated the application of two major approaches in testing for differential item functioning (DIF), or measurement invariance (MI), on Dutch-Flemish PROMIS datasets on Pain Behavior and Pain Interference item banks, developed to be used in computerized adaptive testing (CAT). The Pain Behavior item bank consists of 39 items, and the Pain Interference item bank of 40 items. Both are measured on a six-point Likert scale, although participants who had no pain at baseline (category 1) in the Pain Behavior item bank were excluded from the analysis, which effectively resulted in a five-point scale. We applied item response theory (IRT) and confirmatory factor analysis (CFA) models on responses of approximately 6600 participants from the general population, chronic pain patients, and rheumatoid arthritis patient groups, with the goal of determining whether items exhibited different measurement properties across groups. Through the IRT approach, we found four items, out of a total of 79, showing uniform DIF with minimal impact on theta scores in three comparisons. Through the CFA approach, all items were found to be measurement invariant across groups using ΔCFI being equal to or less than 0.002 as a measure of goodness of fit. As items flagged for DIF had minimal impact on theta scores, it is likely there may not be a significant impact on theta estimates in CAT if items were used without added calibration. We have found that the IRT and CFA approaches show results with a high degree of similarity.

Introduction

PROMIS stands for Patient Reported Outcome Measurement Information System. It is a set of self-report measures, which aim to evaluate patients' health outcomes. The PROMIS initiative consists of multiple item banks testing numerous physical, mental and social health components. Over the past 15 years, PROMIS has been developed in the US as a practical self-reporting test, where physicians and other health care providers could use those test score results to compare them over time, and to further monitor the level of physical, social or mental health in both adults and children (by using PROMIS developed pediatric measures). PROMIS tests aim to be suitable for both the general population and individuals with specific chronic conditions (Cella et al, 2010).

PROMIS item banks have been developed specifically to be applied in a computerized adaptive testing (CAT) setting. CAT is a specialised method of test delivery, where the goal is to estimate a person's latent trait level as accurately as possible, through administering as few items as needed. In CAT, questions are selected by the computer based on a respondent's current estimated latent trait value, which is based on the respondent's answers to previous questions. There is a principal assumption that item parameters are equal across groups. If item parameters differ between groups, CAT algorithms cannot use the same population-level item parameters for all respondents, as this would yield biased estimates of respondents' position on the latent trait. As CAT algorithms base selection of the next item to be administered on the current estimate of a respondent's position on the latent trait, the biased estimates would negatively affect both item selection as well as trait estimation, making the consequences of differential item functioning in CAT potentially more severe than in traditional test administration (Fayers, 2007).

Although CAT has long been used in the field of educational testing, its application has been lacking in the health and medical field, where currently most questionnaire testing

instruments in use have been developed based on classical test theory (CTT). CTT based instruments require to be completed in full, which can make administration and assessment time consuming, expensive, and often impractical. Unlike in traditional CTT questionnaire administration, administering a CAT questionnaire requires only a minimal number of answered items (Fayers, 2007). CATs substantially reduce the number of items needed to be answered in order to get as accurate as possible estimate of health outcomes (Fries *et al*, 2005; Reeve *et al*, 2007). Provided that sufficient testing, development and parameter calibration have taken place on an item bank, the end result is a self-report measure of very low cost and high ease of delivery (Revicki *et al*, 2009).

Due to CAT's potential benefit for the health care system, PROMIS item banks have been translated to other languages and have been further adapted in numerous countries by researchers working on a language specific PROMIS project. The approach to translation is universal, ensuring that there is only one version for each language, instead of a specific version for each country using the same language. As such, the Dutch version is a product of a Dutch-Flemish PROMIS project group. Researchers in the Netherlands and Flanders have translated the PROMIS item banks and have been thoroughly testing them on Dutch and Flemish populations (Terwee *et al*, 2014). Two such item banks are dedicated to measuring Pain Behavior and Pain Interference.

PROMIS Pain Item Banks

Pain behavior refers to types of pain related behaviors that can range from verbal complaints about experiencing pain, to facial expressions, body posturing, and overall limitations in being able to participate in certain activities (Revicki *et al*, 2009). The most widely used pain behavior self-report measure that has been validated in empirical research, is the Pain Behavior Check List (PBCL) developed by Kerns *et al* (1991). The drawbacks of PBCL is that although it consists of 49 short worded self-report items about pain related

behaviors, it captures only a limited scope of pain behavior categories, and does not capture behaviors such as isolating oneself from others when in pain, massaging a painful area, or crying. The goal behind the PROMIS Pain Behavior item bank development was designing a self-report measure that could be integrated into clinical studies, and that could reliably assess a full scope of pain behaviors. The initial repository of pain behavior items was based on literature findings, clinical reviews and qualitative research conducted with patients experiencing various types of pain. Following review by research experts and further cognitive debriefing interviews, the final pain behavior item bank consisted of 52 items covering movement, social interaction, and behaviors of affective and verbal kind. (Revicki *et al*, 2009).

Similarly, a Pain Interference item bank has been developed by Amtmann *et al* (2010), which contains items assessing the degree to which pain interferes with a person's physical and mental wellbeing, and social activities. There exist quite a few short item tests measuring pain interference, such as a nine-item scale from the West Haven-Yale Multidimensional Pain Inventory (WHYMPI; Kerns *et al*, 1985), seven item scales from the Brief Pain Inventory (Daut *et al*, 1983) and the Pain Disability index (Pollard, 1984), a six item Pain Impact Questionnaire (PIQ-6; Becker *et al*, 2007), and a three-item scale from the Chronic Pain Grade (Von Korff *et al*, 1992). The Pain Interference item bank was developed from a "library" of pain interference items (n=644) that were identified through literature and feedback from patients experiencing pain. Following revision from expert researchers in domains of pain assessment, language translation and psychometrics, and additional review through interviews to evaluate item clarity, and appropriateness of the content, 56 items were chosen to constitute the final item bank. Having extensive item banks that can be calibrated for use in computer adaptive testing would curtail the number of questions down to only the necessary and the most relevant ones. (Amtmann *et al*, 2010).

Two Approaches: IRT and CFA

The aim of PROMIS is to develop self-reported measures that can be used across populations, but this requires the items to measure the same traits in the same way across different groups. Two frameworks allowing for empirical comparison of the measurement properties of items across groups are item response theory (IRT) and confirmatory factor analysis (CFA). Both IRT and CFA may also be referred to as latent trait theory, in which observed item scores are assumed to measure a continuous underlying latent trait or ability.

Item Response Theory

The three main assumptions of IRT are that the model is unidimensional, that all items are locally independent, and that a response given by a person to an item can be modelled mathematically by the item response function. This item response function gives the probability of selecting a response category, given the ability level (theta) of a person. In IRT, the primary goal is to test people and position them along a continuum of the latent trait (or ability) being measured. In addition, IRT models allow for positioning items based on their 'difficulty' on the same latent trait continuum as respondents, which in turn allows for selecting items that approximate a respondent's latent trait value as close as possible, as in CAT. Although a latent trait cannot be directly observed, IRT assumes it can be measured (Hambleton *et al*, 1991). If the value of the parameters characterizing the item response function differ between groups, we encounter differential item functioning (DIF). DIF analyses are used to examine if people from different groups with the same level of trait have different probabilities of selecting a certain response category of an item.

Confirmatory Factor Analysis

CFA is a particular form of factor analysis, which can be used to determine whether a set of observed variables (e.g., item or subscale scores) all measure the same underlying

latent variable, trait or factor (Brown and Moore, 2012; Fox, 1983). In the CFA approach, the association between observed item scores and latent trait(s) is characterized by several parameters: factor loadings, item thresholds, item residual variances. Equality of these parameters across groups are examined, in order to evaluate the presence of measurement invariance (MI). When assessing levels of invariance, we constrain each parameter to equality between groups. If this yields a significant decrease in model fit, equality constraints for certain items need to be released, and it is concluded that that specific item measures the latent variable differently across groups (Long, 1983). This procedure can be more specifically referred to as multi-group categorical CFA.

When dealing with ordered-categorical data, such as the PROMIS item banks, ordered-categorical CFA and IRT are comparable since both allow for testing the equality of thresholds (CFA) or difficulty parameters (IRT) (Kim and Yoon, 2011), but they use different estimation methods and identify the latent variable's scale differently, which may lead to discrepancies in detecting the lack of invariance.

Previous Research

Kim and Yoon (2011) have run a series of comparisons between multiple group categorical CFA and IRT with Monte Carlo generated data. They found that the higher the sample size, the better both methods performed in terms of detecting true positives. When looking at measurement invariance in factor loadings, CFA outperformed IRT when the degree of DIF was small; however, by accounting for false positive rates as well, the performance of CFA deteriorated, and IRT was shown to be more reliable. In this respect, IRT was proved to be better, especially when sample size and DIF degree were large. However, data distributions in empirical studies on mental and physical health may differ from those in simulation studies. Latent traits may show more skew or kurtosis than the normal distribution. Furthermore, measurement errors may be correlated or heteroscedastic,

all of which would violate the assumptions of the fitted model. The current study therefore aims to compare the performance of DIF and MI detection on empirical data.

As PROMIS and its item banks are still relatively novel, not many studies are available that specifically test for DIF between populations, as the majority of earlier studies were focused on calibration of the items (Reeve et al, 2007; Cella et al, 2010; Rose et al, 2014). Comparisons for age and gender have been performed by Amtmann et al (2010), where they have found in total nine items to have non-uniform DIF in the Pain Interference item bank. Revicki et al (2009) tested for age, gender and education related DIF in the Pain Behavior item bank. They have found one item was detected for gender, and five further items were detected for age related DIF, all of which were uniform. Furthermore, no study so far on PROMIS item banks has been performed comparing measurement invariance and differential item functioning testing.

Our main research goal is to test and compare the results of measurement invariance and differential item functioning analysis of the Dutch-Flemish PROMIS item banks of Pain Behavior and Pain Interference.

Regardless of the method (IRT or CFA) used, the outcome of the analysis would inform whether the items in the PROMIS item banks measure the same latent trait in the same way across groups. In other words, whether the items are suitable to be administered in CAT. An item in which parameters are not equal between groups, needs to be handled in a way to reduce bias. Depending on the level of DIF or a lack of MI for an item, the options can range from having separate parameters for given populations in question, to lowering the priority of the item as to reduce the likelihood of it appearing in the CAT, or removing it entirely.

Method

Datasets

In total, we have examined five datasets comprising of chronic pain patients (CP), general population (GP), and rheumatoid arthritis patients (RA). Details are summarized in Table 1.

Table 1. Datasets used in analyses.

Datasets	Item bank	n	Mean	Std. Dev.	Male %	Female %
Chronic Pain Reade[1]	PB	929	51.21	12.47	21.31	78.69
Chronic Pain Reade[1]	PI	973	51.66	12.60	22.05	77.95
Chronic Pain (clinics)[2]	PB	1595	47.40	13.73	42.13	58.87
Chronic Pain (clinics)[2]	PI	1650	47.44	13.80	41.50	58.50
General Population[3]	PB	783	51.93	14.69	43.30	56.70
General Population[4]	PI	1049	51.28	15.11	45.09	54.91
Rheumatoid Arthritis[5]	PB & PI	2144	58.44	12.67	31.16	68.84

Some datasets include both item banks, while others needed to be joined as they are representative of the same group, but participants were recruited from different institutes. Given that there are three separate groups and two item banks, and we examined two latent variable methods, we were able to make 12 group comparisons; six by performing DIF analysis between groups, and six by analyzing measurement invariance between those same groups. The datasets consisted of male and female participants in the age range from 18 to 94 years of age, of Dutch and Flemish background.

For both the Pain Behavior and Pain Interference item banks we started with two datasets of chronic pain patients. One cohort was collected at Reade (Pain Behavior n = 929; Pain Interference n = 973), an outpatient secondary care centre for rheumatology and

rehabilitation in the Netherlands, and the other was collected from patients registered at practices of 31 participating physicians specializing in muscoskeletal medicine in the Netherlands (Pain Behavior n = 1595; Pain Interference n = 1650). After both analyses resulted in no items flagged for DIF between the two groups with regards to recruitment location, the datasets were combined into a final chronic pain patient dataset (n=2524) for both the Pain Behavior and Pain Interference item banks.

Within the Pain Behavior item bank, analyses were run between the chronic pain patient dataset (n = 2524), a dataset representing the general population collected from Desan (n = 783), and the rheumatoid arthritis dataset (n = 1456) combined of both Dutch and Flemish participants. The Dutch cohort consisted of RA patients from the Amsterdam Rheumatoid Arthritis (AMS-RA) cohort who have been registered since 2000 in Reade. The Flemish cohort consisted of RA patients from the arthritis cohort from KU Leuven, Belgium. Having three groups, we performed 3 comparisons.

Similarly, within the Pain Interference item bank, we compared the chronic pain patients combined dataset (n = 2623), the general population dataset (n=1049), and the rheumatoid arthritis dataset (n=1917).

Measures/Instruments

The participants completed a paper-and-pencil or web-based survey which included the Dutch-Flemish versions of the Pain Behavior and Pain Interference item bank questionnaires. The Pain Behavior item bank consists of 39 items, which are all graded on a six-point Likert scale. The items are questions posed regarding different pain related behaviors that have taken place in the past seven days prior to answering the questionnaire. The possible answers range from Not having pain (1), Never (2), Rarely (3), Sometimes (4),

Often (5), and At all times (6). Participants who had no pain at baseline (1) were excluded from the analyses. This exclusion effectively resulted in a five-point Likert scale. The Pain Interference item bank consists of 40 items. There are three different 5-point Likert response scales depending on the item question phrasing:

A)	B)	C)
Not at all (1)	Never (1)	Never (1)
A little bit (2)	Rarely (2)	Once a week or less (2)
Somewhat (3)	Sometimes (3)	Once every few days (3)
Quite a bit (4)	Often (4)	Once a day (4)
Very much (5);	Always (5);	Every few hours (5).

IRT/DIF Analysis

To analyze polytomous items, as used in our instruments, the proportional odds logistic regression model is used (Agresti, 2007). For each item, there is an intercept only null model along with three nested models formed in hierarchy:

$$Model\ 0: logit\ P\ (u_i \geq k) = \alpha_k$$
$$Model\ 1: logit\ P\ (u_i \geq k) = \alpha_k + \beta_1 * ability$$
$$Model\ 2: logit\ P\ (u_i \geq k) = \alpha_k + \beta_1 * ability + \beta_2 * group$$
$$Model\ 3: logit\ P\ (u_i \geq k) = \alpha_k + \beta_1 * ability + \beta_2 * group + \beta_3 * ability * group$$

These nested models are built upon the null model (Model 0) by adding the effect of ability (trait, latent variable) in Model 1, the effect of group in Model 2, and finally, the interaction effect between ability and group in Model 3. Logistic regression is then applied to detect DIF, as described below (Swaminathan and Rogers, 1990). This IRT approach is executed in the R package *lordif* developed by Choi (2011, 2016), which specializes in DIF analysis through a combination of ordinal logistic regression and IRT principles (Meade & Lautenschlager,

2004). *lordif* achieves DIF analysis by testing each item through a data run in which all items' parameter estimates are constrained to be equal across groups, except for those of the item that is tested for DIF.

Once an item is flagged for DIF, its differential functioning can be described as uniform (Model 2), provided the effect is constant, or non-uniform (Model 3), if the effect shows variation depending on the trait level. Uniform DIF is the simplest of DIF types where one group would show a consistently higher likelihood to endorse a response category across the total range of trait levels. On the other hand, in non-uniform DIF, whether a group shows a higher or lower likelihood to endorse a response category depends on the level of the latent trait (Walker, 2011). In such case, a group can have, for example, a small advantage at the lower end and a major advantage at the higher end, or an advantage at the lower end and a disadvantage at the higher end. Non-uniform DIF occurs when there is an ability and group membership interaction (Model 3).

In general, uniform DIF for a given item is tested by comparing log likelihood values between Models 1 and 2 (Figure 1), and non-uniform DIF is tested by comparing log likelihood values between Models 2 and 3. Both of those comparisons have 1 degree of freedom (*df*). The difference between log likelihood values of Models 1 and 3 provides an overall test of 'total DIF effect', with 2 *df*. If this test yields a p-value below the pre-specified alpha level, DIF has been detected and we subsequently look towards specific significance in χ^2_{12} for uniform, and χ^2_{23} for non-uniform DIF (Walker *et al*, 2001; Jodoin *et al*, 2001).

The estimator used is maximum likelihood estimation (ML). ML is the most widely used estimation method as it is a normal theory estimator, where samples are assumed to be of an adequate size, observations are assumed to be independent, the model is correctly specified and data are multivariate normal and continuous. ML employs an iterative estimation process where differences between observed sample covariance matrix and the

implied model matrix are minimized (Mindrila, 2010). However, in non-normality conditions, ML can produce biased fit indices, inflated chi-squares, and underestimation of standard errors (Hoogland & Boomsma, 1998).

In IRT, the ability is generally assumed to be measured on a standard scale with a mean of 0 and standard deviation of 1. These specifications identify the scale for theta, and also the scale for item parameters (Stocking & Lord, 1983).

GRM model

In IRT, the graded response model is used to express the probability of choosing a response category given the persons' ability level. The GRM was developed by Samejima (1968) to extend a two-parameter logistic model, originally intended for dichotomous, to polychotomous items. In the GRM, each item has a single discrimination parameter and a threshold for all response categories minus one. A single-factor CFA on polychotomous ordinal items is equivalent to the graded response model (Samejima, 1969; Dodd et al, 1989).

CFA/MI Analysis

A one-factor CFA model can be described as a unidimensional model with assumed continuous latent response variates X_{ij}^{*}, that underlie the observed scores X_{ij}:

$$X_{ij}^{*} = \tau_j + \lambda_j \xi_i + \varepsilon_{ij}$$

where τ_j is the intercept of item *j*, λ_j is the factor loading of item *j*, ξ_i is person *i*'s common factor score, and ε_{ij} is a residual.

These latent response variates are manifested as discrete scores with a set of thresholds:

$$X_{ij} = c, \text{if } v_{jc} < X_{ij}^{*} \leq v_{j(c+1)}$$

with C being the number of categories, the number of thresholds is always one less (C-1). v_{jc} indicates the C ordered-categorical responses of the j^{th} item (Kim & Yoon, 2011).

In the multi-group CFA approach, we are also comparing three models between groups:

1) a 'configural fit' model, which tests whether the same factorial structure (i.e. pattern of zero and non-zero loadings) holds between groups,

2) a 'metric invariant' model, which in addition tests whether equality of factor loadings holds between groups, but allows for differences in thresholds between groups, and

3) a 'scalar invariant' model, which tests whether there are both equal thresholds and equal loadings between groups.

We have executed the CFA approach is steps in R package *lavaan* developed for latent variable modelling by Rosseel (2012).

Assessing model fit in CFA

Traditionally, the most commonly applied test of factorial invariance in CFA is the chi-square difference test $\Delta\chi^2$. Chi-square (χ^2) is a statistic that can be used to measure how much of a difference exists between observed and expected values. A χ^2 test is also referred to as a goodness of fit test and it determines whether the data matches the fitted model. The difference in χ^2 ($\Delta\chi^2$) tests whether a more complex model fits significantly better than a simpler model. The value of $\Delta\chi^2$ is calculated by subtracting the χ^2 value of the more complex model from that of the less complex model. To calculate the degrees of freedom (*df*) for the $\Delta\chi^2$ test (Δdf), we subtract the *df* from the more complex model from the *df* of the less complex model. If the $\Delta\chi^2$ (Δdf) test between the two models is statistically significant, then the more complex model is a better fit than the less complex model.

However, with large datasets, small differences in model fit are more likely to be significant (Bentler & Bonett, 1980). In order to avoid false positives, we will be relying on

model fit indices: comparative fit index (CFI), difference in CFI (ΔCFI), standardized root mean squared residual (SRMR), and root mean square error of approximation (RMSEA) (Chen, 2007; Cheung & Rensvold, 2002).

CFI analyzes the model fit by examining the discrepancy between the data and the hypothesized model, while adjusting for the issues of sample size inherent in the chi-squared test of model fit. CFI takes into account sample size that performs well even when samples size is small. In examining baseline comparisons, the CFI depends in large part on the average size of the correlations in the data. If the average correlation between variables is not high, then the CFI will not be very high. Values for this statistic range between 0.0 and 1.0 with values closer to 1.0 indicating good fit. A CFI value of 0.95 or higher is desirable and recognized as indicative of good fit. In using ΔCFI as a measure of goodness of fit, we rely on ΔCFI values between different model fits being equal to or less than 0.002 to be able to accept the null hypothesis of invariance (Meade *et al*, 2008).

SRMR is an absolute measure of fit and is defined as the standardized difference between the observed correlation and the predicted correlation. A value of .08 or smaller is a guideline for good fit, and a value of zero would indicate a perfect fit. RMSEA is a measure of goodness of fit for statistical models, where the goal is to have an approximate or close fit with the model, rather than an exact fit, which is often not practical for large populations. RMSEA tells us how well the model would fit the population covariance matrix. Values of 0.01, 0.05, and 0.08 indicate excellent, good, and mediocre fit, respectively. 0.10 is used as a cutoff point for poor fitting models. As such, lower values indicate a better fitting model. (Hu & Bentler 1998, 1999; Kenny et al. 2015). These measurement standards were designed with educational field in mind specifically, and thus in other fields, such as health, these fit standards may be difficult to achieve.

The estimator used for fitting the CFA models is diagonally weighted least squares (DWLS), which can deal with ordinal data. DWLS is a robust weighted least squares method (WLS), based on a polychoric correlation matrix of variables used in the analysis. DWLS uses only the diagonal of weights in inversion, and all weights in the estimation of fit and standard error (Li, 2016). This technique is used when performing analysis of items on self-report tests that use a Likert rating scale, exactly like PROMIS item banks, and can be used with small sample sizes, large models, as well as skewed and ordinal data. (Mindrila, 2010).

In CFA, the most common method of latent variable scaling is to use one marker variable, and to fix its loading to 1 for setting the scale. The same marker variable's intercept is fixed to 0. In this way the latent variable scale is related to the marker variable (Beaujean, 2012).

Comparing DIF and MI models

Of note, the configural invariance fit model is equivalent to Model 1 in IRT, which tests for the presence of a common ability factor. An item showing a lack of scalar invariance would be equivalent to the item being flagged for uniform DIF (Model 2). In a similar fashion, an item showing a lack of metric invariance would be equivalent to the item being flagged for a non-uniform DIF (Model 3).

Results

IRT/DIF

A total of four items were flagged for DIF, in three out of six comparisons. All these items were flagged for uniform DIF. The results are summarized in Table 2. Below, we will discuss the DIF found for the Pain Behavior and Pain Interference item banks, respectively.

Table 2. Differential Item Functioning Analysis Results for Pain Behavior and Pain Interference

PROMIS item banks	Groups Compared	Items with DIF	McFadden R^2 values	Pr	Parameter Slopes* and Thresholds	DIF Type
Pain Behavior	Rheumatoid Arthritis Patients (RA) vs General Population (GP)	PAINBE24 (item 14) – *"In the past 7 days... When I was in pain I moved stiffly..."*	R^2_{12} = .0260 R^2_{13} = .0261 R^2_{23} = 0	χ^2_{12} = 0 χ^2_{13} = 0 χ^2_{23} = .796	RA = 1.88*; -1.45, -0.53, 0.75, 2.25 GP = 1.57*; -0.63, 0.14, 1.29, 2.89	Uniform
		PAINBE25 (item 15) – *"In the past 7 days... When I was in pain, I called out for someone to help me..."*	R^2_{12} = .0213 R^2_{13} = .0323 R^2_{23} = .0011	χ^2_{12} = 0 χ^2_{13} = 0 χ^2_{23} = .014	RA = 1.54*; -0.41, 0.62, 2.15, 3.59 GP = 1.81*; 0.4, 1.28, 2.44, 3.72	Uniform
	Chronic Pain Patients (CP) vs General Population (GP)	PAINBE45 (item 33) – *"In the past 7 days... When I was in pain I thrashed..."*	R^2_{12} = .0225 R^2_{13} = .0230 R^2_{23} = .0005	χ^2_{12} = 0 χ^2_{13} = 0 χ^2_{23} = .131	CP = 2.08*; 1.29, 1.89, 2.64 GP = 1.97*; 0.7, 1.48, 2.43	Uniform
Pain Interference	Chronic Pain Patients (CP) vs General Population (GP)	PAININ20 (item 16) – *"In the past 7 days... How much did pain feel like a burden to you..."*	R^2_{12} = .0238 R^2_{13} = .0238 R^2_{23} = 0	χ^2_{12} = 0 χ^2_{13} = 0 χ^2_{23} = .646	CP = 2.84*; -1.62, -0.61, 0.07, 1.23 GP = 2.8*; -1.23, -0.12, 0.53, 1.64	Uniform

Pain Behavior Item Bank

In the pain behavior item bank, items PAINBE24 and PAINBE25 were flagged for DIF for the comparison between the general population and the rheumatoid arthritis patient group, and item PAINBE45 for the comparison between the general population as the chronic pain patient group.

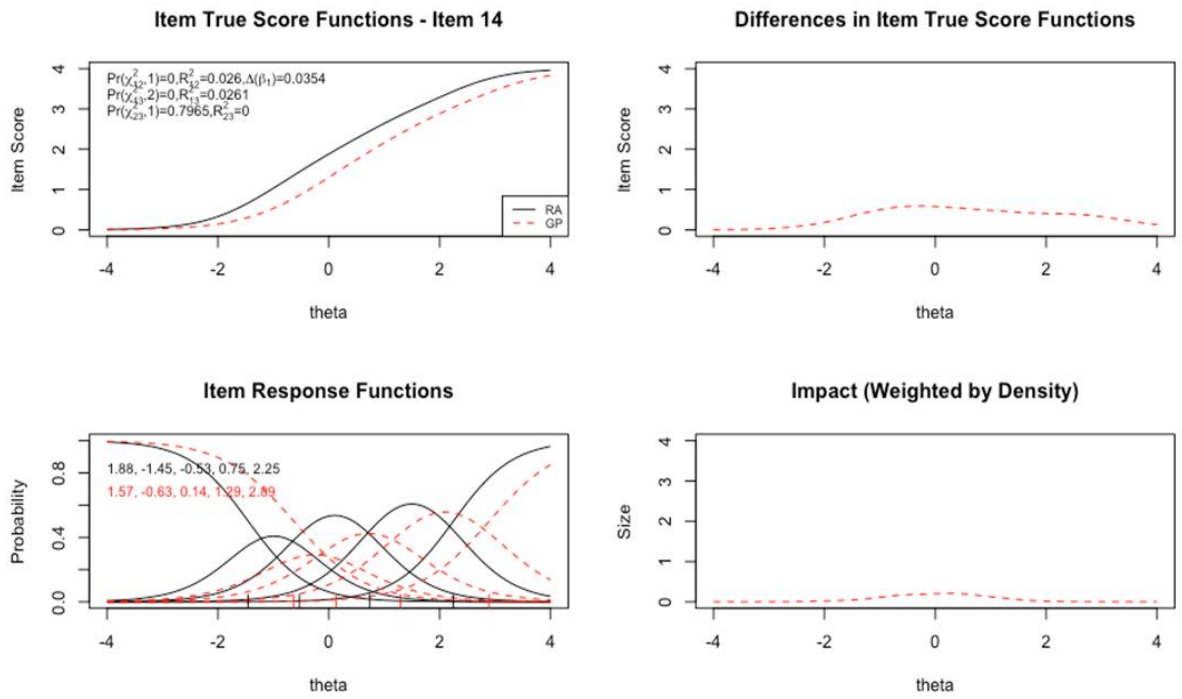

Figure 1. Item true score functions graph from the Pain Behavior item bank, comparison of rheumatoid arthritis patients (RA) and general population (GP) for PAINBE24 (item 14) – *"In the past 7 days... When I was in pain I moved stiffly..."*.

In item 14 (PAINBE24 - *"In the past 7 days… When I was in pain I moved stiffly …")* from the Pain Behavior item bank, comparing the general population with the rheumatoid arthritis group, we see that the general population group has uniformly higher category

thresholds than the rheumatoid arthritis patient group (Figure 1). This shows us that given the same trait level (theta), a person from the general population group is more likely than a person from the rheumatoid arthritis patient group to endorse a higher response category on moving stiffly when in pain.

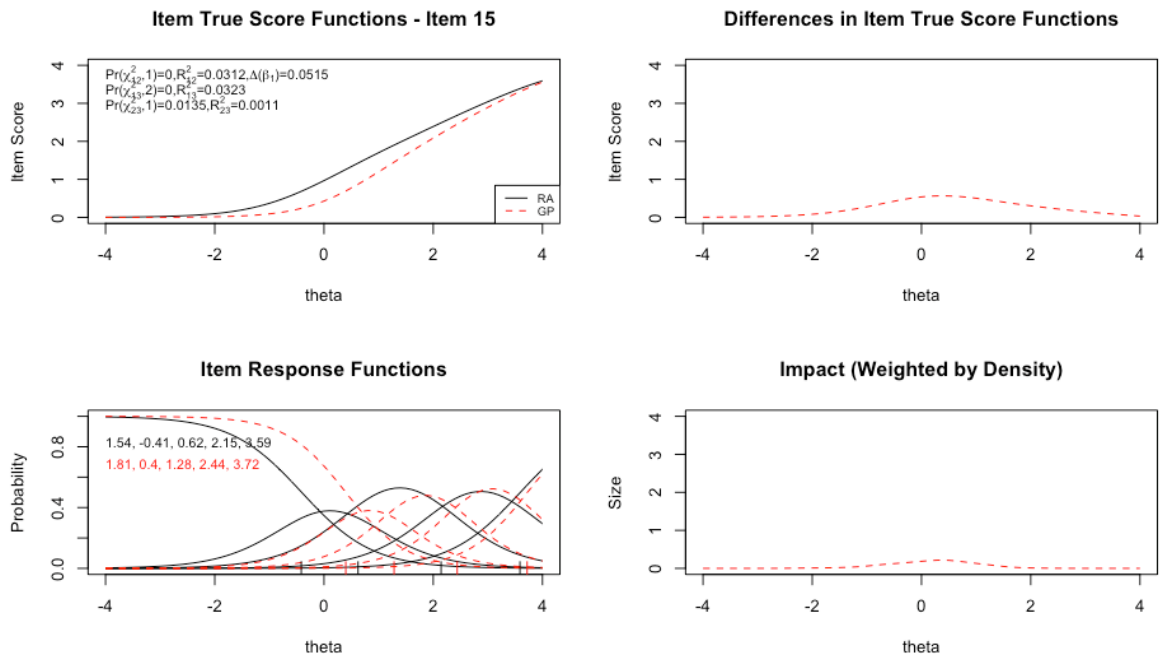

Figure 2. Item true score functions graph from the Pain Behavior item bank, comparison of rheumatoid arthritis patients (RA) and general population (GP) for PAINBE25 (item 15) – *"In the past 7 days... When I was in pain, I called out for someone to help me..."*

In item 15 PAINBE25 (item 15) – *"In the past 7 days... When I was in pain, I called out for someone to help me..."* from the Pain Behavior item bank, in comparing the general population with the rheumatoid arthritis group, the general population group has uniformly higher category thresholds than the rheumatoid arthritis patient group (Figure 2). This shows us that given the same trait level (theta), a person from the general population group is more

likely than a person from the rheumatoid arthritis patient group, to endorse a higher response category on the item for calling out for help when being in pain.

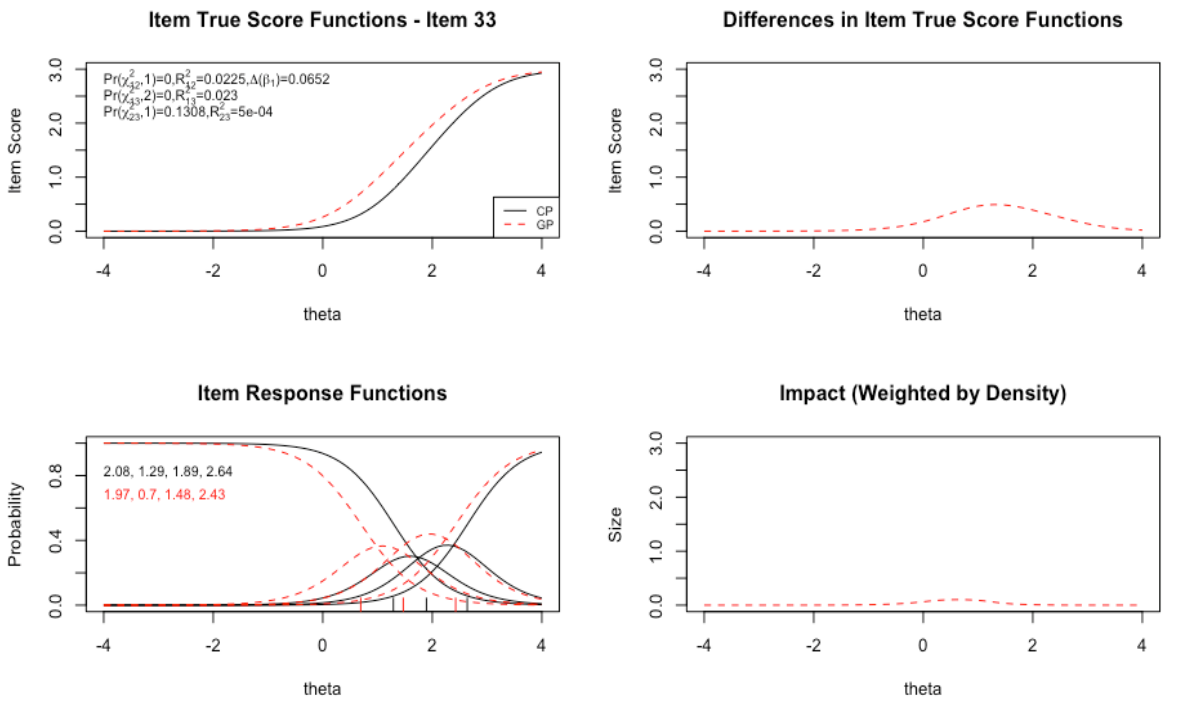

Figure 3. Item true score functions graph from the Pain Behavior item bank, comparison of chronic pain patients (CP) and general population (GP) for PAINBE45 (item 33) – *"In the past 7 days... When I was in pain I thrashed..."*.

In item 33 PAINBE45 (item 33) – *"In the past 7 days... When I was in pain I thrashed...",* from the Pain Behavior item bank, in comparing the chronic pain group with the general population group, we see only three thresholds as no persons in either groups have elected the highest category answer. The chronic pain group has uniformly higher category thresholds than the general population group (Figure 3). This shows us that given the same trait level (theta), a person from the chronic pain group is more likely than a person from the general population group) to endorse a higher response category on item for confirming that when they were in pain, they thrashed.

Pain Interference Item Bank

In the pain interference item bank, the only item flagged for DIF was PAININ20 when comparing the general population with the chronic pain patients.

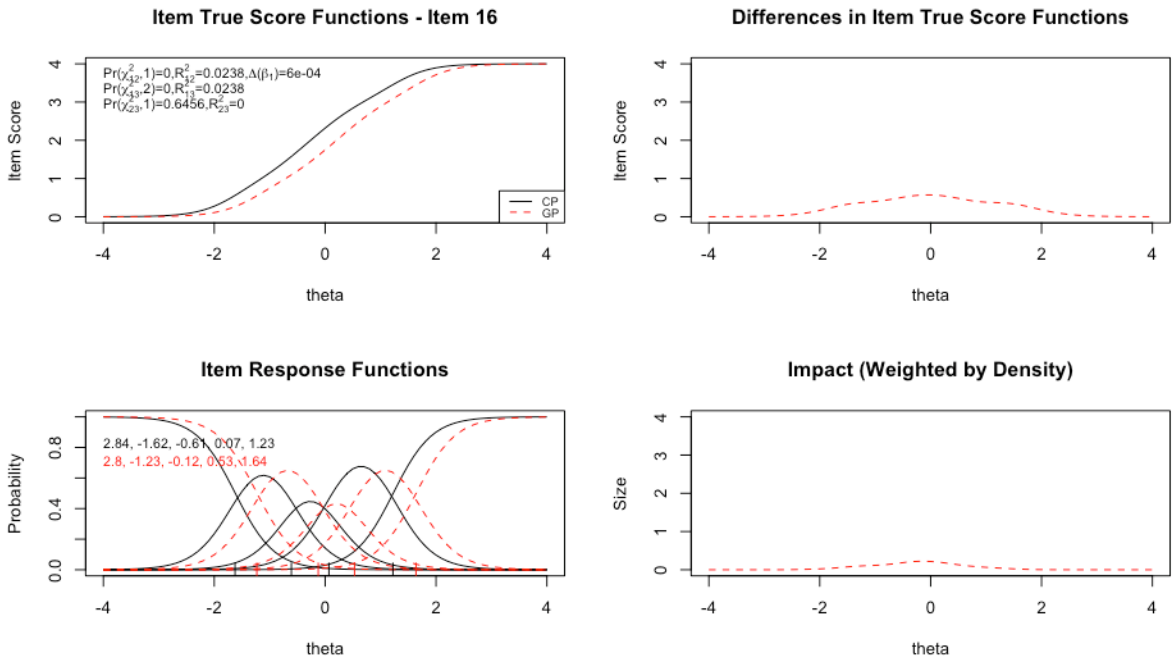

Figure 4. Item true score functions graph from the Pain Interference item bank, comparison of chronic pain patients (CP) and general population (GP) for PAININ20 (item 16) – *"In the past 7 days... How much did pain feel like a burden to you…"*.

In item 16 PAININ20 (item 16) – *"In the past 7 days... How much did pain feel like a burden to you…"*, from the Pain Interference item bank, in comparing the chronic pain group with the general population group, the general population group has uniformly higher category thresholds than the chronic pain group (Figure 4). This shows us that given the same trait level (theta), a person from the general population group is more likely than a

person from the chronic pain group to endorse a higher response category on the item for pain feeling like a burden.

The difference in item true score functions are in general found in the mid to higher levels of theta (upper right section of Figures 1-4). Overall, the item characteristic curves for flagged DIF items show only minimal density-weighted impact between groups (lower right section of Figures 1-4). As datasets are large, this indicates that these differences in measurement parameters may have negligible effect on latent trait estimates as only a few subjects have that theta level.

CFA/MI

By relying on a standard of the ΔCFI .002 criterion, we found no lack of measurement invariance for any of the items in the PROMIS pain item banks. The results are summarized in Table 3. Below, we will discuss the model fitting results for each of the item banks, separately.

Table 3. Measurement Invariance Analysis Results for Pain Behavior and Pain Interference

			Fit Indices					
PROMIS item banks	Groups Compared	Model Fit	χ^2	df	CFI	ΔCFI	RMSEA	SRMR
Pain Behavior	Chronic Pain Patients (CP) vs General Population (GP)	Joined dataset model*	19496.428	702	0.974	-	0.090	0.069
		Configural Fit	20741.583	1404	0.974	-	0.091	0.072
		Metric Fit	22560.431	1442	0.972	0.002	0.094	0.075
		Thresholds Fit	22132.492	1556	0.972	0	0.089	0.072
	Chronic Pain Patients (CP) vs Rheumatoid Arthritis Patients (RA)	Joined dataset model *	24388.905	702	0.971	-	0.092	0.070
		Configural Fit	24529.632	1404	0.972	-	0.091	0.071
		Metric Fit	25973.338	1442	0.970	0.002	0.092	0.074
		Thresholds Fit	26355.133	1557	0.970	0	0.089	0.071
	Rheumatoid Arthritis Patients (RA) vs General Population (GP)	Joined dataset model *	12111.468	702	0.982	-	0.085	0.067
		Configural Fit	12333.386	1404	0.983	-	0.083	0.067
		Metric Fit	12945.652	1442	0.982	0.001	0.084	0.069
		Thresholds Fit	13554.159	1556	0.982	0	0.083	0.067
Pain Interference	Chronic Pain Patients (CP) vs General Population (GP)	Joined dataset model *	70287.428	740	0.990	-	0.160	0.077
		Configural Fit	73441.641	1480	0.989	-	0.163	0.082
		Metric Fit	77146.420	1519	0.988	0.001	0.165	0.084
		Thresholds Fit	74873.646	1638	0.989	0.001	0.156	0.082
	Chronic Pain Patients (CP) vs Rheumatoid Arthritis Patients (RA)	Joined dataset model *	81960.936	740	0.993	-	0.156	0.071
		Configural Fit	80707.665	1480	0.993	-	0.154	0.073
		Metric Fit	87771.289	1519	0.992	0.001	0.158	0.075
		Thresholds Fit	83347.156	1638	0.993	0.001	0.148	0.073
	Rheumatoid Arthritis Patients (RA) vs General Population (GP)	Joined dataset model *	38540.201	740	0.995	-	0.131	0.056
		Configural Fit	40921.358	1480	0.996	-	0.134	0.059
		Metric Fit	43089.544	1519	0.995	0.001	0.136	0.060
		Thresholds Fit	41494.143	1638	0.996	0.001	0.128	0.059

* Joined dataset model for both groups prior to equality constraints between groups

Pain Behavior Item Bank

For all group comparisons, we first assessed the model of the joined dataset for the two groups, with all items loading onto the one latent variable, pain behavior. For the joined dataset containing data from the chronic pain and general population groups, we see that most importantly, CFI is above 0.95 which is very good, SRMR is also acceptable at lower than 0.08, but RMSEA could be better at 0.090. When testing configural invariance between groups, we had to collapse all scores of 6 from the Chronic Pain (CP) group to five, as there was nobody in the General Population group (GP) with a score 6 for two questions. This was the case for items PAINBE40 (collapsed 26 cases), and PAINBE41 (collapsed 20 cases). The configural invariance model fit indices are quite good with regards to CFI and SRMR, except RMSEA again continues to be out of the range of good fit. This will be a constant throughout the comparisons.

In the metric invariance model, a further restriction of loadings being equal between the two groups is added. Now we can focus on the ΔCFI between the metric and configural invariance models. Here the ΔCFI is .002 due to rounding errors. In the thresholds invariance model, we add a further condition of thresholds being equal between the two groups. There is no change in CFI between thresholds and metric fit. Thus we conclude the item loadings and thresholds are invariant between groups.

When testing the configural fit, again, in the Chronic Pain group, 20 scores of 6 had to be collapsed into 5s for question PAINBE41, as nobody in the Rheumatoid Arthritis group had equivalent scores. The CFI remains high at 0.972, with SRMR at 0.070, while RMSEA continues to be higher at 0.091. Looking at the metric fit, we find that ΔCFI is 0.002 due to rounding errors, which means loadings between groups are invariant. Furthermore, there is no change in CFI between the thresholds fit and metric fit, confirming there is also no difference between thresholds for the two groups.

Comparing the Rheumatoid Arthritis patient group with the General Population group, the starting model fit has a high CFI at 0.982, with SRMR at 0.067, and RMSEA at 0.085. CFI remains high at 0.983, with SRMR staying at 0.067, and RMSEA at 0.083. The difference in CFI between the metric and configural fit is only 0.001, confirming that loadings between the two groups are invariant. There is no difference in CFI values between the thresholds and the metric fit, confirming that thresholds are also invariant between groups.

Pain Interference Item Bank

In the comparison between the Chronic Pain and General Population groups, the initial model has an extremely high CFI at 0.990, with SRMR at 0.077, but with a much higher RMSEA at 0.160. The configural fit showed slight worsening of RMSEA at 0.163 and SRMR of 0.082, but CFI remains high at 0.989. The comparison of metric to the configural fit shows a difference in CFI of only 0.001, which indicates that the loadings between the two groups are invariant. Similarly, the difference in CFI between the thresholds fit and the metric fit is also at 0.001, which also indicates that thresholds between the two groups are invariant.

In the second comparison between the Chronic Pain and Rheumatoid Arthritis patient groups, the starting model has a CFI value of 0.993, with SRMR at 0.071, which is good being below 0.080, but RMSEA is again high at 0.156. In the configural fit, where we are designating the distinction between the two groups, the CFI value remains the same, with only slight variation in the remaining indices. In the metric fit, the loadings are found to be invariable as ΔCFI between the metric and the configural fit is only 0.001. The situation is the same when testing for thresholds fit. There is no difference in thresholds between the two groups, as difference in CFI is only 0.001.

Finally, in the comparison between the Rheumatoid Arthritis patient group and the General Population group, the CFI is high at 0.0995, SRMR is lowest so far at 0.056, but

RMSEA is still relatively high at 0.131. The configural fit remained steady with CFI at 0.0996, SRMR at 0.059, and RMSEA at 0.134. ΔCFI between the metric fit and the configural fit, and the thresholds and the metric fit are both 0.001, indicating that both the loadings and thresholds between the two groups are invariant.

Discussion

We assessed differences in measurement parameters in PROMIS item banks of Pain Behavior and Pain Interference. We used IRT and CFA techniques to establish any item measurement invariance and compare the results obtained with the two methods. Using the CFA approach, we have found that all items to be measurement invariant between groups, with regards to loadings or thresholds, while in the IRT approach, we have found three items flagged for DIF in the Pain Behavior item bank, and one item flagged for DIF in the Pain Interference item bank. Sample sizes were large and therefore there was high power to detect differences, although their impact on theta estimates appears minimal.

From the Pain Behavior item bank, we can conclude that the Chronic Pain group is more likely to thrash when in pain (PAINBE45), than is the General Population, given the same level of theta. The General Population is more likely to endorse that when they were in pain, they moved stiffly (PAINBE24), than is the Rheumatoid Arthritis patient group. This is somewhat unusual as stiffness is one of the main characteristics of rheumatoid arthritis. Similarly, the General Population is more likely to call for help when in pain (PAINBE25), than is the Rheumatoid Arthritis patient group. An explanation might be that the general population thinks about different kinds of pain where help is needed, while RA patients think of pain caused by their rheumatoid arthritis for which crying for help does not make sense. A further explanation could also be that as rheumatoid arthritis is a chronic type of disability, a person suffering from it would be more used to dealing with pain in everyday life, and therefore, experiencing pain would not be anything out of the ordinary. On the other hand, a

person with no chronic condition, would not be used to the constant presence of pain, and therefore any painful experience out of the ordinary could be perceived as cause for alarm. Cognitive debriefing regarding the notion of types of pain could be helpful in assessing the root of the difference in this case.

From the Pain Interference item bank, the General Population group is also more likely to endorse having felt pain that felt like a burden (PAININ20) than is the Chronic Pain patient group. A possible explanation could be due to chronic pain patient group being accustomed to living daily with chronic pain, they are less likely to recognize their pain as a burden out of the ordinary, while the general population is more likely to experience acute type of pain which they would recognize as being out of the ordinary.

As PROMIS and its item banks are still relatively novel, not many studies are available that specifically test for DIF between populations, as majority of articles available is focused on calibration of the items. Comparisons for age and gender have been done by Amtmann et al (2010), where they have found in total nine items to have non-uniform DIF in the Pain Interference item bank. Revicki et al (2009) tested for age, gender and education related DIF in the Pain Behavior item bank. They have found one item was detected for gender, and five further items were detected for age related DIF, all of which were uniform. As the comparisons we have made are between several groups, our results are quite promising considering the much lower number of items showing DIF.

The strength of our study was having a variety of large datasets being highly representative of the Dutch and Flemish populations. However, there could have been some selection bias, especially regarding the General Population group. It is possible that the people recruited for the general population group had a higher instance of pain and mobility issues than the actual general population, which could account for some anomalous findings.

Interestingly, contrary to the findings of Kim and Yoon (2011), we have found that in our case, it was the IRT approach that flagged more items for differential item functioning, whereas in the CFA approach, we didn't find any items that showed a lack of measurement invariance. However, this was because we have used ΔCFI of .002 or less as a measure of goodness of fit in order to prevent finding a high degree of false positives to begin with in CFA approach. Had we instead used the traditional chi square test approach, we would have found a lack of measurement invariance at each step for a multitude of items due to the size of the dataset and the large number of degrees of freedom.

We have made detailed comparisons and thorough analyses of DIF in the item banks. Although items flagged for DIF could be advised to be calibrated in order to ensure that ability is measured equally across groups, in the case of the items that were flagged, their impact is minimal on theta. This means that even without any specific calibration, their use in CAT would most likely not have any significant impact on the overall ability results. We can suggest that for the most part, DIF is likely to be negligible and the items may be freely used across varying population. Ultimately, both methods guided us to similar results and the same conclusion.

References

Agresti, A. (2007). *An introduction to categorical data analysis*. John Wiley & Sons, Hoboken, NJ.

Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W. H., Choi, S., Revicki, D., ... & Lai, J. S. (2010). Development of a PROMIS item bank to measure pain interference. *Pain*, *150*(1), 173-182.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. Routledge. New York, NY.

Becker, J., Schwartz, C., Saris-Baglama, R. N., Kosinski, M., & Bjorner, J. B. (2007). Us0ing item response theory (IRT) for developing and evaluating the Pain Impact Questionnaire (PIQ-6™). *Pain Medicine, 8* (3), 129-144.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, *88*(3), 588.

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. *Handbook of structural equation modeling*, 361-379, New York, NY, US: Guilford Press.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., ... & Cook, K. (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.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling*, *14*(3), 464-504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233-255.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, *39*(8), 1-30.

Choi, S. W., Crane, P. K., & Choi, M. S. W. (2016). Package 'lordif'.

Daut, R. L., Cleeland, C. S., & Flanery, R. C. (1983). Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases. *Pain*, *17*(2), 197-210.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement*, *13*(2), 129-143.

Fayers, P. M. (2007). Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. *Quality of Life Research*, *16*(1), 187-194.

Fox, R. J. (1983). *Confirmatory factor analysis*. John Wiley & Sons, Ltd.

Fries, J. F., Bruce, B., & Cella, D. (2005). The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. *Clinical and Experimental Rheumatology*, *23*(5), S53.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory (Measurement methods for the social sciences series, Vol. 2), Sage Publications, Newbury Park, CA.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329-367.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods*, *3*(4), 424.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329-349.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486-507.

Kerns, R. D., Turk, D. C., & Rudy, T. E. (1985). The West Haven-Yale Multidimensional Pain Inventory (WHYMPI). *Pain*, *23*(4), 345-356.

Kerns, R. D., Haythornthwaite, J., Rosenberg, R., Southwick, S., Giller, E. L., & Jacob, M. C. (1991). The Pain Behavior Check List (PBCL): factor structure and psychometric properties. *Journal of Behavioral Medicine*, *14*(2), 155-167.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*(2), 212-228.

Li, C. H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological methods*, *21*(3), 369.

Long, J. S. (1983a). Confirmatory factor analysis: A preface to LISREL. Beverly Hills, CA: Sage

Meade, A. W., & Lautenschlager, G. J. (2004, April). Same Question, Different Answers: CFA and Two IRT Approaches to Measurement Invariance. In *19th Annual Conference of the Society for Industrial and Organizational Psychology* (Vol. 1), Chicago, IL.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568-592.

Mîndrila, D. (2010). Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society*, *1*(1), 60-66.

Pollard, C. A. (1984). Preliminary validity study of the Pain Disability Index. *Perceptual and Motor Skills*, *59*(3), 974.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., ... & Liu, H. (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*, *45*(5), S22-S31.

Revicki, D. A., Chen, W. H., Harnam, N., Cook, K. F., Amtmann, D., Callahan, L. F., ... & Keefe, F. J. (2009). Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain*, *146*(1), 158-169.

Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware Jr, J. E. (2014). The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology*, *67*(5), 516-526.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36.

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Report Series*, *1968*(1).

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

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied psychological measurement*, *7*(2), 201-210.

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

Terwee, C. B., Roorda, L. D., De Vet, H. C. W., Dekker, J., Westhovens, R., Van Leeuwen, J., & Boers, M. (2014). Dutch–Flemish translation of 17 item banks from the patient-reported outcomes measurement information system (PROMIS). *Quality of Life Research*, *23*(6), 1733-1741.

Von Korff, M., Ormel, J., Keefe, F. J., & Dworkin, S. F. (1992). Grading the severity of chronic pain. *Pain*, *50*(2), 133-149.

Walker, C. M., Beretvas, S. N., Ackerman, T. A. (2001). An examination of conditioning variables used in computer adaptive testing for DIF. *Applied Measurement in Education, 14*, 3-16.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364-376.