# An examination of the psychometric properties of the of PROMIS-57 questionnaire Norwegian version

Stein Arne Rimehaug


HELSEF 4502
Masteroppgave i interdisiplinær helseforskning,
Institutt for helse og samfunn, Medisinsk fakultet

UNIVERSITETET I OSLO

15. mai 2020

# An examination of the psychometric properties of the of PROMIS-57 questionnaire Norwegian version



2020

An examination of the psychometric properties of the of PROMIS-57 questionnaire Norwegian version

http://www.duo.uio.no/

Trykk: Reprosentralen, Universitetet i Oslo

# Table of Contents

## Preface

In 2014, a renowned expert in rehabilitation outcome measurement, dr. Allen Heinemann, had a meeting to share his recommendations with me and colleagues from Norway at his research center at Northwestern University in Chicago. He told us about many different questionnaires, some familiar ones, some not. One stood out, as he told us about a research initiative to build a new, better license free system out of bits and pieces of previously used questionnaires for measuring patient outcomes. He explained how this was made possible by using statistical methods from education research. We had never heard of PROMIS, and nor about Item Response Theory, but this started a path of discovery that has been very inspiring for me, personally. International conferences and translation efforts, collaboration across the nations, successes and setbacks. It has also been fascinating to observe that people that can get into long, heated arguments over topics as dry as statistical analysis, and which model is better.

This Master's program has allowed me to dive deeper into the world of health research from many different perspectives, and also cross the line and participate in class in Item Response Theory at the Centre for Educational Measurement (CEMO) at the University of Oslo. All the mornings in 2019 that I exited the commuter train at Blindern station to either cross left, to HELSAM, or go right to CEMO, and experienced two very different learning cultures within the same University, has also been quite fascinating.

I would like to thank my two advisors, Hilde Stendal Robinson and Mari Klokkerud and inspiring educators and co-students at both faculties. Thanks to the students at HELSAM for sharing light-hearted philosophical and critical thoughts. Thanks to the international students at CEMO for practical and mental support through the intense weeks of trying to understand equations that seemingly use the entire Greek alphabet. Thanks to former colleagues Jan Egil Nordvik and Ingvild Grimstad for the exciting early years of getting to know PROMIS. Thanks to Carolyn Terwee and Felix Fischer for practical advice early on in this work, and Aaron James Kaat for his statistical analysis advice towards the end. Thanks to my employer Sunnaas sykehus for giving me the time and opportunity to do this, and eternal gratitude especially to my wife Alison and children Martin and Madelen for all their support and patience.

## Abstract

Purpose:

The aims of the cross-sectional study were to explore reliability and validity of the Norwegian PROMIS-57 questionnaire in a general population sample, n=408, and to examine Item Response properties and factor structure.

Methods:

Reliability measures were obtained from factor analysis and Item Response Theory (IRT)methods, correlations between PROMIS-57 and RAND36 were examined for concurrent and discriminant validity, factor structure and IRT assumptions were examined with factor analysis methods. IRT Item and model fit and graphic plots were inspected, and Differential Item Functioning (DIF) for language, age, gender and education level were examined.

Results:

PROMIS-57 demonstrates excellent reliability and satisfactory concurrent and discriminant validity. Factor structure of seven domains was confirmed. IRT assumptions are met for unidimensionality, local independence, monotonicity and invariance with no Differential Item Functioning (DIF) of consequence for language or age groups. Estimated Common Variance (ECV) per domain and CFA model fit supports unidimensionality for all seven domains. Acceptable Graded Response model fit and IRT plots.

Conclusions:

The psychometric properties and factor structure of Norwegian PROMIS-57 are satisfactory, and this questionnaire along with PROMIS 29 and the included 8 or 4 item short forms for physical function, anxiety, depression, fatigue, sleep disturbance, social participation ability and pain interference are ready for use in research and clinical care in Norwegian populations. Further studies on longitudinal reliability and sensitivity in patient populations and for Norwegian item calibration and reference scores are needed.

## Abbreviations

IRT – Item Response Theory

CFA – Confirmatory Factor Analysis

EFA – Exploratory Factor Analysis

PHO – PROMIS Health Organization International

DIF – Differential Item Functioning

GRM – Graded Response Model

GRSM – Generalized Rating Scale Model

HRQOL - Health Related Quality of Life

PROM – Patient Reported Outcome Measures

PROMIS – Patient Reported Outcome Measurement Information System

RMSEA – Root Mean Square Error of Approximation

SRMSR - Standardized Root Mean Square Residual

TLI - Tucker Lewis Index

CFI - Comparative Fit Index

BIC – Bayesian Information Criteria

PF – PROMIS Physical Function short form or domain

ANX – PROMIS Anxiety short form or domain

DEP – PROMIS Depression short form or domain

FAT – PROMIS Fatigue short form or domain

SLP – PROMIS Sleep Disturbance short form or domain

SOC – PROMIS Social Roles and Activities Abiity short form or domain

PAIN – PROMIS Pain Interference short form or domain

# Introduction

## The aims and structure of this thesis

### Aims of the study

This is a cross-sectional study with the purpose to explore the psychometric properties of the Norwegian version of the PROMIS-57 questionnaire in a convenience sample of respondents from the general population in Norway. It is written as a "kappa" thesis accompanying an article designed for publication by the guidelines for the peer reviewed journal Quality of Life Research.

### Research questions

**What are the psychometric properties and validity of the Norwegian version of PROMIS profile 57, and its concurrent validity against RAND-36?**

Detailed research questions for the article:

- What is the internal consistency and reliability for each of the seven PROMIS domains and PROMIS-57 and PROMIS 29, respectively, and as a whole?
- What is the concurrent validity of the seven PROMIS-57 domains against scales/items measuring the same in RAND-36?
- What is the discriminant validity of each of the seven PROMIS-57 domains measured against other domains from both PROMIS-57 and RAND 36
- Does confirmatory factor analysis confirm the factor structure of 7 domains?
- Examining IRT Item Characteristic Curves, and Test Information Function plots, do they exhibit acceptable parameters for each PROMIS-57 domain?
- Examining Standard Error plots for each domain, is reliability diminished in PROMIS 29 compared with PROMIS-57?
- Is the data free of group invariance, as expressed by Differential Item Functioning for language DIF, (as well as age, gender and education), using DIF analysis to the extent sample size allows

Additional aims for the kappa thesis:

- To present background information to help put the study into a research context
- To discuss the rationale for the choice of statistical methods
- To explore the results of analysis to explain variation from the main conclusion

### Structure of the thesis

This thesis starts with explaining central terms, some background on PROMIS, and how factor analysis and Item Response Theory has been instrumental in the development of PROMIS. Included is some basic description of Item Response Theory, followed by a description of the data collection and variables in the data set, and a section discussing the selection of statistical methods. The results of the analysis are mostly covered in the article, but before presenting the article itself, I have included a more in-depth presentation and discussion of some of the results that somewhat deviate from the main conclusion. References for the thesis are presented separately, while the article has its own references. Finally, appendices and supplementary material.

## Background

### PROMIS-57 and RAND-36

Two questionnaires were used in this study; PROMIS-57 (Patient Reported Outcome Measurement Information System 57-item Profile), and RAND 36-item Health Survey 1.0 (RAND-36) (R. D. Hays, Sherbourne, & Mazel, 1993). RAND-36 is a license free questionnaire almost identical to Short Form-36, arguably the most commonly used questionnaire for measuring quality of life in health research (R. D. a. R. Hays, B. B . , 2008). PROMIS-57 is a more recent alternative. It was created with newer test development methods that sets it apart from most other questionnaires that are commonly used in a quality of life measurement. It is not actually a single questionnaire, but rather a collection of seven short form questionnaires, and each of the seven should be scored separately. PROMIS-57 is only small portion of the PROMIS system. PROMIS consists of not just PROMIS-57, but a wide range of PROMIS questionnaires and item banks, described on the PROMIS web site as “a set of person-centered measures that

evaluates and monitors physical, mental, and social health in adults and children. It can be used with the general population and with individuals living with chronic conditions." (www.nihpromis.org, 2020). That description does not say directly that PROMIS measures health related quality of life (HRQoL), yet it does belong to that category of questionnaires.

### A Health-Related Quality of life questionnaire, or not?

Most published studies involving PROMIS include something about quality of life in the introductory description, and quite a few are published in "Quality of life research" journal. Marcel Dijkers commentary "What's in a name?" The indiscriminate use of the "Quality of life" label" (Dijkers, 2007) made a point that the term quality of life had become too commonly applied without attempting to define it more closely, and that it can take on different meanings for different people. According to Dijkers, popular "quality of life" questionnaires like Short Form-36 are applied in research without much thought to whether it is able to capture what makes life good for the respondents. Volumes more has been written on the subject, and this is one possible reason developers of a new questionnaire like PROMIS-57 would simply avoid labelling it as an instrument for measuring health related quality of life, and leave it to the researcher to operationalize which aspect of life to measure. Thus, a broad discussion about "what is quality of life" and how it is different from "health related quality of life" may fall outside the scope of this study of the psychometric properties of a questionnaire that is not labelled as a quality of life measurement instrument, anyway. However, it is important to explore whether it measures what it is supposed to measure, as that is the core meaning of validity.

### Central concepts, terms and definitions

PROMIS questionnaires measure qualities that are not immediately observable and measurable; conditions that an individual may experience to different degrees at different times, such as common emotional states, pain, fatigue or anxiety (Schnohr, Rasmussen, Langberg, & Bjorner, 2017). These so-called latent constructs emerge from the shared human experience, and their understanding can be influenced by current culture and discourse, both in the mind of the researcher and the respondent. '*Latent*' in this context means anything that is not directly and

immediately measurable (Khanna et al., 2011). The science around validation of questionnaires tries to make sure that the constructs we hope to measure are grounded in reality, are coherent and unidimensional, and are measuring the same latent constructs in different people. In health and psychology research, different terms for such latent life constructs are used interchangeably; such as latent variables, domains, scales, and constructs, and are not always clearly defined. In this thesis, a *construct* means the general idea of a particular concept to be measured, a *latent variable* is the quantitative measurement of such a construct, and *domains* are the specifically defined latent variables, also called scales, that are included in the PROMIS-57 questionnaire. The range of possible values from low to high within a domain is referred to as a *scale*. In IRT and in HRQOL research, an *item* is a single question within a questionnaire, and the answer options for any item are called *response categories*. Consequently, '*test level analysis'* assesses the entire questionnaire all at once, '*scale level analysis'* means one domain at a time, while '*item level analysis' means* examining the properties of each item separately.

## What is PROMIS?

There is an important distinction between the acronyms PROMs and PROMIS: 'Patient Reported Outcome Measure' (PROM) is the common term for any questionnaire used in health care and health research that enables patients to self-report health, symptoms and emotional states, and is described in more detail in an introductory article by Weldring and Smith (Weldring & Smith, 2013). 'PROMs' is just the plural form of PROM, not to be confused with 'PROMIS'. With the help of PROMs, the measured improvement in a person's self-reported condition is used in health research and clinical care to assess the outcome of health interventions, such as medications, surgery, or rehabilitation. The use of PROMs allows clinicians and researchers to include different aspects of self-reported health related quality of life (HRQOL), both in the assessment of a patient's current status, and change in HRQOL over time. Disease specific PROMs are "designed to identify specific symptoms and their impact on the function of those specific conditions" (Weldring & Smith, 2013), while generic PROMs ask questions that are well suited to use over a wide range of diagnoses, including some well-known

QoL measures, such as SF-36 (the Medical Outcome Study 36-item Short-Form Health Survey) and the Euro-Qol EQ-5D.

## The PROMIS initiative

After decades of PROMs being used in research, the NIH (National Institutes of Health) in the US launched a research initiative to develop a new flexible questionnaire system, that eventually became known as the "PROMIS" initiative (Cella et al., 2007). PROMIS is an acronym for Patient Reported Outcome Measurement Information System. Looking closer at that term helps explain what PROMIS is. The first part, of course, indicates that this is a type of PROM. The last part of the PROMIS name, "measurement information system", alludes to the fact that it is a system for combining PROM items (or questions), in new and more flexible ways, rather than using long, fixed questionnaires.

## The background for PROMIS

The motivation behind developing PROMIS came out of dissatisfaction with existing PROM questionnaires, and the emergence of statistical analysis methods that made it possible to do something about it (Cella et al., 2010). It is fairly common in health research to use both generic and disease-specific PROMs together (Weldring & Smith, 2013). This often results in a high respondent burden, especially if many questionnaires are long and used together. Many of the questions may appear repetitive or irrelevant to the person answering them. The intention behind the development of PROMIS was to reduce response burden and improve measurement precision with the help of new statistical methods (Cella et al., 2007). Educational research had long ago adopted new psychometric and measurement methods, mainly factor analysis, item banking and Item Response Theory (IRT) (Reckase, 1979), that the PROMIS developers of PROMIS wanted to take advantage of. I will now go on to describe those newer methods and concepts, how they inspired the PROMIS initiative, and ultimately why these methods are used in this study on the psychometric properties of the Norwegian PROMIS-57.

## Item Response Theory and Item banking in educational research

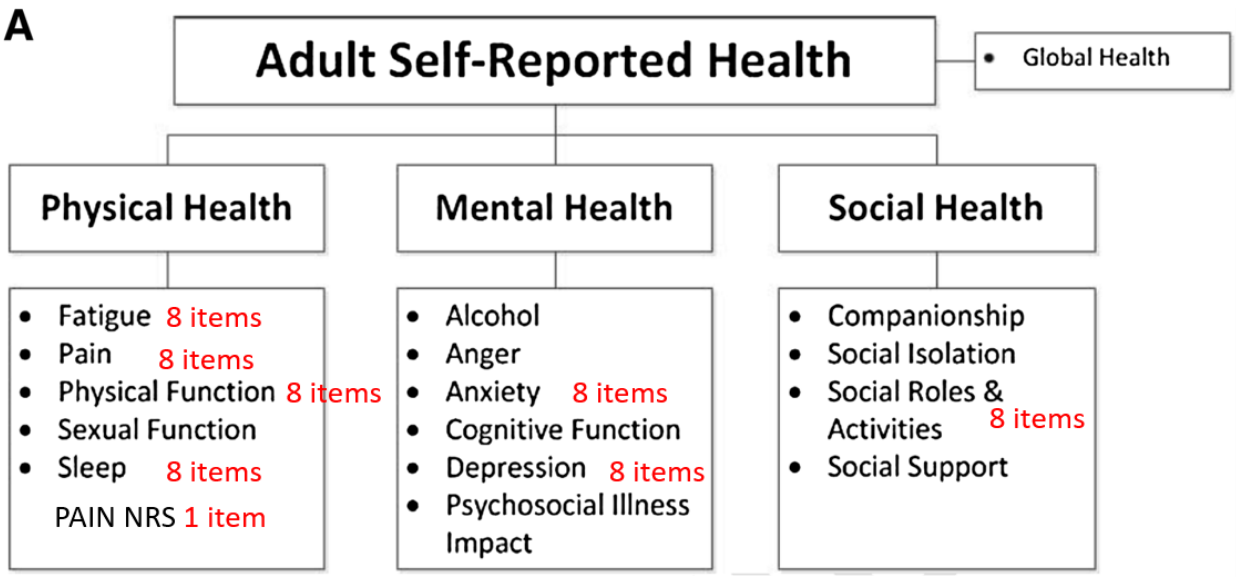
Already in 1997, the head researcher behind PROMIS, David Cella, presented the idea of using item banking and computer programs to perform shorter, yet more accurate self-reported measurement for patients with HIV (Revicki & Cella, 1997). The article refers to the successful applications of Computer Assisted Testing in personality testing, and item banking from educational research, and outlines how it would be feasible to do something similar in health research. Now, 23 years later Cella has seen his vision through, and multiple PROMIS item banks, built on this idea, totalling over 1000 items are in use in health research for a wide variety of patients (Smith & Jensen, 2019). Item banking is a concept that originated in educational research, and is made possible by Item Response Theory (IRT) statistical methods (Riley et al., 2010). The need to have a larger bank of items to draw from emerged in large scale college entry exams in the US. If students get totally different questions on a test, they can't reveal to the next round of students what questions to expect. IRT methods can establish the difficulty level of every single question on the same scale, so that students receiving different questions still can be fairly scored. There has to be separate difficulty scales for different school subjects, and thus separate Item banks of hundreds of items (questions) for each subject. This method is behind well-known large-scale test systems like the SAT (Scholastic Aptitude Test) and PISA (Programme for International Student Assessment) (Braun & von Davier, 2017). The PROMIS initiative came about when applying the same logic to HRQoL research and questionnaire development, and creating item banks for different conceptual component of HRQoL (Cella et al., 2007). Those conceptual components are referred to as *domains*.

## The item banks of PROMIS-57

The PROMIS initiative developed item banks and new test methods, by collecting and re-using items from hundreds of other existing questionnaires, but keeping only the items considered to be best performing, through a long development process, using qualitative methods, Exploratory Factor Analysis (EFA) and IRT (Cella et al., 2010). PROMIS now consists of several clearly defined item banks (Riley et al., 2010), each containing many dozen items. The purpose is not to expose an individual to all them, but rather select just a by using 'Computer Assisted Testing'

software or a fixed short form questionnaire selected from the item bank. Figure 1 shows the domain structure of the PROMIS system of Item banks.

Figure 1: PROMIS domain and Item bank structure




Red text indicates which PROMIS item banks the different PROMIS-57 domains/items belong to.

## Defining the PROMIS-57 domains

The different PROMIS domains, and the conceptual understanding of the seven domains included in PROMIS-57 are defined in an article from 2010 (Cella et al., 2010). I am including a verbatim quote to avoid altering the definitions. Descriptions of other domains not included in PROMIS-57 are omitted in this long quote:

> "**Physical function** is defined as one's ability to carry out various activities that require physical capability, ranging from self-care (activities of daily living) to more vigorous activities that require increasing degrees of mobility, strength, or endurance. Physical

function is conceptually multidimensional, with four related subdomains: mobility (lower extremity function), dexterity (upper extremity function), axial (neck and back) function, and ability to carry out instrumental activities of daily living.

**Fatigue:** In the health outcomes measurement perspective, fatigue is defined as an overwhelming, debilitating, and sustained sense of exhaustion that decreases one's ability to carry out daily activities, including the ability to work effectively and to function at one's usual level in family or social roles… …Fatigue is divided conceptually into the experience of fatigue (such as its intensity, frequency, and duration), and the impact of fatigue upon physical, mental, and social activities.

**Pain:** Pain is an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage. Pain is what the respondent says it is—that is, the "gold standard" of pain assessment is self-report. Pain is divided conceptually into components of quality (referring to the nature, characteristics, intensity, frequency, and duration of pain), impact upon physical, mental and social activities, and behaviors one engages in to avoid, minimize, or reduce pain.

…

The PROMIS **Sleep Disturbance** item bank focuses on perceptions of sleep quality, sleep depth, and restoration associated with sleep; perceived difficulties with getting to sleep or staying asleep; and perceptions of the adequacy of and satisfaction with sleep. The Sleep Disturbance item bank does not include symptoms of specific sleep disorders, nor does it provide subjective estimates of sleep quantities (e.g., the total amount of sleep, time to fall asleep, or amount of wakefulness during sleep).

…

**Depression:** The PROMIS item bank for depression focuses on negative mood (e.g., sadness, guilt), decrease in positive affect (e.g., loss of interest), information-processing deficits (e.g., problems in decision-making), negative views of the self (e.g., self-criticism, worthlessness), and negative social cognition (e.g., loneliness, interpersonal alienation).

**Anxiety:** The PROMIS item bank for anxiety focuses on fear (e.g., fearfulness, feelings of panic), anxious misery (e.g., worry, dread), hyperarousal (e.g., tension, nervousness, restlessness), and somatic symptoms related to arousal (e.g., cardiovascular symptoms, dizziness).

**Social function:** Social function is defined by PROMIS as involvement in, and satisfaction with, one's usual social roles in life's situations and activities. These roles may exist in dyadic or family relationships, parental responsibilities, work responsibilities and social activities. Social function has also been referred to with terms such as role participation and social adjustment. …a conceptual division of social function into "ability to participate" and "satisfaction with participation." Each of these two components has sub-components that divide social roles such as work and family responsibilities, and more discretionary social activities such as leisure activity and relationships with friends." (Cella et al., 2010)

## Item Response Theory

### Item Response Theory in health research

Item response theory (IRT) is not really a single theory, as much as it is a type of statistical analysis models that have been used in educational research and test development for decades, but have been more slowly adopted in health sciences. Many health studies, especially European ones, have utilized analysis with Rasch models to test questionnaires. Rasch can also be considered a restricted form of IRT, while more complex forms of Item Response modelling, and Item banking were long absent from health research (R. D. a. R. Hays, B. B . , 2008). One 2009 article describes how IRT at that time was just starting to be applied in clinical health studies (Steven P. Reise & Waller, 2009). Questionnaires developed with IRT, whether Rasch or other models, are claimed to be more sensitive and accurate (Stover, McLeod, Langer, Chen, & Reeve, 2019).

## Item Response Theory basics

In Item Response theory, the imagined or measured position somewhere on a scale of any latent variable, the difficulty or severity of what is being measured, is denoted with Greek letter θ and referred to as the *theta*. In IRT models, the structural elements are called *parameters*, and the theta is the first parameter. Rasch analysis is a type of IRT model called *1PL model*, meaning that it uses only this first parameter. 1PL IRT analysis estimates the theta for each item, and also for each response category within an item, placing them on an interval scale. The second parameter is discrimination, which refers to how well an item is able to separate individuals located above or below the item's theta level. In IRT graphic plots, discrimination is represented by how steeply the curves rises on the y-axis, so discrimination may also be referred to as the *slope*. The steeper the slope, the better the discrimination ability of the item. IRT analysis methods that considers the discrimination in addition to the theta, are called *2PL models*. In this study, a 2PL model called Graded Response Model (Stover et al., 2019) is mainly used, sometimes referred to as Samejima's model, GRM or just 'graded'. In an IRT context, non-IRT methods like internal consistency, correlation studies, and factor analysis are referred to as Classic Test Theory (CTT). This study employs combination of some of the CTT and IRT statistical analysis methods that were also used to develop PROMIS.

## Computer assisted testing, or not

Educational research, and eventually IRT based health research, proved that an effective way to decrease the response burden would be to to use a CAT computer program (Computer Adapted Testing) (Senders, Hanes, Bourdette, Whitham, & Shinto, 2014) to iteratively select items for each individual, by letting the response to an item guide the selection of the next. Since the calibrated weight of each item has been estimated, the software can select just three or four items from an item bank and place the individual more precisely than a longer questionnaire (Segawa, Schalet, & Cella, 2019). This is really the essence of what makes PROMIS different and attractive, but PROMIS CAT is out of reach, for now, in Norway. The technological infrastructure needed to use such CAT modules is difficult to implement, and translating all of the PROMIS item banks would be an enormous undertaking. As an alternative, PROMIS

researchers developed static Short Forms of different lengths, consisting of 4, 6, or 8 items selected from an item bank (Cella et al., 2019). Each short form covers a single domain, such as pain, sleep, anxiety or depression. In order to meet clinician's and researcher's expectation of a more traditional generic fixed HRQOL questionnaire, PROMIS profile forms (PROMIS-57 and 29) were established, by combining selected short forms for the most commonly investigated domains. Many more PROMIS item banks exist in the PROMIS system, totaling more than one thousand items, so these profile questionnaires are not utilizing the full potential of item banking. They are using only seven of the item banks and a more limited, fixed set of items. Yet, the more generic PROMIS-57 and PROMIS 29 are being translated and used in research worldwide. These profile forms, PROMIS-57 and PROMIS 29, are the subject of this study. PROMIS-57 has been translated into Norwegian previously, but the translated version has not been the subject of a validation study, until this one.

### Scoring method: T-scores

A person's raw score for each of the seven PROMIS-57 domain scales can range between 8 and 40. The IRT analysis transforms the raw score to a calibrated Theta score, and by convention placing 0 as the estimated mean, and Theta=1 as an expression of 1 Standard Deviation (SD) above 0. Since the position on the scale is established by the IRT analysis, items in a questionnaire can be selected to make it cover the theta range of the target population better. Custom made PROMIS questionnaires can be assembled, using items that match the expected theta levels of the respondents, and leaving out items that are irrelevant or out of range, exemplified by the PROMIS-FatigueMS form for Multiple Sclerosis (K. F. Cook et al., 2012). PROMIS-57, on the other hand was created to be generic, and possible to use across many conditions and diagnoses. The process of selecting the seven domains for PROMIS-57, and also which items to include in each short form is well described by Cella et al. (Cella et al., 2019). Since PROMIS-57 was meant to be used across a wide range of conditions and diagnoses, the chosen items would need to cover a range of thetas that would be appropriate for many patient populations. This was accomplished by selecting items not around the average score for healthy people, but centered around 1 SD – one standard deviation in the less desirable direction (Cella

et al., 2019). Another feature of the PROMIS scoring is that the scores are converted to a T-score metric, through calibrated scoring, in order to make interpretation of scores easier. A T-score of 50 on any questionnaire is the reference population mean of the scale (reference theta=0), and every 10-point theta deviation from 50 equals one Standard Deviation (1SD) removed from that average. T-scores can be obtained either by using a look-up table (see appendix 2) from www.healthmeasures.org or uploading anonymous data to a scoring server at the same site.

### Previous similar research

There is a large body of published research on the psychometric properties of various PROMIS instruments. While a systematic article presentation is outside of the scope of this article, searches in PubMed and Google Scholar, resulted in about 200 highly relevant articles, though not all are used as final references. Search strategies used included key concepts in IRT, technical terms related to the statistical methods, and terms related to PROMIS.

### Studies describing and validating the development of PROMIS scales

There are studies and overview articles from the development of PROMIS that have set the standard for analytic methods, especially the overview articles for the entire PROMIS process (Cella et al., 2007) and (Cella et al., 2010), and for the PROMIS-57 (Cella et al., 2019). There is certainly a risk of bias in these articles from the developer, so I have made a point of checking the theoretical rationale and methods also from other sources.

### Validation studies in other languages and patient populations

Many validation studies in different patient populations have been published, and also articles presenting validation of PROMIS item banks and questionnaires after translation. Most of these use similar types of CTT and IRT methods, while others have used Rasch analysis. Many validation studies have been done on PROMIS item banks, while only a few assess PROMIS CAT modules, profile and short forms.  There is at least one study validating PROMIS-57 as

such (Tang et al., 2019), studying internal consistency reliability, correlations for convergent validity, known-groups comparison, and applying factor analysis. A few more studies of psychometric properties have been done on PROMIS-29 (Coste, Rouquette, Valderas, Rose, & Leplege, 2018), (F. Fischer et al., 2018), (Hinchcliff et al., 2011), (Katz, Pedro, & Michaud, 2017) and (Rose et al., 2018). Several studies are validating some of the PROMIS short forms that are included in PROMIS-57, in patient populations such as hepatitis (Evon et al., 2018), cancer (Hahn et al., 2016), (Jensen et al., 2016), and (Teresi, Ocepek-Welikson, Kleinman, Ramirez, & Kim, 2016), joint pain (Hackney, Klinedinst, & Resnick, 2019), arm disability (Hung et al., 2018), obesity (Kudel et al., 2019), fibromyalgia (Merriwether et al., 2017), ankle surgery, (Stephan, Mainzer, Kummel, & Impellizzeri, 2019), CFS/ME (Yang, Keller, & Lin, 2019), and diverse populations (Ameringer et al., 2016). Finally, I have studied several articles validating entire item banks, mostly to highlight methodological issues, as the results from these are not directly comparable to results from short forms. Many of the studies have provided valuable comparative information, and insight into the methods used, and are referred to in this thesis, when relevant.

# Methods

## Data collection and ethics considerations

The data for this study was collected by colleagues at Sunnaas sykehus. The local data protection officer (*Informasjonssikkerhetsleder og personvernombud*) was informed of the project plan for data collection as an anonymous online survey using the www.nettskjema.no portal, in order to have secure and anonymous collection of data. This collection method did not require further ethics approval or database approval, as there was no intervention, and was considered to involve no personal information protection issues. There was no identifying personal information in the data set and no Internet Provider (IP) address, email, geographic location or any other identifying information collected or stored in the system.

Efforts were made between December 2018 and March 2019 to recruit a convenience sample intended to approximate a cross-section of the general population. Recruitment was done through a one-time advertisement in Aftenposten/A-magasinet Sunday edition (circulation 770 000) in January 2019, multiple posts on Facebook pages for Sunnaas sykehus, followed by over 6000 people, and a promoted post on Facebook page of Regional kompetansetjeneste for rehabilitering which reached 1086 people. These Facebook posts were in addition shared by some of the people reached and their "friends of friends. Respondents were instructed to reply only once. Respondents were informed at the introduction page that clicking "next" to start completing this questionnaire would in effect imply consenting to the anonymous responses being used to test the quality of the questionnaire, as well as used in published research.who were directed to an anonymous online questionnaire, including information about the purpose of the study, and a consent statement. (see "Samtykke og instruksjoner appendix 1).

## Sample descriptive variables

Respondents demographic information was collected: gender, year of birth, education level categories, work status categories, income categories, cohabitation, self-reported presence of mental and/or physical conditions, and whether taking medications prescribed by a doctor. See Table 1 and a further descriptions in the in the results section in the thesis and in the article.

## The response variables from PROMIS-57 and RAND 36

All respondents completed Norwegian versions of PROMIS Profile 57 and RAND 36-item Health Survey 1.0 (RAND-36) (R. D. Hays et al., 1993). In PROMIS-57, there are eight items for each of these seven areas. These 56 items all have 5-category Likert scale responses, meaning symmetrical negative to positive response or lower to higher options, such as "Not at all, A little bit, Somewhat, Quite a bit, Very much." Every PROMIS item is scored 1-5, in such a way that a '5' is always 'more' of the measured variable, regardless whether it is desirable, like social participation, or undesirable, like pain. The 57$^{th}$ item is a pain intensity numeric rating scale (NRS) from 0-10. A complete list of all 57 items in Norwegian and English with response options can be found in the appendix, part 3.

## PROMIS-57 item scores and domain scores

The individual responses to every PROMIS-57 item were used in this study for descriptive statistics, factor analysis and IRT analysis. They also form the basis for sumscores (=raw scores) for each domain scale ranging from 8 to 40 points. Each sum-score was converted to T-scores, by uploading the just the scores without personal information to an online scoring service that returns raw score, Theta level, SE, and T-score for every person. These scores were obtained for comparison, but the analysis was done mostly on raw scores. Thus, the available PROMIS variables for study are each individual respondents' item scores,seven domain sumscores, seven T-scores, plus the pain NRS 0-10 scores. In addition, the SE estimates per domain for each individual from the PROMIS scoring service.

## RAND 36 description and variables

In addition to PROMIS-57, the RAND 36-item Health Survey 1.0 (RAND-36) was also collected from each respondent. RAND-36 contains the same items as the original SF-36 (R. D. Hays et al., 1993), but has a different scoring. The 3, 5 or 6 category responses converted to sum-scores, using the official RAND-36 scoring syntax (R. D. Hays & Morales, 2001), so that higher scores

indicate more desirable health on a 0-100 scale. These 36 items and ten sum-scores for each respondent are available for analysis.

### Additional US data sets – WAVE 1 and HUI profiles

An extension of IRT is Differential Item Functioning analysis (DIF). This checks whether different populations are measured fairly, and can shed light on whether individual items in a questionnaire contribute to the score in the same way for men/women, older/younger people, across education levels, etc. The DIF method can also be used to assess language DIF, that is if each item contributes to the score in the Norwegian questionnaire the same way they do in the English original. Language DIF can be tested when sufficiently large samples complete the same set of items in each language. A US dataset was needed in order to do the language DIF analysis for PROMIS-57. For this purpose, two US data sets were obtained and prepared, in addition to the main data set from Norway. PROMIS experts recommended two data sets that are available for download on a Harvard University Dataverse server: The *Wave 1* (Cella, 2015) and *PROMIS profiles HUI (Cella, 2017)*. These contain many thousands of respondents and hundreds of items, but only the respondents that had been asked all the items in one PROMIS short form were selected, in order to avoid biasing the results with differing sample sizes and respondent characteristics between item. There are between 800 and 3000 qualifying respondents in these remaining sub-samples.

## Methods for statistical analysis

### Standards and recommendations

There are three sources for direct recommendation for what methods to use for validating translated PROMIS measures. PROMIS Health Organization (PHO), the organization overseeing development, translation and adaptation of PROMIS measures have issued two different documents, the PROMIS Standards for release (PHO, 2014) and the PROMIS® Instrument Development and Validation Scientific Standards Version 2.0 (revised May 2013), especially appendix 8 – 11 (PHO International, 2013). These documents both set recommendations for

validation studies, the second one in greater detail. Another guiding source for the choice of methods Item Category Curves or this study has been the COSMIN risk of bias checklist (Mokkink et al., 2018), a much-used international reference for what should be reported in studies involving PROM questionnaires. These recommendations have guided the choices of methods, though there are a few options to consider within these recommendation.

### Scoring

The recommended scoring method for PROMIS is conversion to T-scores. As already mentioned, this gives a scaled, calibrated scoring that makes interpretation of scores easier, and the same for all scales, once familiar with the system. Result from shorter or longer questionnaires, and from CAT modules are directly comparable when using T-scores. The international reference to be used then is from a US population sample, and studies have found the cross-cultural bias by doing so to be small (F. Fischer et al., 2018) and (H. F. Fischer et al., 2017). However, because of some risk of cultural bias, the T-scores were obtained, but raw scores were used for most of the analyses, unless stated otherwise.

## Methods for reliability testing

### Internal consistency and the limited value of Cronbach's alpha

Internal consistency is the concept of how well the items fit together. If some items within a scale measure something different than intended, this would reduce the reliability of the questionnaire. Cronbach's alpha is commonly used to test the reliability of questionnaires, and reported in many studies as a measure of internal consistency, but according to some sources also misused and over-interpreted. Sijtsma (Sijtsma, 2009) provides an overview of the shortcomings and misunderstanding surrounding reporting av Cronbach's alpha, even going as far as saying it is NOT a measure of internal consistency. Cronbach himself took the opportunity at the 50$^{th}$ anniversary of the original publication to point out that Cronbach's alpha now being overused and overinterpreted as a reliability measure (Cronbach & Shavelson, 2004).

In some of the published psychometric studies of PROMIS questionnaires, other measures of reliability are reported instead of alpha. The rationale for these alternative methods for quantifying the internal consistency and other aspects of reliability should be considered.

### Reliability measures from factor analysis – omega

MacDonalds Omega is starting to become more commonly reported, and the "From alpha to omega" article by Dunn et al (Dunn, Baguley, & Brunsden, 2014), claims that omega is as a better alternative to alpha, partly because it is not biased towards higher numbers of items. Dunn argues that if reporting Cronbach's alpha without including omega, one must also include the CI's (Confidence Intervals), and check that the underlying model is "Tau-Equivalent", meaning that every item has about the same factor loading in factor analysis. Alpha is very commonly reported without that information, but since the evidence against that practice is convincing, the omega indices are also explored and reported in this study. Omega is reported sometimes as omega_h (hierarchical), other times as omega_t (total), and denoted with the Greek letter as ω_h and ω_t. These are both derived based on the squared factor loadings from factor analysis. Omega is based on a particular kind called bi-factor analysis, which tests the response patterns in relation to both a main factor and other interfering factors. Omega_h, then, is based on the relation to the main factor (the squared factor loadings), whereas omega_t takes into account all the factors. In her review of evidence and reporting practices for alpha in the literature, Taber also argues that "internal consistency" is a poorly defined concept, and that Cronbach's alpha originally was meant to be a cross-sectional reliability measure of *equivalence* – or "whether different sets of test items would give the same measurement outcomes", and "how much the test score depends upon general and group, rather than item specific, factors". This is actually very close to what is measured with the omega_h reliability measure, but without the undesirable effect of favorable bias for longer questionnaires.

Omega calculation is not included in SPSS, but the open source statistical software system 'R' (R Core Team, 2018) puts the different omega measures within reach by using the 'omega' function in the R package 'psych' or the 'reliability' function in 'semTools' R package. Since R was already used to perform IRT and other analyses, in this study, I had the software to also obtain omega measures.

## Graphic representation of reliability across the Theta range

These other reliability alternatives still look at reliability as a single number, and measures at the average level of the trait to be measured, whereas IRT methods will assess reliability across all different levels of the latent construct, and reveal how reliability varies at different theta levels. This is best represented graphically on the scale level (per domain), and in PROMIS studies it is represented in plots with Theta levels on the x-axis, and on the y-axis either SE curves or reliability curves.

Another common IRT graphic at scale level is the Test information function (TIF) plot, which shows the combined effect of the discrimination information from the items. At item level, IRT trace line plots of Item Category Curves (ICC's – *not to be confused with Intraclass Correlation Coefficient*) are commonly reported, both in Rasch and in other IRT analysis. The ICC graph displays the IRT parameters for a single item, with a separate curve for each response category. At a glance, it can show where on the theta each response category is most sensitive. The height of the ICC curves represents the information, based on the discrimination parameter.

## Choice of reliability measures and expected results:

Many of the published PROMIS studies report only IRT based reliability measures, and do not even mention Cronbach's alpha. However, since separate audiences recognize the different reliability measures, all of these are included in the article. Readers with knowledge only of so-called Classical Test Theory methods will look for the Cronbach's alpha. The other reliability indices, omega_t, omega_h, empirical marginal reliability, and graphic representations of reliability from IRT, are included to be available for comparison with other studies also using omega, and to better represent reliability in an IRT context.

Quite frequently .7 is referred to as the accepted minimum for alpha, but this is an over-simplification, since Cronbach's original reference article actually says that a value between .7 and 1.0 can be considered satisfactory, and >.9 excellent, provided certain conditions are met.

What is acceptable depends on the circumstances. Hays and Reeve (R. D. a. R. Hays, B. B . , 2008) argue that a higher internal consistency reliability, at least >.9, is needed for measures used to make decisions based on the score of an individual, as is the case when using PROMs in clinical practice. A higher alpha should be expected for a more focused single construct questionnaire (i.e. measuring only anxiety), such as the seven short forms nested within PROMIS-57, and measuring alpha for a broad questionnaire that combines measurement of different constructs is not really appropriate (Taber, 2018). A classic psychometric book from 1994 (Nunnally & Bernstein, 1994) is sometimes cited for the .7 alpha threshold, but actually raises the bar further all reliability measures: "If important decisions are made with respect to specific test scores, a reliability of .90 is the bare minimum, and a reliability of .95 should be considered the desirable standard." p.264 (Nunnally & Bernstein, 1994).

In the article, Cronbach's alpha is represented as one for each of the seven domain scales, and one total alpha value for the entire PROMIS-57. A priori expectations for total alpha and domain alpha are not the same, as the domains cover very narrow constructs, and are expected to have very high alpha (>.9). The entire PROMIS-57 scale can be viewed as a general measure of HRQOL, combining different physical and mental scales into one. Thus, for the overall scale a somewhat lower alpha, is expected, yet well above .7, as the number of items will likely bias the result upwards.

## Methods for validity in a Classical Test Theory context

The purpose of validating a questionnaire is to check that it is actually measuring what it is intended to measure. As mentioned, the intention was to comply with the PHO standards for validation after translation, (PHO, 2014). Some of this can be done qualitatively, as was the case during the translation of PROMIS-57, but that is outside the scope of this quantitative theses.

### Concurrent and discriminatory validity

The validation in this study is first done first by comparing correlation coefficiants between scores for concurrent and discriminant validity, using Spearman's rho, since most score

distributions were non-normal. The rationale, method and results for this is described in the article.

### Factor structure

While internal consistency and reliability measures go a long way to show that the chosen attributes are coherently and accurately measured, they do not prove that each different item measures what they are supposed to. PROMIS-57 consists of seven domains, meant to cover three main areas within HRQOL; physical, mental and social (Cella et al., 2019).

Does this structural assumption hold up for the Norwegian translation as well? If not, the validity of the translation can be questioned. Some items may after translation capture something else than the original wording, or there may be cultural difference influencing what types of words correspond to the intended latent trait. Factor analysis was performed to shed some more light on this, and check how well this seven-factor structure holds up. In addition, factor analysis is helpful when checking some of the assumptions that need to be satisfied in order to employ IRT. There are many considerations and many choices for factor analysis.

### All-at-once or seven separate analyses?

One dilemma, then, is it more useful, to run the factor analysis for all of PROMIS-57 at once looking for the number of factors to match the seven domains, or to run a separate factor analysis on each of the domains, looking to confirm just a single factor. In this study, we did both. The benefit of doing an all-domains-at-once analysis is to examine whether items unintentionally co-vary across domains, or "stick to their own" as they should. A factor analysis performed on one domain at a time, will scrutinize the items even more closely within that one domain.

### Explore or confirm?

Exploratory factor analysis (EFA) seeks to use respondent data to develop hypotheses about the structure, to explore the covariance among items, in order to get a step closer to knowing what constructs we are actually measuring. Confirmatory Factor Analysis (CFA) compares the data to

an already established structure, to get a sense if any of the items are influenced by unknown covariables.

In the article only CFA, confirmatory factor analysis, is reported. Exploratory factor analysis was also performed. Principal component analysis is the first step of the analysis, to get a better sense of the structure, and how much of the variation is explained by the most obvious groupings of items, called factors. The output variables also become a part of the evidence for unidimensionality of each of the seven domains in PROMIS-57, one of the assumptions for utilizing IRT.

A third option, bi-factor analysis, (Steven P. Reise, Scheines, Widaman, & Haviland, 2013) emerged along the way. It is the method for obtaining the omega reliability measures, but also a very useful tool for a variety of exporatory factor analysis.

### Exploratory factor analysis

Exploratory factor analysis (EFA) starts with assessing the data with Principal Component Analysis (PCA) to find the best fitting number of factors (factor extraction), based on the eigenvalue of each factor, then rotates the factors to investigate the factor loadings – how strong the connection is for every item to each of the factors. Factors emerge from the analysis, not as the "truth", but as one of several plausible ways to group the items, and with factor loadings showing how strong a connection each factor has to each item. If factor loadings for an item are not clearly favoring the same factor as the other items in what we consider a domain, that could possibly indicate that external covariables are influencing the score of that item. Even when the expected factor structure is known, EFA can be used to explore the factor loadings with the chosen number of factors.

### Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA), on the other hand, is done by first specifying the model, this means setting up a model of the expected structure for the construct, then checking how well

the data fits this model. In either case, model fit indices are collected, to help assess how well the chosen method fits the data, and to what extent the results can be trusted.

If a seven-factor CFA model fits the data, meaning that the patterns of responses match expected responses estimated by the factor analysis, then it is an indication that this domain structure is appropriate for the translated instrument, as well. The fit indices become a measure of how well the factor model set up matches the data.

### Choice of Confirmatory Factor Analysis estimator

CFA cannot be performed in SPSS, so Lavaan package v 6.05 in R was used. There are a number of "estimators" to choose from, and it is important to choose one that is appropriate for the data. It is appropriate to consider domain scores to be on an interval scale, whereas each individual item, scored 1-5, must be treated as ordinal. Not all CFA estimators perform well with ordinal scales. Also, our data, and PROMIS data in other studies (Katz et al., 2017), are not normally distributed. The better estimator then, based on recent psychometric studies, is Weighted Least Squares with Mean and Variance
adjustment for the CFA model (WLSMV), and not Maximum Likelihood (ML) based ones (Li, 2016). Li describes the rationale for choosing WLSMV like this: "WLSMV, on the other hand, is specifically designed for categorical observed data (e.g., binary or ordinal) in which neither the normality assumption nor the continuity property is considered plausible. Although WLSMV makes no distributional assumptions about *observed* variables, a normal *latent* distribution underlying each observed categorical variable is instead assumed." The continuity property is not so much a concern since items with Likert scale scores are clearly at least ordinal and probably interval in nature, but avoiding bias from skewed variables is a high priority and worth the extra trouble of tracking down correct coding and interpretation for the WLSMV estimator.

### Interpretation of Confirmatory Factor Analysis output variables

There are two kinds of output from this CFA model that can help validate our translated questionnaire; fit indices and the residuals from the covariance matrix. The CFA is comparing

estimated data from the model with our actual data, and a number of "fit indices" are used to assess how close they match. The paper from Hu & Bentler is frequently cited for its fit indices (Hu & Bentler, 1999) and suggested cutoff values for "acceptable" and "good" fit, the Comparative Fit Index (CFI, >0.95 for good fit), Tucker-Lewis Index (TLI, >0.95 for good fit), and the Root Mean Square Error of Approximation (RMSEA, <0.06 for good fit). These fit indices are from other areas of science, and not all of the indices and suggested cutoff values are equally relevant for psychological measurement and HRQOL instruments with Likert scale responses. (K. Cook, Kallen, & Amtmann, 2009). The residual covariance matrix derived from the CFA gives an additional indication about the factor structure. The residuals should be small, or else it indicates that we have not captured the structure. (Reeve et al., 2007) Small residuals can be used an indicator of Local Independence, which will be covered later.

### Bi-factor model

Bi-factor analysis is testing to see how much variation can explained by a general factor, and two or three group factors are set up in addition, to capture the residual variance, or variation not explained by the general factor. (Steven P. Reise et al., 2013) The minimal way of using it is to extract the omega values, and an indicator called Explained Common Variance (ECV). The ECV is simply the percentage of variance explained by the general factor, and is an expression of unidimensionality, meaning to what extent a single factor relates to all the tested items, (Sijtsma, 2009). When the bi-factor analysis is performed in R, an RMSEA fit index is also reported for comparing the fit of an alternative single factor model and for the fit of the bi-factor model, to indicate which model better fits the data. Also, there is graphic representation of the factor model.

### Factor analysis as premise for Item Response Theory

Much of the quantitative validation work in this study is centered around factor analysis and IRT. A number of assumptions about the data have to be met in order to trust the IRT results, and factor analysis is needed to satisfy these assumptions. Output from EFA helps establish

unidimensionality, and the residual covariates from CFA help establish Local Independence, a concept that will be covered shortly.

## Methods for validity in an IRT context

### Item Response Theory assumptions

The necessary assumption for IRT are only briefly covered in the article. In order to use IRT analysis and be able to trust the results, a number of assumptions have to be met. Unidimensionality and local independence should be evident among items within a scale, and monotonicity and invariance should be apparent in the scores. The article covers the methods used for this, and a detailed discussion of each is beyond the scope of this thesis. Methods from previous research on PROMIS measures that are recommended for this (PHO International, 2013) were mostly chosen.

Unidimensionality, i.e. that each scale or domain measures a single latent trait (Stochl, Jones, & Croudace, 2012) is not an absolute term, so the point is to show that a scale is sufficiently unidimensional. Local independence means that each item is contributing uniquely to the latent trait being measured and are not influenced by something outside of what is being measured (Reeve et al., 2007). Both these assumptions can be tested with different methods, in this study they were tested through factor and bi-factor analysis.

Monotonicity – that the "probability of endorsing or selecting an item response indicative of better health status should increase as the underlying level of health increases" (Reeve et al., 2007) is best tested by the Mokken scale analysis – which has its own package in R (Ark, 2007). The last IRT assumption is invariance. A scale is said to have "measurement invariance (also known as measurement equivalence) across groups if subjects with identical levels of the latent construct have the same expected raw-score on the measure" (Hirschfeld & Von Brachel, 2014).

Invariance can be tested with multigroup factor analysis, although in this study the IRT based differential item functioning (DIF) analysis method was chosen, using the R package lordif (Choi, Gibbons, & Crane, 2011). However, DIF analysis is not just an assumption check for IRT, but an important part of checking the quality and validity of the translation in its own right.

## Item Response Theory vs. Rasch

While most previous validation studies of PROMIS have been performed 2PL IRT models like the graded response model (GRM), some have also used Rasch analysis (Hung, Voss, Bounsanga, Crum, & Tyser, 2017), (Hung et al., 2018), (Petrillo, Cano, McLeod, & Coon, 2015). Rasch model is quite restricted, allowing only one parameter to be estimated, and the basic idea is that items either fit well or fit poorly, and should be removed. The other IRT methods are used with more flexibility. There are many IRT models, and modifications to each, and the basic concept is that no perfect fit exists between model and data, so the model with the best fit should be used. Whether to use a 2PL IRT model or Rasch model for analysis can best be determined after reviewing fit indices for both. Fit indices check estimated data against observed data in different ways, thereby testing the chosen model against the data. In factor analysis, these indices check that the chosen factor structure and number of factors fit the data. In IRT, the fit indices confirm that the estimated IRT parameters are reliable, provided that the assumptions also have been satisfied.

## Model fit indices for choice of Item Response Theory model

Each item can be tested for how well they fit in the chosen model. There is wide consensus among previous studies to use the s-x2 test for item misfit, a variant of chi-square tests, but performed on estimated IRT parameters, (Depaoli, Tiemensma, & Felt, 2018), and the cutoff is simply a significance value <.001.

Model fit, on the other hand is not an absolute "yes or no", but rather a matter of degrees of fit. The model fit indices were obtained by running IRT analysis for each model in each of the seven domains, using mirt package in R, and looking up the output for "M2 test type C2" (Chalmers, 2012), then considering those preliminary results to assess the strength of each model. The same cutoffs for fit indices that are used for factor analysis, also apply to IRT model fit (Hu & Bentler, 1999). The different indices measure different aspects of model fit. The Root Mean Square Error of Approximation (RMSEA) was applied to test how well the model fits the data relative to its degrees of freedom. It is common for HRQOL questionnaires to not meet the established RMSEA cutoff of <.06 (K. Cook et al., 2009).

Standardized Root Mean Square Residual (SRMR or SRMSR) were used to indicate how well the model captures the data, after comparing observed and predicted correlation matrices.

Tucker Lewis Index (TLI) and Comparative Fit Index (CFI) values express estimated differences between the examined model and a hypothetical (null) model where none of the components in the model are related. They share the same cutoff value CFI > .95, interpreted as indicating that estimates and observations are highly correlated, and indicating the model fits the data.

The Bayesian Information Criteria (BIC) is different from the others, mainly in that it gives relative values of very different magnitudes, that can only be interpreted in comparison between models. The model with the lowest possible BIC has better fit (K. Cook et al., 2009).

The most important point of all this is that any single one of these methods cannot be relied upon alone, rather it is the total picture that indicates model fit. The original article that established these thresholds also says they should not be applied as absolute cutoffs, especially not RMSEA (K. Cook et al., 2009). The more of these indices that meet or approximate the thresholds, the better.

### Choice of method for Differential Item Functioning analysis

The DIF analysis is another IRT analysis that is used to quantify the invariance, meaning fairness of measurement for all different groups of people. The discrimination parameter and difficulty (severity) parameters for two groups are estimated, and the difference is quantified and also represented graphically, after controlling for theta level of the measured trait.

Two methods were used in this study for performing DIF analysis, both using 'lordif' package in R which uses ordinal logistic regression models (Choi et al., 2011). There is a clear recommendation from the PHO (PHO, 2014) to use lordif and with McFadden's pseudo R2 change of 2% as the criteria for flagging DIF. First a chi-square based DIF analysis was applied, then the "pseudo R2" method. The chi-square method is a lot more sensitive, flagging issues that turn out to be of no consequence for the scoring. The advantage of doing chi-square first, in spite of possible type 1 errors, is that it allows the identification of 'anchors', items that are sure to be

free of DIF, and those can be used as a reference in the ‘pseudo R2’ method, to test the less certain items against the certain ones. The process of iteratively selecting anchors and assessing the magnitude of DIF is well described by Teresi (Teresi et al., 2009).

# Results

## Sample characteristics

All 408 complete responses were collected and used in the analysis. Since there was no tracking of IP address or other identifying characteristics, there is a slight possibility that the same person has responded twice, but not likely, considering the burden of having to respond to over 100 items. Our intention of reaching a general population sample was partly accomplished, however respondents appear to have a higher education and possibly more health problems than the average population, and also a higher proportion of women (74%) – see Table 1 in the article for details.

| **Table 1. Sample characteristics: age, gender, cohabitation, income, education and health status (N=408).** | | |
|---|---|---|
| **Age –** mean (SD * / min-max) | | 52 (13 / 19-88) |
| | | **n(%):** |
| **Women** | | 310 (76) |
| **Living alone** | | 102 (25) |
| **Income source** | Employed, part or full time: | 215 (53) |
| | Retired | 57 (14) |
| | Permanent disability | 79 (19) |
| | Sick leave, short or long term** | 42 (10) |
| | Other ** | 15 (4) |
| **Income level*** ** | Low (<350k NOK) | 124 (31) |
| | Middle (350k-600k NOK) | 183 (45) |
| | High (>600k NOK) | 96 (24) |
| **Education** | College level or higher | 298 (73) |
| | Intermediate | 89 (22) |
| | Elementary only (>10yr) | 21 (5) |
| **Health problems, self reported** | Physical health problems | 166 (41) |
| | Mental health problems | 18 (4) |
| | Both physical and mental | 94 (23) |
| | No health problems | 130 (32) |

*SD=Standard deviation, **= away from work >12 month duration, «arbeidsavklaringspenger», ***=homemaker, student, no response or marked as «other»

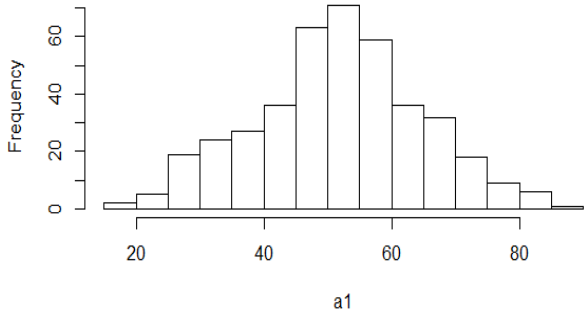
## Age distribution

Only adults were recruited, and the age range of respondents is 19-88. The age histogram (Fig. 1.1a) shows a normality distribution, while the total population age curve for adults is fairly flat from age 18 to 50 (Fig. 1.1b), then declining beyond that, meaning 18-30 year olds are underrepresented in the sample.

Figure 1.1: Age distribution of the PROMIS-57 study sample(1a) and Norwegian population (1b)

Figure 1.1a

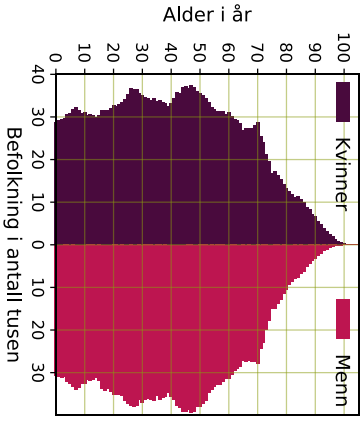
Age distribution all respondents, n=408




1.1b

Age distribution for all citizens Norway 2019

Source: Statistisk sentralbyrå (SSB, 2020)




## Gender association with other variables

Since there are many more women in the sample, significance testing was done to explore whether that could introduce other types of systematic bias in the data. Significance testing was performed to check for gender difference in age, cohabitation status, income, and prescription medication use. The difference in respondent age with Independent samples t-test is significant at .002, women on average 4.7 years younger.

Chi-square test for independence (with Yates Continuity Correction) indicated no significant association between gender and living alone, x2 (1, n=408) = .462, no significant association between gender and three income categories, x2 (1, n=408) = .032, no significant association between gender and taking medication prescribed by a doctor, x2 (1, n=408) = .064.

## Results from statistical analysis on Sleep Disturbance

The results of the analysis are well covered in the article, although all the details could not be included because of word restrictions. Rather than repeating and expanding the presentation of

all results, this section of the thesis will highlight some results that deviate from the overall picture, and deserve more attention. Among the seven domains in PROMIS-57, one has somewhat less convincing results than the others. Sleep Disturbance has worse – or at least different - psychometric properties than the other domains. The results for Sleep Disturbance will be presented in more detail, beyond what is in the article, followed by a discussion of these results, and some of the statistical methods used.

Table 2: PROMIS-57 per domain score, reliability and validity variables

| | PF | Anxiety | Depression | Fatigue | Sleep Dstrb | Social | Pain Intf | Pain Intensity NRS |
|---|---|---|---|---|---|---|---|---|
| Mean T-score (SD*) | 47.6 (10.6) | 50.8 (11.0) | 51.3 (13.0) | 52.3 (11.0) | 52.6 (9.9) | 48.3 (12.2) | 55.0 (11.8) | 3.5** (2.8) |
| ceiling% | 36.4 | 0.0 | 1.0 | 2.4 | 1.5 | 21.5 | 6.8 | 1.2 |
| floor% | 0.5 | 24.2 | 29.6 | 17.6 | 2.7 | 3.9 | 29.1 | 19.6 |
| Cronbach's alpha | .97 | .96 | .97 | .98 | .92 | .98 | .98 | |
| McDonalds omega ω_t (ω_h)*** | .97 (.96) | .96 (.95) | .97 (.96) | .98 (.98) | .92 (.91) | .99 (.99) | .99 (.99) | |
| IRT Marginal reliability | .87 | .91 | .89 | .94 | .92 | .93 | .90 | |
| IRT discrimination mean (per item min-max)**** | 5,9 (4.3 - 7.8) | 4,7 (3.8 - 6.1) | 4,8 (3.7 - 5.6) | 7,4 (4.8 - 10.8) | 4,0 (1.3- 10.2) | 7,5 (5.2 - 10.0) | 8,5 (5.6 - 10.6) | |
| Eigenvalue ratio from 1-factor EFA | 12:1 | 10:1 | 16:1 | 33:1 | 5 :1 | 54:1 | 26:1 | |
| Explained Common Variance (ECV) | .88 | .86 | .88 | .96 | .74 | .96 | .94 | |

*=Standard Deviation ** Pain NRS mean, not T-score *** ω_t=omega total, ω_h = hierarchical
All confidence intervals (95%) for alpha and omega are <+/-.01, except Sleep: +/-.02
****IRT Discrimination parameter from Graded Response model

## Sleep disturbance results at the scale level

**Distribution:** The Sleep disturbance scale score is normally distributed in the sample, with no floor or ceiling effect.

Sleep44, Sleep67, and Sleep72 have zero-inflated (skewed) scores, while the other five items do not, resulting in the appearance of an overall normality of the scale score.

**Reliability:** The different internal consistency and reliability measures are excellent.

**Correlations:** Sleep disturbance correlates only moderately with the other six PROMIS-57 domains: $r_s$.54 - $r_s$.61, and with the RAND 36 sumscores: $r_s$ .44 - $r_s$ .62. (Table 5 from the article)

Table 5: Spearman rank correlations within PROMIS-57 domains and against RAND-36 sumscores

| PROMIS: | Physical function | Anxiety | Depression | Fatigue | Sleep | Social* | Pain** | Pain NRS*** |
|---|---|---|---|---|---|---|---|---|
| Physical function | 1.000 | -.409 | -.501 | -.750 | -.541 | .822 | -.815 | -.741 |
| Anxiety | -.409 | 1.000 | .759 | .591 | .547 | -.532 | .438 | .449 |
| Depression | -.501 | .759 | 1.000 | .642 | .546 | -.585 | .509 | .462 |
| Fatigue | -.750 | .591 | .642 | 1.000 | .608 | -.857 | .728 | .688 |
| Sleep | -.541 | .547 | .546 | .608 | 1.000 | -.593 | .547 | .533 |
| Social* | .822 | -.532 | -.585 | -.857 | -.593 | 1.000 | -.774 | -.691 |
| Pain** | -.815 | .438 | .509 | .728 | .547 | -.774 | 1.000 | .918 |
| Pain NRS | -.741 | .449 | .462 | .688 | .533 | -.691 | .918 | 1.000 |
| PROMIS: / RAND 36: | Physical function | Anxiety | Depression | Fatigue | Sleep | Social* | Pain** | Pain NRS*** |
| RAND36 PF PHYSICAL | .880 | - .329 | - .422 | - .675 | - .513 | .751 | - .781 | - .731 |
| RAND36 RP ROLEPHY | .786 | - .420 | - .479 | - .738 | - .509 | .794 | - .737 | - .688 |
| RAND36 BP BODILYPAIN | .793 | - .414 | - .468 | - .713 | - .526 | .741 | - .927 | - .918 |
| RAND36 GH GENERAL | .776 | - .524 | - .558 | - .776 | - .620 | .785 | - .718 | - .681 |
| RAND36 VT VITALIT | .715 | - .560 | - .632 | - .864 | - .617 | .827 | - .670 | - .622 |
| RAND36 SF SOCIAL | .785 | - .517 | - .587 | - .827 | - .597 | .885 | - .743 | - .683 |
| RAND36 RE ROLEMOT | .389 | - .545 | - .584 | - .524 | - .441 | .488 | - .417 | - .432 |
| RAND36 MH MENTAL | .467 | - .727 | - .806 | - .644 | - .560 | .574 | - .480 | - .451 |

*= Social roles and activities ability **= Pain interference ***Pain intensity numeric rating scale

**Unidimensionality**:

From bi-factor EFA analysis the Explained Common Variance (ECV) is 74 (threshold>60). All other domains: 86 – 96.

Bi-factor analysis model fit sensitivity testing: single factor EFA gives RMSEA= .27, general factor and three main factors RMSEA= .06, indicating good model fit for the multidimensional model and poor for a strictly unidimensional one.

The ratio of general factor eigenvalue-to-max group factor SLP: 5.6 :1, all other domains 10:1 – 54:1, (suggested threshold >4 :1).

Table 1.2 PROMIS-57: CFA with single factor, WLSMV estimator, and scaled indices (thresholds)

| | RMSEA (<.06) | SRMR* (<.08) | CFI (>.95) | TLI (>.95) |
|---|---|---|---|---|
| Sleep | .22 | **.07** | **.99** | **.98** |
| Other domains | .08 - .16 | **.01 - .02** | **.99 – 1.00** | **.99 – 1.00** |

*=unscaled **Bold** = cutoff criteria are met

**IRT model fit:**

The fit indices applied to the GRM IRT model finds fairly unsatisfactory model fit for SLP, with these values (cutoffs): RMSEA .22 (<.08) SRMSR .08 (<.06) CFI .86 (>.95) TIL .90 (>.95) Model fit for all seven short forms and three different IRT models in Table 1.3

Table 1.3: PROMIS-57model fit indices for comparing three IRT models Rasch / Graded Response / Generalized Rating Scale, n=408

| Thresholds: | Physical Fct | Anxiety | Depression | FAT | SLP | SOC | PAIN |
|---|---|---|---|---|---|---|---|
| BIC (lowest=best) | 5200/5108/**5068** | 5352/5258/**5202** | 5536/5447/**5350** | 5838/**5500**/5501 | 8057/**7731**/7781 | 5674/5367/**5299** | 4863/5220/**4824** |
| RMSEA <.06 | .107 / .115/ .116 | .091/.082/.076 | .095/.098/.086 | .138/.103/.106 | .209/.227/168 | .136/.116/.095 | .145/.186/.138 |
| SRMSR <.08 | .086 /**.027 / .040** | .092/**.025/.034** | **.075/.029/.030** | **.012/.013/.025** | .123/.081/.103 | .120/**.013/.020** | .119**/018/.027** |
| TLI >.95 | **.098 /.098 /.097** | **.099/.099/.099** | **.098/.099/.099** | **.098/.099/.099** | .877/.856/.921 | **.978/.983/.989** | **.974/.958/.977** |
| CFI >.95 | **.098 /.098 /.097** | **.099/.099/.099** | **.099/.099/.098** | **.098/.099/.098** | .882/.897/.884 | **.978/.988/.984** | **.975/.970/.966** |
| # of criteria met: | 2 / 3 / 4 | 2 / 3 / 4 | 3 / 3 / 4 | 3 / 4 / 3 | 0 / 1 / 0 | 2 / 3 / 4 | 2 / 3 / 5 |

**Bold** = within suggested thresholds (Hu & Bentler, 1999)

## Sleep disturbance results at item level:

**Local independence** appears to be supported, based on the CFA residuals method at .2 cutoff, but the other issues led me to additionally explore lower cutoff of >.1, and two other methods: the Q3 and the Chen & Thissen(Chen & Thissen, 1997). These more sensitive methods flag Sleep items for Local Dependence (LD, means a lack of local independence) to a greater degree than other domains. With the Q3 method, 21,4% of possible Sleep item pairs are flagged, but no pairs in the other domains. Chen & Thissen results are

Fatigue 2 pairs out of 28 possible=7.1%

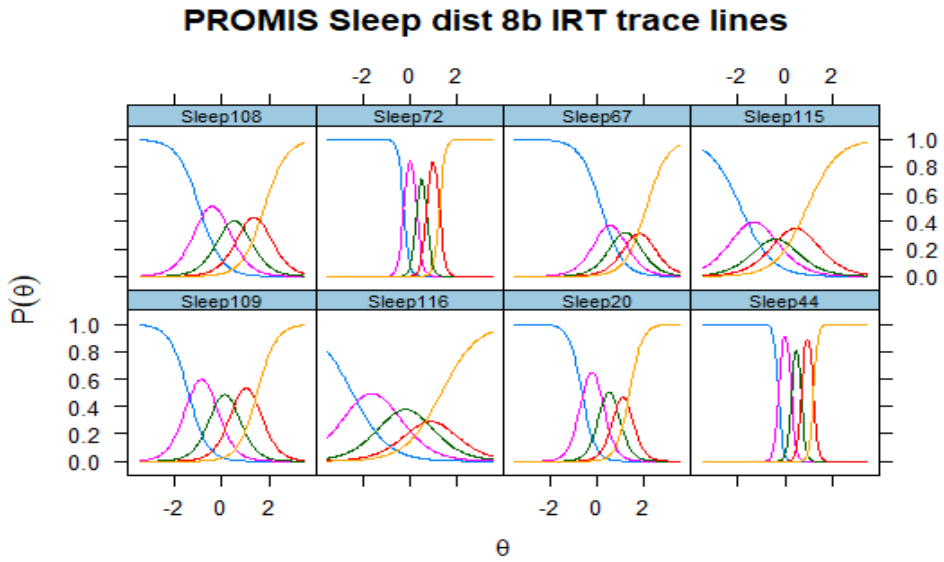
Sleep 1 pairs out of 28 possible=3.6%

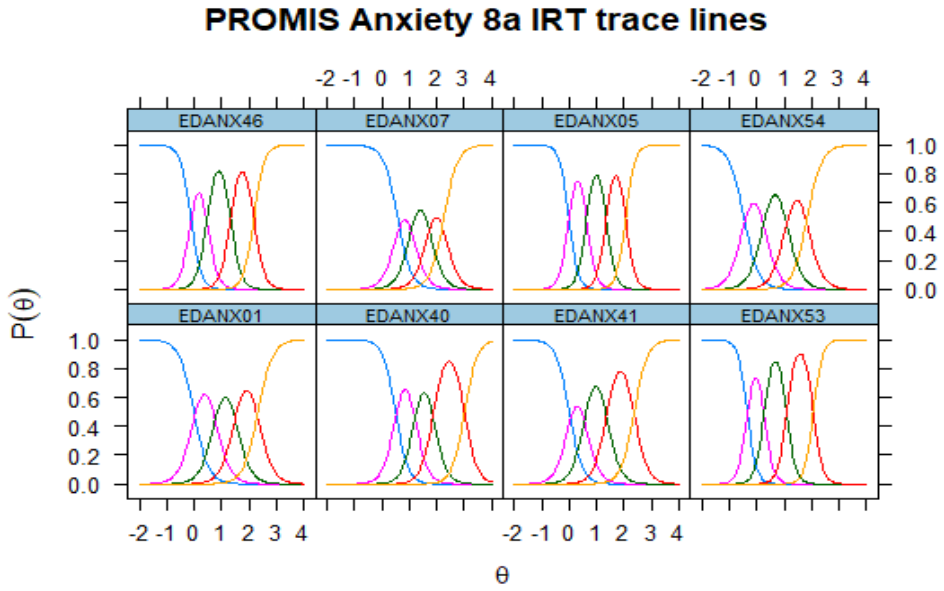
Social 1 pairs out of 28 possible=3.6%

The Items Sleep109 and Sleep115 are included in LD pairs by all three methods, and Sleep116 by two.

**IRT trace line plots** of Item Category Curves (ICC's) for every item in the SLP are displayed in Fig. 1.2 shows plots for the ANX items for comparison. ICC's are neatly distributed for most other domains, but are at-a-glance identifiable as highly variable for the Sleep Items, some low and stretched out (Sleep67 and 115), indicating higher measurement error and reduced IRT information, while others are too steep (Sleep 44 and 72), indicating exaggerated discrimination parameters. Others, still have totally overlapped curves (Sleep 115 and 116), indicating that as many as 5 response categories may not be necessary, though just one or two items is insufficient justification for changing the number of response elements. Item characteristic Curves from An

Figure 1.2 PROMIS-57 IRT ICC (=Trace line plots) curves for Sleep and Anxiety




**Item fit** Two SLP items, Sleep 44 and 72 are significant (<.001) for poor item fit, when assessed using the S-X2 statistic, which calculates the differences between observed and expected responses under the GRM model. A p-value of the S-X2 statistic <0.001 for an item is considered as item misfit (Luijten et al., 2019). No other domains have misfit when testing one domain at a time.

## Discussion

The main findings are discussed in the article, but the results for the Sleep Disturbance domain are different and interesting, and there are some questionable results from the analysis. It may be related to a more limited set of problems.

### Representative sample?

The 408 respondents is arguable a very small percent of the total potential recruitment, but hopefully representative to a sufficient degree. For testing the IRT characteristics of a questionnaire it is not essential that respondents belong to a pure general population sample, and a mix of healthy and chronically ill respondents can actually be beneficial for testing the full range of latent variables (Amtmann et al., 2010). If anything, our sample turned out, in spite of their percentage of self-reported health problems, to have too many healthy individuals for the analysis, causing some of the scales to have a skewed distribution, floor or ceiling effects, and even what is called zero-inflated distribution; a group of respondents scoring at the very lowest end, while the rest of the sample is normally distributed (Smits, Öğreden, Garnier-Villarreal, Terwee, & Chalmers, 2020). Age is normally distributed, which means young adults are under-represented, as shown in Fig.2. However, many patient populations have very few individuals 18-30, so this is perhaps less of an issue than if the sample was missing the middle-aged to older age range.

The much smaller number of men than women may also threaten how representative the sample is. Most of the other demographic variables except age were found to not be significantly different between men and women. While there may be other unexplored differences, the lack of significance relieved some of the concern with having a gender imbalanced sample. In many patient populations, there are more women than men, so representativeness would also have been more of a concern had there been mostly men in our sample.

## Normality and zero-inflation.

Sleep disturbance is the only domain within PROMIS-57 where the score appears to be normally distributed. The other six domains are not. A skewed score distribution can be viewed as a floor or ceiling effect in a questionnaire, that has failed to capture one side of a normal distribution in the total population. It might also be an accurate reflection of an asymmetric latent construct in the population, where the "normal" or at least the most common condition is to be symptom free. Most PROMIS short forms come through as normally distributed when applied to a sick or disabled sample, as in a study of rheumatoid arthritis patients (Bartlett et al., 2015), but tend to show ceiling effect in a healthy population.

## Why is only the Sleep Disturbance score normally distributed in the sample?

The sample distribution may or may not truly represent the total population distribution. The "normal" distribution of the PROMIS-57 Sleep Disturbance (SLP) domain indicates that few people in the sample are free of sleep problems. This could simply be "the sign of the times", since smart phones provide 24/7 information and entertainment overload, causing more people now to have trouble sleeping and falling asleep to some degree, maybe more than when the scale was developed 10 years ago. A study of 50 000 Norwegian students (Sivertsen et al., 2019) found a high prevalence of lack of sleep and a substantial increase in insomnia over the last ten years. SLP being more normally distributed than other PROMIS domains was also found in another fairly recent study of PROMIS 29 applied to different patient populations (Katz et al., 2017). Sleep Disturbance tended towards normality, while the other domains were mostly skewed. Katz also reported floor/ceiling effects for other domains, but not SLP.

## Distribution for each item

Looking at item by item distribution can also provide some important insights. While SLP as a scale is normally distributed, all the individual SLP items do not contribute equally to this normality. Histograms of every single SLP item show that the items do not contribute equally to that normality. Sleep44, Sleep67, and Sleep72, the three items that do not have a normal

distribution appear to be conceptually inter-related. They are all about falling asleep, while the others are about staying asleep and getting sufficient rest. This makes these three items interesting to look at in terms of Local Independence. Item 44 and 72 also have an almost identical wording and meaning. This is also the case in the English language original, as pointed out by Teresi (Teresi, Ocepek-Welikson, Cook, et al., 2016), so it is not brought on by the translation.

### Score polarity reversal - monotonicity

Articles by Jensen et al (Jensen et al., 2016) and Teresi and Jones (Teresi & Jones, 2016) both pointed out that the items in SLP short form 8 are a mix of positively and negatively worded items, which appears to confuse some respondents. The scoring is designed to make sure the scores still are high for a greater amount of the measured domain. However, some individuals may not pay full attention to the words, and assume that they all have the same polarity, or direction. In the context of PROMIS-57, respondents get to the SLP items after answering 32 other items, by which time some respondents may have become a bit more careless with their responses, and not notice that responding "quite a bit" means great sleep in one item and poor sleep in the next. This is supposed to be picked up by the monotonicity testing, reported in the article as acceptable. In addition, this was tested by creating two new sub-scores for Sleep, one for the negative and one for the positive items. There turned out to be correlation coefficient of .73 between the two, while the average correlation between Sleep items is only .58. Together with the somewhat lower internal consistency measures for SLP, there may be a slight degree of polarity reversal issues, but not enough to conclude that this causes any kind of bias.

### Assumptions for Item Response Theory analysis

IRT analyses, generating discrimination and theta parameter from the GRM 2PL models, should not be performed without first checking the necessary assumptions for IRT modelling. The article discusses IRT results, but the four assumptions are mentioned only very briefly, so the details of unidimensionality and local independence are covered a bit more in this section.

## Dimensionality

Unidimensionality is a necessary assumption for doing IRT analysis, but also an important element in the validity of a questionnaire. As a whole, PROMIS-57 is only supposed to be moderately unidimensional, to the limited extent that HRQOL is a single dimension. Relative to for example measuring mathemathics skills or body height, HRQOL can be viewed loosely as a different separate dimension, but under scrutiny, it becomes apparent that it consists of many separate dimensions. Dimensionality is a relative term, then, and there are a few different approaches to examining how close, yet unique and separate the items are. The results presented in the article support unidimensionality for Sleep Disturbance (SLP), also, but with less firm indices than the other domains.

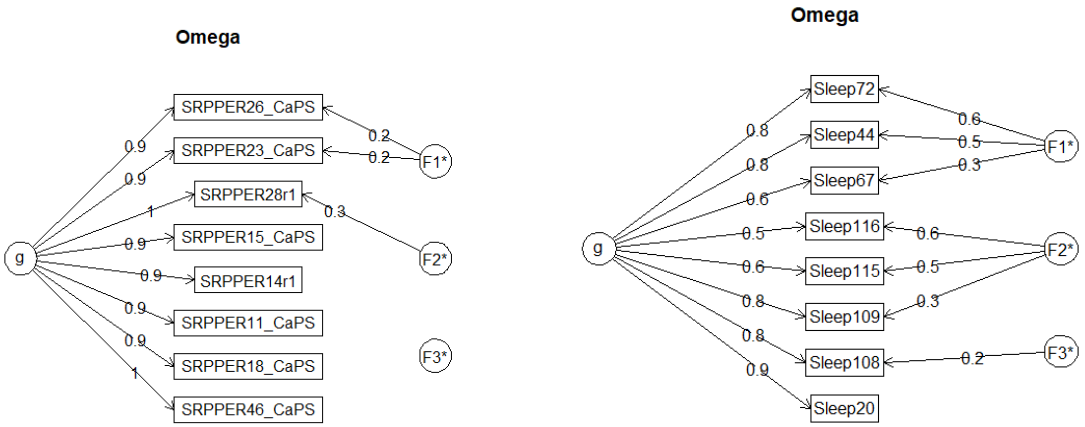
The bi-factor model analysis is a good method for looking at dimensionality. A unidimensional set of items will in a bi-factor model have a large portion of its variability explained by the general factor of the model. The explained common variance (ECV) provides information about the degree of unidimensionality based on the observed variance-covariance matrix of a bi-factor model (Sijtsma, 2009). The 74% ECV for SLP in this study indicates acceptable unidimensionality. Two recent PROMIS studies described lower ECV than that, 67 and 68, as evidence for "sufficiently unidimensional" item banks (van Bruggen, Lameijer, & Terwee, 2019 )"and (Lameijer et al., 2020), based on a cutoff of 60 from a reference study (Steven P. Reise et al., 2013). The other domains have even higher ECV (86-96) in this sample.

The bi-factor model gives the opportunity for items to show connection to group factors in addition to the general factor. Testing a firmly unidimensional domain with bi-factor model will show support mostly for the general factor, as visualized with the Social domain (SOC) results. Group factors loadings to the right of the figure for SOC are too weak to be considered (<.3), and general factor loadings to the left very strong (>.9). Not so with SLP.

Inspecting the visual output in Fig 1.3 helps illustrate this. Factors can have a loading between 0=no relation, and 1.00=strongest possible. There are two clear group factors with between .3 and .6 factor loadings, and the general factor loads between .5 and .8. for the different items. The three items already mentioned that measure problems falling asleep are forming a group factor, and the three items that are form the other group on closer scrutiny are all about getting enough sleep.

Figure 1.3 :

Bi-factor model visualization for Social roles and activities (left) and Sleep disturbance (right)




In conclusion, sufficient unidimensionality is supported for all PROMIS-57 domains. At the same time, unidimensionality is not a straightforward concept, and there are clearly subcomponents to the Sleep Disturbance domain to a greater degree than the other domains, but not to the extent that it would bias the result of an IRT model. Multi-dimensionality may have a negative effect on reliability estimates and IRT parameters, and may need to be explored with method beyond the already generous scope of this thesis.

### Local Dependence – LD

Local independence is also assumption for IRT. As already mentioned, it assumes that the items are only related to the construct (the dominant factor) being measured and not to other constructs (any other factors). This implies that, after controlling for the dominant factor, there should be no significant covariance between item responses. LD analysis checks the residual covariances for every possible pair of items against each other, applying a predetermined threshold value.

### Additional LD analysis

There are many different approaches to producing a covariance matrix for LD analysis, and a few additional methods that appeared in different reference articles are included, to explore whether they would give the same results. For the most part they all concluded that six PROMIS-57 domains are sufficiently locally independent, while SLP may have some issues.

The different methods have different ways of examining the data, and different sensitivity, so there is a definite risk of type 1 errors, or highlighting issues of no practical consequence. The most accepted method in PROMIS studies is looking at the residual covariance matrix from a single factor CFA (in R package 'lavaan') and applying a <.2 cutoff (Reeve et al., 2007). This method applied to this data identifies no LD in any domain. Less restrictive cutoff values suggested in early PROMIS studies were then applied: .13 is suggested by Amtmann et al (Amtmann et al., 2010), and >.1 is suggested for questionnaire development (PHO International, 2013). This identified only two of the PROMIS item pairs in the sample, one >.13 and one more >.1

Another common method, Yen's Q3(Christensen, Makransky, & Horton, 2017), used for LD identification in at least 14 international PROMIS studies, is based on the residual matrix from the IRT model, using 'mirt'package in R. The cutoff is relative to the covariance residuals by first calculating the mean residual, then setting the threshold to .2 above the mean (Christensen et al., 2017). This flags six pairs, 21% of possible pairs in SLP, for LD, and also indicates LD pairs in some of the other domains; PF 2/28=7.1%, ANX 0%, DEP 1/28=3.6%, FAT 1/28=3.6%, SLP 6/28=21.4%, SOC 1/28=3.6%, PAIN 2/28=7.1% Another variety of Q3 is called the Jack-knife Slope Index (JSI). All of these more restrictive methods flag Sleep items for Local Dependence (LD) to a greater degree than other domains, as 21,4% of possible SLP pairs are flagged with the Q3 method. Those happen to be the ones with reversed scoring, asking about positive sleep questions. There is no established threshold for acceptable percentage of item pairs with LD. This makes the Local independence of the Sleep items questionable, and there is a possibility of some biased IRT estimates for the SLP, though also the possibility of "false positive" type 1 errors.

## IRT - Why the steep slopes?

Steep slopes in IRT plots are equivalent of high discrimination parameters, the second parameter in 2PL IRT. Discrimination contributes to reliability, in the sense that it increases the amount of estimation information. High discrimination is a sign of well performing items, up to a point. Beyond 5.5, however, it becomes problematic (source: verbal conversation with psychometrician

Aaron Kaat). The high parameter may be accurate, or falsely inflated by some other factor. Possible reasons for type 1 error needs to be considered. There are a few possible reasons for over-inflation of the discrimination. The psychometric literature on this is not abundant, but there are a few possible explanations.

### Possible causes of steep slopes

1. Local independence violations can cause inflated IRT discrimination . There is some amount of LD, but only partly involving the same domains and items that have steep slopes.

2. The sample may be too skewed, or actually zero-inflated, meaning many responders are scoring the very end of the scale (raw score = either 8 or 40), while the rest are normally distributed. There is support for the idea that skewed data with lots of «symptom-free» responders, or "non-cases", give hyperinflated slopes in Reise, Rodriguez, Spritzer& Hays, 2018 (Steven P. Reise, Rodriguez, Spritzer, & Hays, 2018), referring in part also to Wall et al(Wall, Park, & Moustaki, 2015). A recent simulation study by Smits et al (Smits et al., 2020) did not find problems with skewed data, unless they were zero-inflated. The study found inflated discrimination slopes in the graded model with zero-inflated scores, and suggests 1.5 to 2 points increased bias of discrimination when 25% of the sample is a "non-case". This sample has an even greater number of "non-cases", based on the histograms of raw scores per domain (Figure 1 in the article). The one exception is SLP, which appears quite normal-distributed. Three Sleep items are in spite of that heavily skewed, and two of those (Sleep44 and 72) also those have very high discrimination (10.8 and 8.7, respectively) . However, the third one, Sleep67 is the most skewed (skewness=1.3), but has a discrimination of only 2.2. Something more is at play.

3. The sample size could be inadequate for the analysis, in the presence of non-normal distribution. Simulation studies looking into sample size accept n>200, (Depaoli et al., 2018), but caution that this depends on a few other factors. Model complexity and too few respondents endorsing some of the categories can bias the parameters estimated from the model (Forero, 2009). The COSMIN criteria (Mokkink et al., 2018) sample size recommendation for IRT analysis is also >200 per group. The PHO recommendations for basic validation after translation (PHO, 2014) accepts >200 per group for DIF analysis with IRT, as well. However, the

requirements listed in another section of the same document sheds some light on the very high slopes. PHO recommendation for IRT item calibration of translated item banks (as opposed to short forms) versions is much higher, minimum 500 and ideally >1000. I do not have an entire item bank to perform a calibration study on, so I did not at first notice this quote: "PROMIS recommends a minimum of 500 subjects per item (i.e. each item should have been completed by at least 500 subjects). It should be noted that this sample size may be adequate for estimating item parameters, but may be too small for other analyses, such as computing item and test information functions. Also inflated discrimination parameters can be a problem. Therefore, a more optimal sample size would be 1000 to 2000 subjects per item." …

… "Reise and Yu concluded that at least 500 subjects are needed to achieve an adequate calibration under the graded response model. However, for good estimations of the easiest and most difficult items, they recommend 2000 subjects." (PHO, 2014) referring to (Steve P. Reise & Yu, 1990).

The issue, then both with discrimination slopes is quite possibly an insufficient sample size, exaggerated by zero-inflation, and insufficient variation, since that "robs" the sample of respondents to provide information across the entire theta range. For this reason these IRT results may be insufficient for doing a Norwegian calibration of the PROMIS-57 scores. The international recommendation is to use the US PROMIS reference data, unless a national large scale calibration study has been performed. The language DIF analysis results indicate that there is no significant cultural bias between US and Norwegian version. Any bias, whether caused by cultural differences or by a poor translation, would show up as DIF. Still further studies in larger samples should be performed to replicate and support these findings.

# Conclusion:

## General conclusion from the article:

The results and discussion in the article are independently  answering the research questions for this thesis.  The additional discussion har been attempt at getting a little deeper into the question of why one domain shows less convincing results.  Those additional explorations into a difficult territory has somewhat added to the understanding, but mostly highlighted more topics for future research.

By traditional measures, such as Cronbach's alpha and correlation comparisons for concurrent and discriminant validity, the Norwegian PROMIS-57 has excellent reliability and validity.  The same goes for all the shorter versions that are nested within PROMIS-57: PROMIS 29, and the eight, six and four item short forms for each of the domains Physical function, Anxiety, Depression, Fatigue, Sleep Disturbance, Social Roles and Activities Ability, and Pain Interference.  Examining factor structure, IRT parameters and IRT based reliability, and testing invariance across languages by DIF analysis, the results are also satisfactory, although there are some borderline results for the Sleep Disturbance scores.

The answer to the overall research is that Norwegian PROMIS-57 has satisfactory psychometric properties, and can be recommended for use in research and clinical practice, as it has excellent reliability and sufficient validity, including concurrent and discriminant validity, a confirmed factor structure and no detected language DIF and no age DIF.

Item fit and model fit in an IRT context is acceptable, and the Standard Error plot and ICC plots provide visualization of measurement characteristics indicating PROMIS-57 as valid in populations that have somewhat worse symptoms and HRQOL than the general population.

PROMIS 29 can be viewed implicitly as having similar reliability and validity as PROMIS-57, but with some loss of measurement precision as indicated by a slightly narrower range for standard error of measurement, and lower information precision in the IRT model.  Respondents may relate favorably to the shorter questionnaire length, so separate studies should be performed on PROMIS 29, as such.  In line with PROMIS official policy, it is perfectly acceptable to select 8 items (from PROMIS-57) for some domains and 4 items (from PROMIS 29) for others, and

skipping domains altogether, if they are not relevant to the respondents. This will allow for more flexible and patient friendly modes of measurement, while waiting for full item banks and Computer Assisted Testing (CAT) modules to be available in Norwegian.

### About the article

In the following article all the results are presented in a more condensed format. It tis designed for publication in Quality of Life Research. Most requirements for that publication have been observed. However, more tables and figures have been included than are allowed, for illustrative purposes for the reader. Even more tables and figure are added as supplementary material in the appendix that follows the article.

## References for the kappa thesis


Ameringer, S., Elswick, R. K., Jr., Menzies, V., Robins, J. L., Starkweather, A., Walter, J., . . . Jallo, N. (2016). Psychometric Evaluation of the Patient-Reported Outcomes Measurement Information System Fatigue-Short Form Across Diverse Populations. *Nurs Res, 65*(4), 279-289. doi:10.1097/nnr.0000000000000162

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. doi:10.1016/j.pain.2010.04.025

Ark, L. A. v. D. (2007). Mokken Scale Analysis in R. *J Stat Softw, 20*(11). doi:10.18637/jss.v020.i11

Bartlett, S. J., Orbai, A. M., Duncan, T., DeLeon, E., Ruffing, V., Clegg-Smith, K., & Bingham, C. O., 3rd. (2015). Reliability and Validity of Selected PROMIS Measures in People with Rheumatoid Arthritis. *PLoS One, 10*(9), e0138543. doi:10.1371/journal.pone.0138543

Braun, H., & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: psychometric and statistical considerations. *Large-scale Assessments in Education, 5*(1), 17. doi:10.1186/s40536-017-0050-x

Cella, D. (2015). *PROMIS 1 Wave 1* [survey data]. Retrieved from: https://doi.org/10.7910/DVN/0NGAKG

Cella, D. (2017). *PROMIS Profiles-HUI data*. Retrieved from: https://doi.org/10.7910/DVN/P7UKWR

Cella, D., Choi, S. W., Condon, D. M., Schalet, B., Hays, R. D., Rothrock, N. E., . . . Reeve, B. B. (2019). PROMIS((R)) Adult Health Profiles: Efficient Short-Form Measures of Seven Health Domains. *Value Health, 22*(5), 537-544. doi:10.1016/j.jval.2019.02.004

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., . . . Hays, R. (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. *J Clin Epidemiol, 63*(11), 1179-1194. doi:10.1016/j.jclinepi.2010.04.011

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med Care, 45*(5 Suppl 1), S3-s11. doi:10.1097/01.mlr.0000258615.42478.55

Chalmers, R., P. . (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *J Stat Softw, 48(6)*, 1-29. doi:doi: 10.18637/jss.v048.i06

Chen, W.-H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational Statistics, 22*(3), 265-289. doi:10.3102/10769986022003265

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. *J Stat Softw, 39*(8), 1-30.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q(3): Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Appl Psychol Meas, 41*(3), 178-194. doi:10.1177/0146621616677520

Cook, K., Kallen, M., & Amtmann, D. (2009). Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation - Official Journal of the International Society of Quality of Life Research, 18*(4), 447-460. doi:10.1007/s11136-009-9464-4

Cook, K. F., Bamer, A. M., Roddey, T. S., Kraft, G. H., Kim, J., & Amtmann, D. (2012). A PROMIS fatigue short form for use by individuals who have multiple sclerosis. *Qual Life Res, 21*(6), 1021-1030. doi:10.1007/s11136-011-0011-8

Coste, J., Rouquette, A., Valderas, J. M., Rose, M., & Leplege, A. (2018). The French PROMIS-29. Psychometric validation and population reference values. *Rev Epidemiol Sante Publique, 66*(5), 317-324. doi:10.1016/j.respe.2018.05.563

Cronbach, L. J., & Shavelson, R. J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educ Psychol Meas, 64*(3), 391-418. doi:10.1177/0013164404266386

Depaoli, S., Tiemensma, J., & Felt, J. M. (2018). Assessment of health surveys: fitting a multidimensional graded response model. *Psychology, Health & Medicine, 23*(sup1), 1299-1317. doi:10.1080/13548506.2018.1447136

Dijkers, M. (2007). "What's in a name?" The indiscriminate use of the "Quality of life" label, and the need to bring about clarity in conceptualizations. *Int J Nurs Stud, 44*(1), 153-155. doi:10.1016/j.ijnurstu.2006.07.016

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *105*(3), 399-412. doi:10.1111/bjop.12046

Evon, D. M., Amador, J., Stewart, P., Reeve, B. B., Lok, A. S., Sterling, R. K., . . . Fried, M. W. (2018). Psychometric properties of the PROMIS short form measures in a U.S. cohort of 961 patients with chronic hepatitis C prescribed direct acting antiviral therapy. *Aliment Pharmacol Ther, 47*(7), 1001-1011. doi:10.1111/apt.14531

Fischer, F., Gibbons, C., Coste, J., Valderas, J. M., Rose, M., & Leplege, A. (2018). Measurement invariance and general population reference values of the PROMIS Profile 29 in the UK, France, and Germany. *Qual Life Res, 27*(4), 999-1014. doi:10.1007/s11136-018-1785-8

Fischer, H. F., Wahl, I., Nolte, S., Liegl, G., Brahler, E., Lowe, B., & Rose, M. (2017). Language-related differential item functioning between English and German PROMIS Depression items is negligible. *Int J Methods Psychiatr Res, 26*(4). doi:10.1002/mpr.1530

Forero, C. a. M.-O., A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *14*, 275-299. doi:10.1037/a0015825

Hackney, A. J., Klinedinst, N. J., & Resnick, B. (2019). Measuring Fatigue in Older Adults With Joint Pain: Reliability and Validity Testing of the PROMIS Fatigue Short Forms. *J Nurs Meas, 27*(3), 534-553. doi:10.1891/1061-3749.27.3.534

Hahn, E. A., Kallen, M. A., Jensen, R. E., Potosky, A. L., Moinpour, C. M., Ramirez, M., . . . Teresi, J. A. (2016). Measuring social function in diverse cancer populations: Evaluation of measurement equivalence of the Patient Reported Outcomes Measurement Information System((R)) (PROMIS((R))) Ability to Participate in Social Roles and Activities short form. *Psychol Test Assess Model, 58*(2), 403-421. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6136841/pdf/nihms948248.pdf

Hays, R. D., & Morales, L. S. (2001). The RAND-36 measure of health-related quality of life. *Ann Med, 33*(5), 350-357. doi:10.3109/07853890109002089

Hays, R. D., Sherbourne, C. D., & Mazel, R. M. (1993). The RAND 36-Item Health Survey 1.0. *Health Econ, 2*(3), 217-227. doi:10.1002/hec.4730020305

Hays, R. D. a. R., B. B . . (2008). Measurement and Modeling of Health-Related Quality

of Life. In e. Kris Heggenhougen and Stella Quah (Ed.), *International*

*Encyclopedia of Public Health* (Vol. 4, pp. 241-252 ). San Diego: Academic Press.

Hinchcliff, M., Beaumont, J. L., Thavarajah, K., Varga, J., Chung, A., Podlusky, S., . . . Cella, D. (2011). Validity of two new patient-reported outcome measures in systemic sclerosis: Patient-Reported Outcomes Measurement Information System 29-item Health Profile and Functional Assessment of Chronic Illness Therapy-Dyspnea short form. *Arthritis Care Res (Hoboken), 63*(11), 1620-1628. doi:10.1002/acr.20591

Hirschfeld, G., & Von Brachel, R. (2014). Multiple-Group Confirmatory Factor Analysis in R--A Tutorial in Measurement Invariance with Continuous and Ordinal Indicators. *Practical Assessment, Research & Evaluation, 19*(7).

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. doi:10.1080/10705519909540118

Hung, M., Voss, M. W., Bounsanga, J., Crum, A. B., & Tyser, A. R. (2017). Examination of the PROMIS upper extremity item bank. *J Hand Ther, 30*(4), 485-490. doi:10.1016/j.jht.2016.10.008

Hung, M., Voss, M. W., Bounsanga, J., Gu, Y., Granger, E. K., & Tashjian, R. Z. (2018). Psychometrics of the Patient-Reported Outcomes Measurement Information System Physical Function instrument administered by computerized adaptive testing and the Disabilities of Arm, Shoulder and Hand in the orthopedic elbow patient population. *J Shoulder Elbow Surg, 27*(3), 515-522. doi:10.1016/j.jse.2017.10.015

Jensen, R. E., King-Kallimanis, B. L., Sexton, E., Reeve, B. B., Moinpour, C. M., Potosky, A. L., . . . Teresi, J. A. (2016). Measurement properties of PROMIS Sleep Disturbance short forms in a large, ethnically diverse cancer cohort. *Psychol Test Assess Model, 58*(2), 353-370.

Katz, P., Pedro, S., & Michaud, K. (2017). Performance of the Patient-Reported Outcomes Measurement Information System 29-Item Profile in Rheumatoid Arthritis, Osteoarthritis, Fibromyalgia, and Systemic Lupus Erythematosus. *Arthritis Care Res (Hoboken), 69*(9), 1312-1321. doi:10.1002/acr.23183

Khanna, D., Krishnan, E., Dewitt, E. M., Khanna, P. P., Spiegel, B., & Hays, R. D. (2011). The future of measuring patient-reported outcomes in rheumatology: Patient-Reported Outcomes Measurement Information System (PROMIS). *Arthritis Care Res (Hoboken), 63 Suppl 11*, S486-490. doi:10.1002/acr.20581

Kudel, I., Pona, A., Cox, S., Szoka, N., Tabone, L., & Brode, C. (2019). Psychometric properties of NIH PROMIS(R) instruments in bariatric surgery candidates. *Health Psychol, 38*(5), 359-368. doi:10.1037/hea0000697

Lameijer, C. M., van Bruggen, S. G. J., Haan, E. J. A., Van Deurzen, D. F. P., Van der Elst, K., Stouten, V., . . . Terwee, C. B. (2020). Graded response model fit, measurement invariance and (comparative) precision of the Dutch-Flemish PROMIS® Upper Extremity V2.0 item bank in patients with upper extremity disorders. *BMC musculoskeletal disorders, 21*(1), 170-170. doi:10.1186/s12891-020-3178-8

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936-949. doi:10.3758/s13428-015-0619-7

Luijten, M. A. J., Terwee, C. B., van Oers, H. A., Joosten, M. M. H., van den Berg, J. M., Schonenberg-Meinema, D., . . . Haverman, L. (2019). Psychometric properties of the pediatric Patient-Reported Outcomes Measurement Information System (PROMIS(R)) item banks in a Dutch clinical sample of children with Juvenile Idiopathic Arthritis. *Arthritis Care Res (Hoboken)*. doi:10.1002/acr.24094

Merriwether, E. N., Rakel, B. A., Zimmerman, M. B., Dailey, D. L., Vance, C. G. T., Darghosian, L., . . . Sluka, K. A. (2017). Reliability and Construct Validity of the Patient-Reported Outcomes Measurement Information System (PROMIS) Instruments in Women with Fibromyalgia. *Pain Med, 18*(8), 1485-1495. doi:10.1093/pm/pnw187

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research, 27*(5), 1171-1179. doi:10.1007/s11136-017-1765-4

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value Health, 18*(1), 25-34. doi:10.1016/j.jval.2014.10.005

PHO International, P. H. O. I. (2013). PROMIS®Instrument Development and Validation Scientific Standards Version 2.0. Retrieved from http://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf

PHO, P. H. O. I. (2014). *Minimum requirements for the release of PROMIS instruments after translation and recommendations for further psychometric evaluation* Retrieved from http://www.healthmeasures.net/images/PROMIS/Standards_for_release_of_PROMIS_instruments_after_translation_v8.pdf

R Core Team. (2018). R: A Language and Environment for Statistical Computing (Version R version 3.5.2): R Core Team, Vienna, Austria. Retrieved from https://www.R-project.org/

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational Statistics, 4*(3), 207-230. doi:10.3102/10769986004003207

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Group, P. C. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care, 45*(5 Suppl 1), S22-31. doi:10.1097/01.mlr.0000250483.85507.04

Reise, S. P., Rodriguez, A., Spritzer, K. L., & Hays, R. D. (2018). Alternative Approaches to Addressing Non-Normal Distributions in the Application of IRT Models to Personality Measures. *J Pers Assess, 100*(4), 363-374. doi:10.1080/00223891.2017.1381969

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective. *Educ Psychol Meas, 73*(1), 5-26. doi:10.1177/0013164412449831

Reise, S. P., & Waller, N. G. (2009). Item Response Theory and Clinical Measurement. *Annu. Rev. Clin. Psychol., 5*(1), 27-48. doi:10.1146/annurev.clinpsy.032408.153553

Reise, S. P., & Yu, J. (1990). Parameter Recovery in the Graded Response Model Using MULTILOG. *Journal of Educational Measurement, 27*(2), 133-144. Retrieved from www.jstor.org/stable/1434973

Revicki, D., & Cella, D. (1997). Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation - Official Journal of the International Society of Quality of Life Res, 6*(6), 595-600. doi:10.1023/A:1018420418455

Riley, W. T., Rothrock, N., Bruce, B., Christodolou, C., Cook, K., Hahn, E. A., & Cella, D. (2010). Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: further evaluation of content validity in IRT-derived item banks. *Qual Life Res, 19*(9), 1311-1321. doi:10.1007/s11136-010-9694-5

Rose, A. J., Bayliss, E., Huang, W., Baseman, L., Butcher, E., Garcia, R. E., & Edelen, M. O. (2018). Evaluating the PROMIS-29 v2.0 for use among older adults with multiple chronic conditions. *Qual Life Res, 27*(11), 2935-2944. doi:10.1007/s11136-018-1958-5

Schnohr, C. W., Rasmussen, C. L., Langberg, H., & Bjorner, J. B. (2017). Danish translation of a physical function item bank from the Patient-Reported Outcome Measurement Information System (PROMIS). *Pilot Feasibility Stud, 3*, 29. doi:10.1186/s40814-017-0146-7

Segawa, E., Schalet, B., & Cella, D. (2019). A comparison of computer adaptive tests (CATs) and short forms in terms of accuracy and number of items administrated using PROMIS profile. *Qual Life Res, 29*(1). doi:10.1007/s11136-019-02312-8

Senders, A., Hanes, D., Bourdette, D., Whitham, R., & Shinto, L. (2014). Reducing survey burden: feasibility and validity of PROMIS measures in multiple sclerosis. *Mult Scler, 20*(8), 1102-1111. doi:10.1177/1352458513517279

Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika, 74*(1), 107-120. doi:10.1007/s11336-008-9101-0

Sivertsen, B., Vedaa, O., Harvey, A. G., Glozier, N., Pallesen, S., Aaro, L. E., . . . Hysing, M. (2019). Sleep patterns and insomnia in young adults: A national survey of Norwegian university students. *J Sleep Res, 28*(2), e12790. doi:10.1111/jsr.12790

Smith, A. W., & Jensen, R. E. (2019). Beyond methods to applied research: Realizing the vision of PROMIS(R). *Health Psychol, 38*(5), 347-350. doi:10.1037/hea0000752

Smits, N., Öğreden, O., Garnier-Villarreal, M., Terwee, C. B., & Chalmers, R. P. (2020). A study of alternative approaches to non-normal latent trait distributions in item response theory models used for health outcome measurement. *Statistical Methods in Medical Research*, 0962280220907625. doi:10.1177/0962280220907625

SSB. (2020). Befolkningspyramide. Retrieved from https://www.ssb.no/befolkning/faktaside/befolkningen

Stephan, A., Mainzer, J., Kummel, D., & Impellizzeri, F. M. (2019). Measurement properties of PROMIS short forms for pain and function in orthopedic foot and ankle surgery patients. *Qual Life Res, 28*(10), 2821-2829. doi:10.1007/s11136-019-02221-w

Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. *BMC Med Res Methodol, 12*(1), 74. doi:10.1186/1471-2288-12-74

Stover, A. M., McLeod, L. D., Langer, M. M., Chen, W. H., & Reeve, B. B. (2019). State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. *J Patient Rep Outcomes, 3*(1), 50. doi:10.1186/s41687-019-0130-5

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education, 48*(6), 1273-1296. doi:10.1007/s11165-016-9602-2

Tang, E., Ekundayo, O., Peipert, J. D., Edwards, N., Bansal, A., Richardson, C., . . . Mucsi, I. (2019). Validation of the Patient-Reported Outcomes Measurement Information System (PROMIS)-57 and -29 item short forms among kidney transplant recipients. *Qual Life Res, 28*(3), 815-827. doi:10.1007/s11136-018-2058-2

Teresi, J. A., & Jones, R. N. (2016). Methodological Issues in Examining Measurement Equivalence in Patient Reported Outcomes Measures: Methods Overview to the Two-Part Series, "Measurement Equivalence of the Patient Reported Outcomes Measurement Information System((R)) (PROMIS((R))) Short Forms". *Psychol Test Assess Model, 58*(1), 37-78. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5625814/pdf/nihms855874.pdf

Teresi, J. A., Ocepek-Welikson, K., Cook, K. F., Kleinman, M., Ramirez, M., Reid, M. C., & Siu, A. (2016). Measurement Equivalence of the Patient Reported Outcomes Measurement Information System((R)) (PROMIS((R))) Pain Interference Short Form Items: Application to Ethnically Diverse Cancer and Palliative Care Populations. *Psychol Test Assess Model, 58*(2), 309-352. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5625836/pdf/nihms855897.pdf

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Jones, R. N., . . . Cella, D. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychol Sci Q, 51*(2), 148-180. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2844669/pdf/nihms136951.pdf

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016). Psychometric Properties and Performance of the Patient Reported Outcomes Measurement Information System((R)) (PROMIS((R))) Depression Short Forms in Ethnically Diverse Groups. *Psychol Test Assess Model, 58*(1), 141-181. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5443256/pdf/nihms855884.pdf

van Bruggen, S. G. J., Lameijer, C. M., & Terwee, C. B. (2019). Structural validity and construct validity of the Dutch-Flemish PROMIS((R)) physical function-upper extremity version 2.0 item bank in Dutch patients with upper extremity injuries. *Disabil Rehabil*, 1-9. doi:10.1080/09638288.2019.1651908

Wall, M. M., Park, J. Y., & Moustaki, I. (2015). IRT Modeling in the Presence of Zero-Inflation With Application to Psychiatric Disorder Severity. *Appl Psychol Meas, 39*(8), 583-597. doi:10.1177/0146621615588184

Weldring, T., & Smith, S. M. S. (2013). Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). *Health services insights, 6*, 61-68. doi:10.4137/HSI.S11093

www.nihpromis.org. (2020). Intro to PROMIS. Retrieved from http://www.nihpromis.org

Yang, M., Keller, S., & Lin, J. S. (2019). Psychometric properties of the PROMIS((R)) Fatigue Short Form 7a among adults with myalgic encephalomyelitis/chronic fatigue syndrome. *Qual Life Res, 28*(12), 3375-3384. doi:10.1007/s11136-019-02289-4

Article:

# Psychometric properties of the of PROMIS-57 questionnaire, Norwegian version.


- Author: Stein Arne Rimehaug[1], Aaron James Kaat[2], Jan Egil Nordvik[3], Mari Klokkerud[4], Hilde Stendal Robinson[5]
- Title: Psychometric properties of the of PROMIS-57 questionnaire Norwegian version
- The Affiliations: [1,5]University of Oslo, [1,4,5]Sunnaas Rehabilitation Hospital, [2]Northwestern University, Chicago.
- Contact: stein.arne.rimehaug@sunnaas.no
- Orcid: [1]0000-0003-4811-2688, [2]0000-0001-8147-1899, [3]0000-0002-4828-2073, [4]0000-0003-0981-648X, [5]0000-0002-0275-2965



Abstract

Trial registration: N/A, not a clinical trial



Keywords

PROMIS, patient-reported outcomes, assessment center, quality of life, clinimetric, psychometric, validity

Declarations

**Funding**: No financial funding to disclose

**Ethics approval** N/A, no need since only anonymously collected data without identifying information was used

**Conflicts of interest/Competing interests**: Stein Arne Rimehaug is national contact person for PROMIS in Norway, James Aaron Kaat and Stein Arne Rimehaug are both members of the Scientific Advisory Committee for PROMIS

**Availability of data and material** (N/A)

**Code availability** R code used available on request, standard analysis from packages mirt, lavaan, psych, Mokken, ggplot and semTools.

**Authors' contributions**: SAR: statistical analysis and manuscript, AJK: psychometric methods advice, analysis quality check, HSR: supervision and presentation advice, MK: presentation advice and practical support, JEN: planning and executing data collection.


……

## ABSTRACT

Purpose: The aims of the cross-sectional study were to explore reliability and validity of the Norwegian PROMIS-57 questionnaire in a general population sample, n=408, and to examine Item Response properties and factor structure.

Methods:

Reliability measures were obtained from factor analysis and Item Response Theory (IRT)methods, correlations between PROMIS-57 and RAND36 were examined for concurrent and discriminant validity, factor structure and IRT assumptions were examined with factor analysis methods. IRT Item and model fit and graphic plots were inspected, and Differential Item Functioning (DIF) for language, age, gender and education level were examined.

Results:

PROMIS-57 demonstrates excellent reliability and satisfactory concurrent and discriminant validity. Factor structure of seven domains was confirmed. IRT assumptions are met for unidimensionality, local independence, monotonicity and invariance with no Differential Item Functioning (DIF) of consequence for language or age groups. Estimated Common Variance (ECV) per domain and CFA model fit supports unidimensionality for all seven domains. Acceptable Graded Response model fit and IRT plots.

Conclusions:

The psychometric properties and factor structure of Norwegian PROMIS-57 are satisfactory, and this questionnaire along with PROMIS 29 and the included 8 or 4 item short forms for physical function, anxiety, depression, fatigue, sleep disturbance, social participation ability and pain interference are ready for use in research and clinical care in Norwegian populations. Further studies on longitudinal reliability and sensitivity in patient populations and for Norwegian item calibration and reference scores are needed.

## Plain language summary:

PROMIS-57 is a questionnaire meant for self-reporting different aspects of health and quality of life. There are sections for physical function, anxiety, depression, fatigue, sleep problems, social participation and pain measurement. This study examined the Norwegian version by having a number of people (408) complete this another commonly used questionnaire, RAND-36, and testing the results with a variety of advanced statistical methods to see if PROMIS-57 is able to accurately and reliably measure these different components of a healthy life. The results indicate that this is the case, and that this questionnaire may be used in research and in health care to help measure the results of treatment or the consequences of living with a health condition or disability.

3467 words - word limit 4000

# Introduction:

The Patient Reported Outcomes Measurement Information Systems (PROMIS) initiative has provided new item banks, short form questionnaires, as well as flexible computerized adapted testing catching on in research and clinical health care (Cella et al., 2019). These new questionnaires were developed in collaboration between the US National Institutes of Health, Northwestern University and others, using modern statistical methods, mainly Item Response Theory (IRT) (Cella et al., 2007). PROMIS also encompasses several hundred items across many item banks, each covering a different human latent trait within the domains already mentioned and many more (Cella et al., 2007).

PROMIS Profile 57 (PROMIS-57) is a collection of eight-item short forms meant to capture important domains of human health related quality of life (HRQOL). The following seven domains are included; physical function (PF), anxiety (ANX), depression (DEP), fatigue (FAT), sleep disturbance (SLP), ability to participate in social roles and activities (SOC), pain interference (PAIN), and a pain intensity numeric rating scale (NRS). All domains have been described previously (Riley et al., 2010). PROMIS-57 has previously been translated into Norwegian, and approved according to rigid standards set forth by the PROMIS Health Organization (PHO International, 2013), but the methological quality has not been examined yet.

PROMIS raw scores can be converted to T-scores using look-up tables or online scoring at www.assessmentcenter.org scoring service (Healthmeasures Scoring service). T-score conversion establishes 50 as a general population mean, and any 10-point deviation corresponds with 1SD – one Standard Deviation difference, for easy-to-understand and consistent scoring across measures. Cultural bias from using US reference T-scores in Europe is minimal (Fischer et al., 2017).

PROMIS-29 is a shorter questionnaire nested within PROMIS-57, consisting of four items each from the same seven domains, and thus can be examined using the same data. RAND-36-item Health Survey 1.0 (RAND-36) (R. D. Hays, Sherbourne, & Mazel, 1993)) is a common HRQOL questionnaire, and reliability and validity is well

established across diverse populations {Hays, 2001 #1389}. It is license and cost free, and covers similar life domains as PROMIS-57.

The aim of this study is to explore the reliability and validity of the Norwegian PROMIS-57 according to criteria issued by the PHO organization (PHO, 2014), using RAND-36 as comparative reference. Each short form embedded in PROMIS-57 was hypothesized to have strong internal consistency, a strong concurrent and discriminant validity against RAND-36, satisfactory IRT properties, factor structure confirmed, no Differential Item Functioning (DIF) for language, age, gender, education level or self-reported health.

# Methods

This study was cross-sectional, and collection of responses was conducted in a sample from the general population. Respondents were recruited in 2019 through a newspaper advertisement and posts on Facebook groups and pages encouraging sharing of a link to an online questionnaire. This questionnaire included a consent statement and information about the purpose of the study.

Participants filled in their responses to all items in the Norwegian PROMIS-57 and RAND-36. PROMIS-57 has 5 category Likert response options, (except for the 0-10 pain intensity NRS item). Raw scores for each short form in PROMIS-57 were calculated for the analyses. Higher scores in any PROMIS scale indicate more of the measured construct, causing some correlations between scores to be negative. RAND-36 contains the same items as the original SF-36 (R. D. Hays et al., 1993), but has a different scoring. The 3, 5 or 6 category responses were converted to sum-scores, using the official RAND-36 scoring syntax (R. D. Hays & Morales, 2001), so that higher scores indicate more desirable health on a 0-100 scale. In addition, the following demographic information was collected: gender, age, education level, employment status, income categories, cohabitation, and presence of mental and/or physical health concern.

# Statistical analyses:

The methods chosen for analysis are intended to match criteria in the COSMIN risk-of-bias checklist (Mokkink et al., 2018), the PROMIS Standards for release of PROMIS® instruments after translation v8 (PHO, 2014) and PROMIS® Instrument Development and Validation Scientific Standards Version 2.0 (PHO International, 2013).

## Reliability and internal consistency

Reliability measures based on factor analysis and IRT, calculating marginal empirical reliability and McDonald's omega coefficient from a bi-factor analysis in R package 'psych' v1.8.12, expecting excellent reliability >.9 for all the above for each of the domains, as found in other studies (Merriwether et al., 2017) (Flynn et al., 2015) (Jensen et al., 2015). Measuring overall consistency for PROMIS-57 is not appropriate, since it is multidimensional, and there is no total score calculation for the questionnaire. IRT Test Information Function and scale Standard Error (SE)

plots were visually inspected to evaluate the reliability of measurement across the range of possible responses for each domain scale (R. D. a. R. Hays, B. B . , 2008). In addition, Cronbach's alpha was calculated.

### Validity

Concurrent validity of PROMIS-57 T-scores per domain were tested against their corresponding against RAND-36 sub-scales using Spearman rho correlation coefficients, considering $r_s$ >.8 as very strong correlation, $r_s$ >.7 as strong, and $r_s$ >.6 as moderate correlation strength.

Discriminant validity was assessed through correlations between dissimilar PROMIS domain scores and RAND-36 sub-scales, expecting for instance physical, social and pain scores to have low to moderate correlations ($r_s$ <.6) with mental measures. Factor validity was examined with Confirmatory Factor Analysis, examining PROMIS-57 for seven factors, and each domain for the relative fit of a single factor.

All 7 domains within PROMIS-57 were separately analyzed with 3 different Item Response Theory (IRT) models, Graded Response Model (Graded), Generalized Rating Scale (GRSM) and Rasch model. The assumptions needed for IRT models (unidimensionality, local independence, monotonicity, and invariance) were checked. EFA with Weighted Least Square Mean and Variance adjusted (WLSMV) analysis performed in 'psych' package in R looking for eigenvalue ratio >4:1 as signs of unidimensionality, and bi-factor analysis, also in 'psych' to extract Explained Common Variance (ECV), indicating what proportion of variation is explained by the general factor (should be >.60 (Reise, Scheines, Widaman, & Haviland, 2013)), and Confirmatory Factor Analysis (CFA) was performed to test the factor structure for unidimensionality, first by a single correlated seven factor CFA for all of PROMIS-57, reversing PF and SOC scoring, then by running a single factor CFA separately for each of the seven PROMIS-57 domains. CFA was performed using R package 'lavaan' v6.05 with the Weighted Least Square Mean and Variance adjusted (WLSMV) estimator. Model fit for the factor analysis and for the IRT models was assessed, looking for the lowest Bayesian Information Criteria (BIC), and Root Mean Square Error of Approximation (RMSEA) <0.06, Standardized Root Mean Square Residual (SRMSR) <0.08, Comparative Fit Index (CFI) > 0.95 and Tucker-Lewis Index (TLI) > 0.95 as reference values (Hu & Bentler, 1999), using scaled and unscaled. Model fit for IRT was obtained though M2 analysis (type C2 because of the sample size) performed in R with 'mirt' package) (Chalmers, 2012)

Local dependency (LD) was examined based on the residuals from the CFA with WLSMV estimator in R package 'psych', flagging any item pair with >.2 residual correlation, as in PROMIS item bank development (Reeve et al., 2007), and with the Chen and Thissen LD index (Chen & Thissen, 1997) in R package 'mirt', considering >.3 as possible LD and >1 as definite LD. Monotonicity was tested using Mokken scale (R package 'mokken' (Ark, 2007)), expecting scalability coefficients (Coef_h) >.3. IRT item fit was examined using 'mirt' v1.31 (Chalmers,

2012) package in R, expecting no items with a S-X2 p-value of less than 0.001, which is indicative of poor item fit. The S-X2 statistic indicates whether each item meets expected response frequencies under the estimated IRT model (Kang & Chen, 2011). Also, IRT plots from Graded Response Model created with 'mirt'. Item Response Function (IRF), Item Characteristic Curves (ICC's) and Item Information curves were visually inspected.

Differential item functioning (DIF) can threaten the validity of a score, if some items bias a sub-population over another. DIF analysis was performed using R package 'lordif' v0.3-3 (Choi, Gibbons, & Crane, 2011) with ordinal logistic regression models and McFadden's pseudo R(2)-change of >/= 2% as critical value, as suggested by the PHO (PHO, 2014). The impact of DIF on item scores and total domain score was examined by inspecting item characteristic curves (ICCs) and test characteristic curves (TCCs), as in previous studies on PROMIS translation validation studies {Crins, 2015 #1175}, {Terwee, 2019 #1356} and (Crins et al., 2019).
Language DIF was performed by comparing the scores in this study against two available PROMIS datasets from US studies, the 'PROMIS Profiles HUI data' (Cella, 2017) and the 'PROMIS 1 WAVE1' (Cella, 2015), including only the respondents that had been subjected to all items within any given short form. Age DIF in the Norwegian sample was studied by grouping respondents as younger (n=206) and older (n=202) around the median age (52). Gender DIF was examined with 310 female and 98 male respondents. Education level DIF was analyzed for the n=299 with college/university level education against the n=109 with high school or lower. Health DIF groups consisted of respondents reporting having "no health problems" (n= 130) vs. physical, mental health problems or both (n =278).

## Results

| **Table 1. Sample characteristics: age, gender, cohabitation, income, education and health status (N=408).** | | |
|---|---|---|
| **Age –** mean (SD * / min-max) | | 52 (13 / 19-88) |
| | | **n(%):** |
| **Women** | | 310 (76) |
| **Living alone** | | 102 (25) |
| **Income source** | Employed, part or full time: | 215 (53) |
| | Retired | 57 (14) |
| | Permanent disability | 79 (19) |
| | Sick leave, short or long term** | 42 (10) |
| | Other ** | 15 (4) |
| **Income level*** ** | Low (<350k NOK) | 124 (31) |
| | Middle (350k-600k NOK) | 183 (45) |
| | High (>600k NOK) | 96 (24) |
| **Education** | College level or higher | 298 (73) |
| | Intermediate | 89 (22) |
| | Elementary only (>10yr) | 21 (5) |
| **Health problems, self reported** | Physical health problems | 166 (41) |
| | Mental health problems | 18 (4) |
| | Both physical and mental | 94 (23) |
| | No health problems | 130 (32) |

*SD=Standard deviation, **= away from work >12 month duration, «arbeidsavklaringspenger», ***=homemaker, student, no response or marked as «other»

408 complete and anonymous responses were collected and all were included in the analysis. Characteristics of respondents are presented in Table 1. The sample self-reports health problems at higher rate than the general population - 32% in the sample report no health problems vs 73% in the HUNT study (Holseter, Dalen, Krokstad, & Eikemo, 2015), and is higher educated -73% college level vs 33% in general population (SSB, 2017). The 4.7 year age difference between genders is significant, whereas gender associations with living alone, income level or taking prescription medications are not. Responses to PROMIS-57 were complete for every item, and all response categories were endorsed in each domain, but category "5" has <10 respondents in five of the DEP and three ANX items. (Histograms of domain scores are in supplementary material 6).

## Reliability:

For PROMIS-57 as a whole, alpha is >.99 for PROMIS-57 and .97 for PROMIS29, with negative scores reversed. The 8-item short forms within PROMIS-57 all have high reliability indices in this Norwegian sample, with McDonald's omega total between .91 and .99, and IRT marginal reliability scores between .87 and .94, and Cronbach's alpha values between .91 and .98, see Table 2 for details.

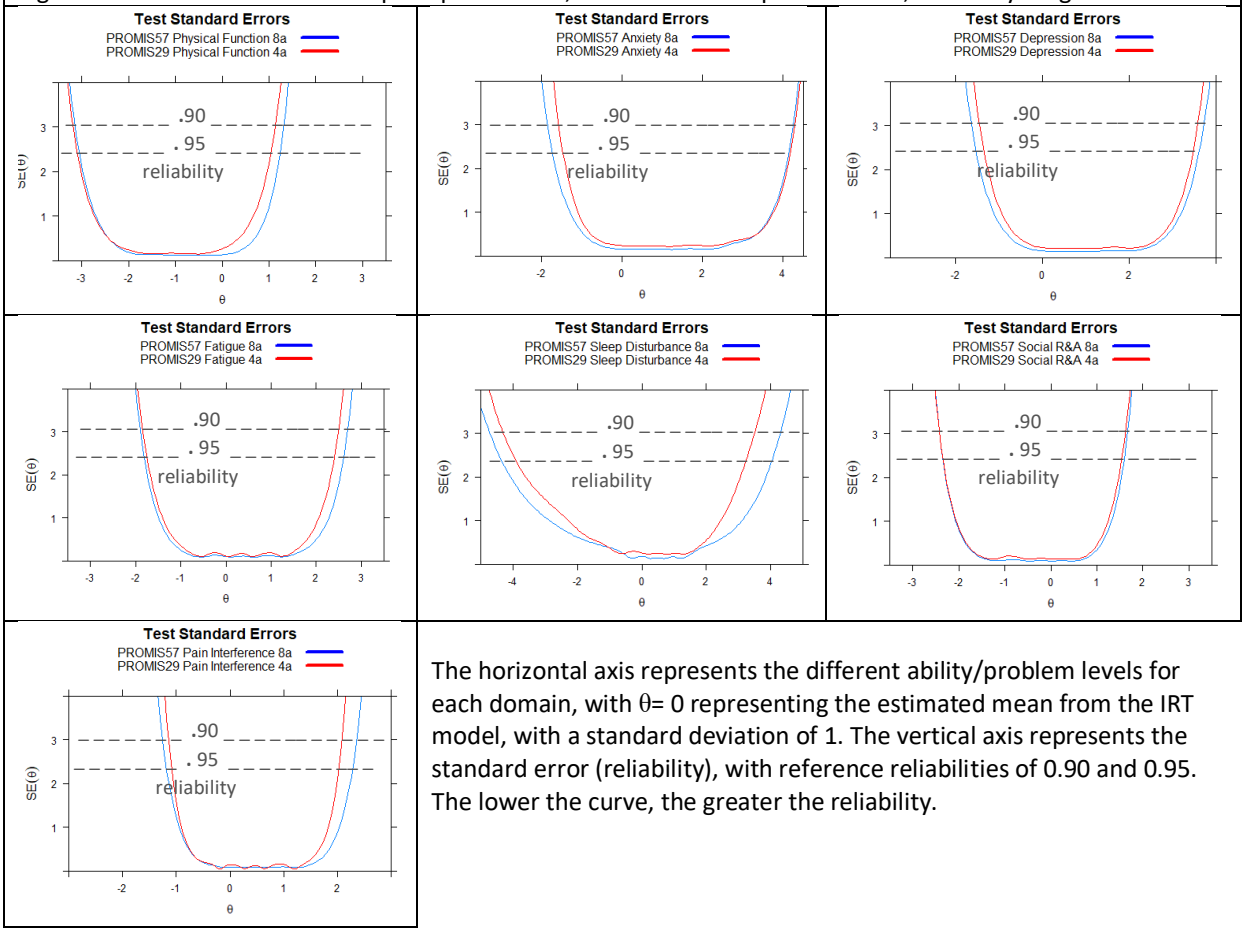
Table 2: PROMIS-57 per domain score, reliability and validity variables

| | PF | Anxiety | Depression | Fatigue | Sleep Dstrb | Social | Pain Intf | Pain Intensity NRS |
|---|---|---|---|---|---|---|---|---|
| Mean T-score (SD*) | 47.6 (10.6) | 50.8 (11.0) | 51.3 (13.0) | 52.3 (11.0) | 52.6 (9.9) | 48.3 (12.2) | 55.0 (11.8) | 3.5** (2.8) |
| ceiling% | 36.4 | 0.0 | 1.0 | 2.4 | 1.5 | 21.5 | 6.8 | 1.2 |
| floor% | 0.5 | 24.2 | 29.6 | 17.6 | 2.7 | 3.9 | 29.1 | 19.6 |
| Cronbach's alpha | .97 | .96 | .97 | .98 | .92 | .98 | .98 | |
| McDonalds omega ω_t (ω_h)*** | .97 (.96) | .96 (.95) | .97 (.96) | .98 (.98) | .92 (.91) | .99 (.99) | .99 (.99) | |
| IRT Marginal reliability | .87 | .91 | .89 | .94 | .92 | .93 | .90 | |
| IRT discrimination mean (per item min-max)**** | 5,9 (4.3 - 7.8) | 4,7 (3.8 - 6.1) | 4,8 (3.7 - 5.6) | 7,4 (4.8 - 10.8) | 4,0 (1.3- 10.2) | 7,5 (5.2 - 10.0) | 8,5 (5.6 - 10.6) | |
| Eigenvalue ratio from 1-factor EFA | 12:1 | 10:1 | 16:1 | 33:1 | 5 :1 | 54:1 | 26:1 | |
| Explained Common Variance (ECV) | .88 | .86 | .88 | .96 | .74 | .96 | .94 | |

*=Standard Deviation ** Pain NRS mean, not T-score *** ω_t=omega total, ω_h = hierarchical
All confidence intervals (95%) for alpha and omega are <+/-.01, except Sleep: +/-.02
****IRT Discrimination parameter from Graded Response model

Plots for the IRT standard error of measurement ranges in Fig. 1 are satisfactory, except for Sleep disturbance 8, where reliability is reduced at both ends of the theta range. SE plots from mirt::fscores are included for PROMIS-57 and 29, showing a small difference in reliability across the range. They are both reliable within a range of the theta (the "ability" or "problem" range) that is relevant to health measurement, from about one standard deviation (SD) better than the population average to at least two SD worse (below 0 for negatively scored PROMIS domains; anxiety, fatigue, pain).

Fig. 1: PROMIS 57 Standard error plots per domain, from Graded Response Model, reliability range




The horizontal axis represents the different ability/problem levels for each domain, with θ= 0 representing the estimated mean from the IRT model, with a standard deviation of 1. The vertical axis represents the standard error (reliability), with reference reliabilities of 0.90 and 0.95. The lower the curve, the greater the reliability.

## Validity

Strong correlation was found between PROMIS Physical function and RAND-36 PF ( .88), PROMIS social and RAND-36 SF (.89), and between PROMIS Fatigue and RAND-36 VT (-.86), PROMIS Depression and RAND-36 MH (-.81), PROMIS Anxiety and RAND-36 MH (-.73), PROMIS Pain Interference and RAND-36 BP (-.93), and between PROMIS Pain intensity NRS and RAND-36 BP (-.92). Details in Table 3.

PROMIS-57 discriminates well between physical and mental scores, as PROMIS anxiety and depression scores correlate only moderately ($r_s$ <.5) with RAND-36 PF and RP, as well as between PROMIS Physical Function and RAND-36 RE and MH, and between PROMIS pain interference and RAND-36 RE and MH. The remaining correlations among PROMIS and RAND-36 dimensions are moderate to strong ($r_s$.5 - $r_s$.8).

Weaker correlations were found, as expected, within PROMIS-57; $r_s$ <.5 between PF/PAIN and ANX/DEP. Moderate correlation ($r_s$ >.6) between SOC and ANX/DEP, between FAT and ANX, and between SLP and all other PROMIS dimensions. As expected, PF, FAT, SOC and PAIN are more related, with correlations well above $r_s$ .7. Details in Table3.

Table 3: Spearman rho correlations $r_s$ within PROMIS-57 domains and against RAND-36 sumscores

| PROMIS: | Physical function | Anxiety | Depression | Fatigue | Sleep | Social* | Pain** | Pain NRS*** |
|---|---|---|---|---|---|---|---|---|
| Physical function | 1.000 | -.409 | -.501 | -.750 | -.541 | .822 | -.815 | -.741 |
| Anxiety | -.409 | 1.000 | .759 | .591 | .547 | -.532 | .438 | .449 |
| Depression | -.501 | .759 | 1.000 | .642 | .546 | -.585 | .509 | .462 |
| Fatigue | -.750 | .591 | .642 | 1.000 | .608 | -.857 | .728 | .688 |
| Sleep | -.541 | .547 | .546 | .608 | 1.000 | -.593 | .547 | .533 |
| Social* | .822 | -.532 | -.585 | -.857 | -.593 | 1.000 | -.774 | -.691 |
| Pain** | -.815 | .438 | .509 | .728 | .547 | -.774 | 1.000 | .918 |
| Pain NRS | -.741 | .449 | .462 | .688 | .533 | -.691 | .918 | 1.000 |
| PROMIS: / RAND 36: | Physical function | Anxiety | Depression | Fatigue | Sleep | Social* | Pain** | Pain NRS*** |
| RAND36 PF PHYSICAL | .880 | - .329 | - .422 | - .675 | - .513 | .751 | - .781 | - .731 |
| RAND36 RP ROLEPHY | .786 | - .420 | - .479 | - .738 | - .509 | .794 | - .737 | - .688 |
| RAND36 BP BODILYPAIN | .793 | - .414 | - .468 | - .713 | - .526 | .741 | - .927 | - .918 |
| RAND36 GH GENERAL | .776 | - .524 | - .558 | - .776 | - .620 | .785 | - .718 | - .681 |
| RAND36 VT VITALIT | .715 | - .560 | - .632 | - .864 | - .617 | .827 | - .670 | - .622 |
| RAND36 SF SOCIAL | .785 | - .517 | - .587 | - .827 | - .597 | .885 | - .743 | - .683 |
| RAND36 RE ROLEMOT | .389 | - .545 | - .584 | - .524 | - .441 | .488 | - .417 | - .432 |
| RAND36 MH MENTAL | .467 | - .727 | - .806 | - .644 | - .560 | .574 | - .480 | - .451 |

*= Social roles and activities ability **= Pain interference ***Pain intensity numeric rating scale

## Unidimensionality (factor validity)

The correlated seven-factor CFA solution using WLSMV estimator for the entire PROMIS-57produced a satisfactory model fit, confirming the original factor structure of seven domains within PROMIS-57. Unscaled fit indices are CFI = .99, TLI = .99, RMSEA = .05, and unscaled SRMR = .04. The average absolute residual correlation is 0.002, and no residual correlations are>.2. From a single-factor CFA using WLSMV estimator performed **separately** for each domain, most scaled and unscaled fit indices are well within the acceptable thresholds for each domain, (Table 4) except for RMSEA, but that is not uncommon for PROMIS and similar questionnaires (Cook, Kallen, & Amtmann, 2009).

Table 4: Single-factor CFA fit, all PROMIS-57domains tested separately with WLSMV estimator

| | rmsea.scaled | srmr | cfi.scaled | tli.scaled |
|---|---|---|---|---|
| PF | 0.129 | **0.022** | **0.998** | **0.997** |
| ANX | 0.080 | **0.019** | **0.998** | **0.998** |
| DEP | 0.124 | **0.023** | **0.996** | **0.994** |
| FAT | 0.115 | **0.010** | **0.999** | **0.999** |
| SLP | 0.223 | **0.074** | **0.986** | **0.980** |
| SOC | 0.124 | **0.011** | **0.999** | **0.999** |

| | | | | | |
|---|---|---|---|---|---|
| PAIN | 0.156 | **0.015** | **0.999** | **0.999** | |
| Cutoffs | <.06 | <.08 | >.95 | >.95 | **Bold** =meets cutoff |

## IRT analysis

Assumptions for IRT were satisfied for all seven short forms. Unidimensionality is supported by Explained Common Variance (ECV) from bifactor models between .74 and .96, and the first to second factor eigenvalue is greater than 4:1 for all domains (The factor structure with seven domains is supported by the EFA, except for slight violation in Sleep disturbance, as evidenced by the weak and double factor loadings already mentioned).
Each domain is considered locally independent, since no item pair residuals from the CFA are >.2 in any domain, and the Chen and Thissen LD index for each domain flags no pairs >1, and only four pairs >.3; two FAT, one SLP, one SOC. (Details in the appendix.) Monotonicity is supported, as Mokken scalability coefficient for each domain scale is between .62 (SLP) and .93 (PAIN), well above the 0.3 cutoff, and no single item lower than .49 (Item Sleep116). Item wording for each item are available in supplementary material, appendix 3.

Model fit indices for three IRT-models were examined in order to select the best IRT model for analysis. The Bayesian Information Criteria (BIC) favors Generalized Rating Scale Model (GRSM) over Graded Response Model (Graded) for all 7 short forms when tested separately, while RMSEA, SRMSR, TLI, CFI provides a mixed result where some of the fit criteria favors Graded over GRSM, but never favors the Rasch model. (Table 5).

Table 5:

PROMIS-57 five model fit indices*, comparing three IRT models Rasch / Graded Response / Generalized Rating Scale, n=408

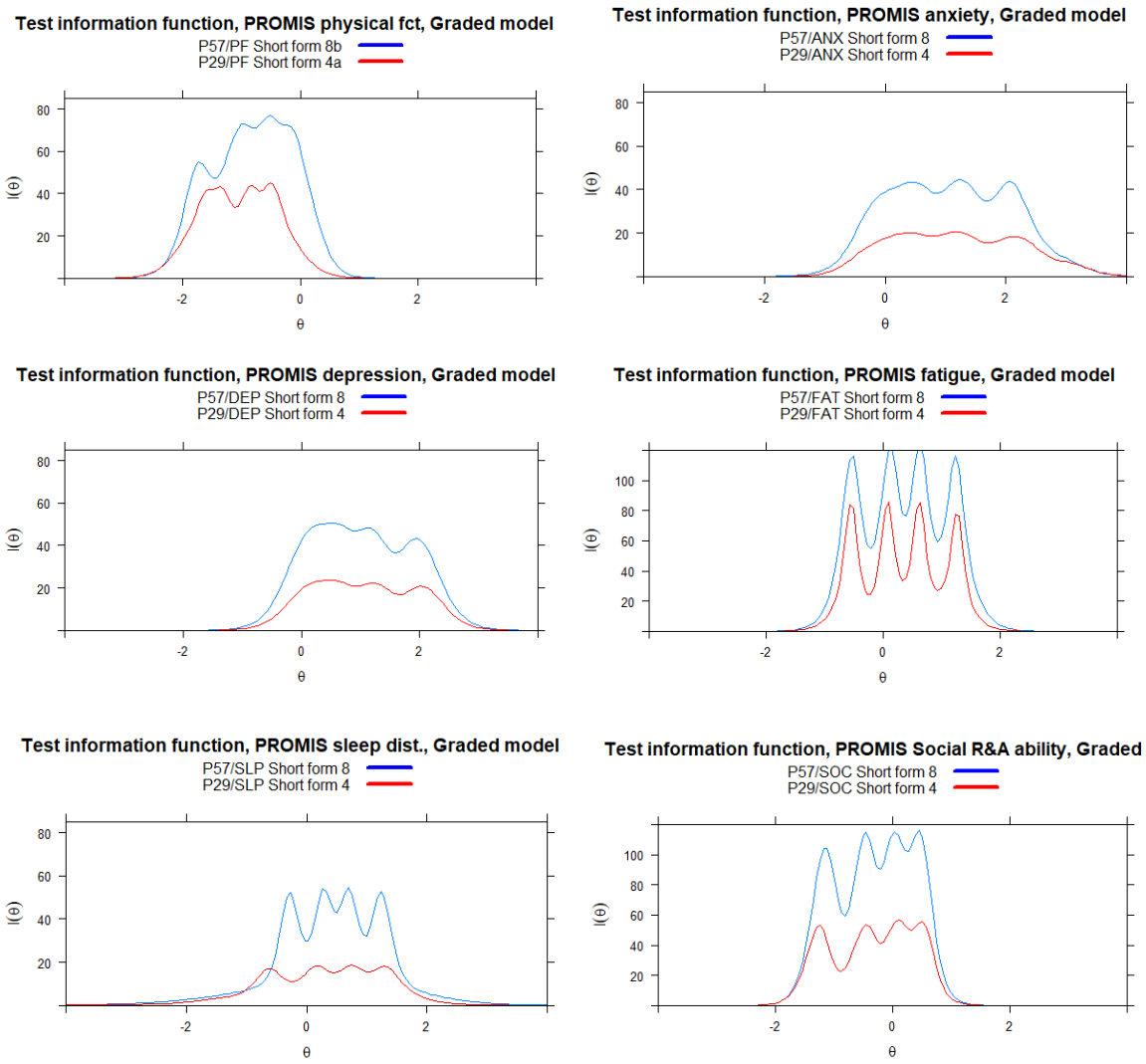
| Thresholds: | Physical Fct | Anxiety | Depression | Fatigue | Sleep | Social | Pain |
|---|---|---|---|---|---|---|---|
| BIC (lowest=best) | 5200/5108/**5068** | 5352/5258/**5202** | 5536/5447/**5350** | 5838/**5500**/5501 | 8057/**7731**/7781 | 5674/5367/**5299** | 4863/5220/**4824** |
| RMSEA <.06 | .107 / .115/ .116 | .091/.082/.076 | .095/.098/.086 | .138/.103/.106 | .209/.227/168 | .136/.116/.095 | .145/.186/.138 |
| SRMSR <.08 | .086 /**.027 / .040** | .092/**.025/.034** | **.075/.029/.030** | **.012/.013/.025** | .123/.081/.103 | .120/**.013/.020** | .119**/018/.027** |
| TLI >.95 | **.098 /.098 /.097** | **.099/.099/.099** | **.098/.099/.099** | **.098/.099/.099** | .877/.856/.921 | **.978/.983/.989** | **.974/.958/.977** |
| CFI >.95 | **.098 /.098 /.097** | **.099/.099/.099** | **.099/.099/.098** | **.098/.099/.098** | .882/.897/.884 | **.978/.988/.984** | **.975/.970/.966** |
| # of criteria met: | 2 / 3 / 4 | 2 / 3 / 4 | 3 / 3 / 4 | 3 / 4 / 3 | 0 / 1 / 0 | 2 / 3 / 4 | 2 / 3 / 5 |

*Bayesian Information Criteria (BIC), Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMSR), Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI)

PROMIS-57 has good IRT Item fit with Graded Response Model, except for two Sleep disturbance items with s-x2 p-values <.001, with or without FDR False Discovery Rate correction (Benjamini & Yekutieli, 2001): the misfitting items are Sleep44 and Sleep72.

Item response curves generated in the mirt package in R to visualize reliability displays well distributed curves, generally without response category curves completely overlapped by others, except item Sleep 116 and PF53 (Physical function). However, steep slopes for some items indicate high discrimination parameters, also evident as spiked Test Information curves. All IRT parameters and plots for the 8 anxiety items are available in supplementary material (Appendix 4)

Fig.4 PROMIS-57 vs PROMIS-29 IRT test information function (TIF) plots comparing




The horizontal axis represents the different ability/problem levels for each domain, with θ= 0 representing the estimated mean from the IRT model, with a standard deviation of 1. The vertical axis represents the combined amount of information from all items that particular scale.

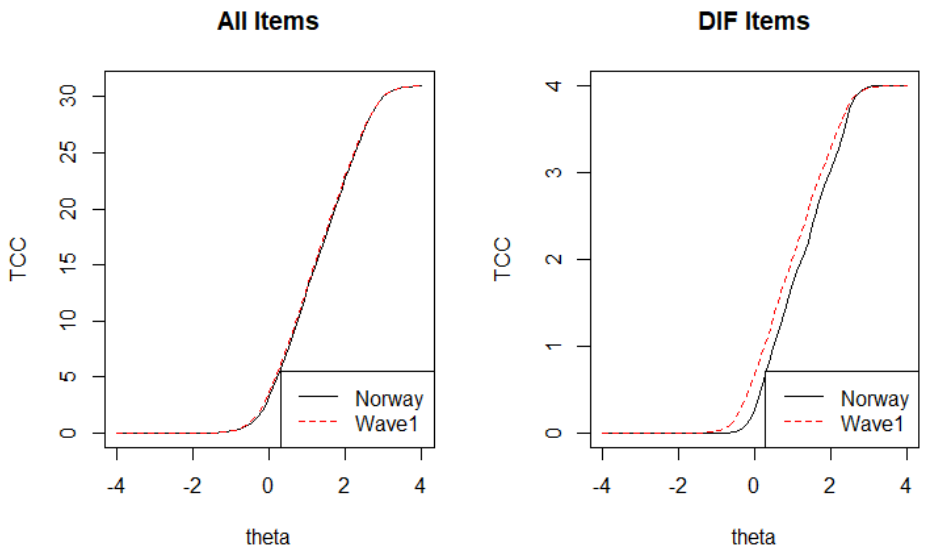
Comparing the Test Information Function (TIF) of PROMIS-57 and PROMIS-29, the information precision is lower in the shorter versions, PROMIS 29 or the included 4 item short forms. Some of the test information function (TIF) curves are unusually spiked, especially with the Graded response Model, related to their also high discrimination parameters.

# Differential Item Functioning

When applying suggested thresholds, no language, age, gender, education DIF of consequence was found. Along the way to this conclusion, however, there are some findings worth exploring.

Three items in PROMIS-57 could not be tested for language DIF, item PFC12 as it is not included in either of the US reference data sets, and EDANX07 and Sleep72 as they were collected from other US respondents than the remaining items. Only respondents that had been presented with the same items in the same domain were selected from the US data sets, n=1214 in Wave1 and n=3409 in Profiles-HUI. DIF analysis with the over-sensitive chi-Square criterion, alpha threshold=0.01, typically flagged one or more items per domain initially. Using the PHO approved R2 criterion method (settings in lordif: pseudo.R2="McFadden", criterion="R2", R2.change = 0.02, model="GRM") , and using as anchors 2-3 DIF free items, as identified by the chi-Sq method (Kopf, Zeileis, & Strobl, 2015), there is language DIF against the US datasets in only one item PAININ09 in all PROMIS-57 short forms. Running DIF analysis without anchors, language DIF was flagged for one item (but not not flagged without anchors), EDANX05 against Wave1 dataset. The same items were not flagged as DIF against the other US dataset (Profiles-HUI). (fig. 5).

Fig 5:




Test characteristic curve (TCC) for language DIF in PROMIS Anxiety. Left graph shows the TCC total consequence of DIF on the scoring of all 8 Norwegian (Norway) and United States (Wave1) PROMIS Anxiety items; the right graph shows the TCC for just EDANX05 items with negligible DIF.

### Gender, age and education DIF

There are some differences between gender on average score of each measure sub-scale, but no gender DIF detected in any of the seven PROMIS short forms in PROMIS-57.

Age: Three PROMIS-57 short forms (Fatigue, Anxiety, and Pain Interference) are free of DIF between older and younger respondents with either method. Physical Function: uniform DIF for one item only with the chi-square (chiSq) method, but none with the pseudo.R2 method. Depression: uniform DIF for two items only with the chiSq method, but not with the R2 method. Two short forms, Sleep disturbance and Social roles show non-uniform age DIF in one item only with the ChiSq method, but not with the R2 method. Education: No items in any short form were flagged for education DIF, comparing with/without college level. Health status DIF: unable to run for PF and ANX as some of the response categories were picked by too few respondents in the healthier group. No health DIF found in the remaining short forms (DEP, FAT, SLP, SOC, or PAIN).

## Discussion:

This the first study to assess the psychometric properties of PROMIS profile and short forms, Norwegian version. PROMIS-57 and 29 and the embedded short forms displayed sufficient validity and reliability for use as a generic clinical measure of HRQOL. The high reliability scores, the omega measures and empirical reliability >.9, and Cronbach's alphas, >.9 support the excellent internal consistency and reliability for PROMIS-57, as in other PROMIS studies (Ron D. Hays, Spritzer, Schalet, & Cella, 2018). Visual inspection of the IRT SE plots provides

further evidence of excellent reliability in the most relevant range for most patient populations, from about population mean to 2SD's worse. PROMIS-29 and its 4-item short forms has similar reliability to PROMIS-57, but with somewhat lower precision beyond 1.5 SD's worse than the mean. Correlations against RAND-36 support the concurrent and discriminatory validity of PROMIS-57. T-scores were used for this to demonstrate the validity of the currently recommended scoring method. Previous studies have also found correlations across PROMIS and RAND-36/SF36 between .66 and .91 for similar constructs (Bingham, 2019) (Hinchcliff et al., 2011) (Crins et al., 2018; Schalet et al., 2015) and between .30 and .61 for dissimilar ones (Rose et al., 2018) (Khanna et al., 2012).

The Norwegian translation has retained the original seven factor structure, and has not introduced significant language DIF bias or age DIF, and probably no gender or education DIF, though any group sample size <200 may have been insufficient to complete rule out type II error. The model fit indices are approaching established criteria of RMSEA<0.06, SRMSR<0.08, CFI>.95 and TLI >0.95 (Hu & Bentler, 1999), (details in Table 5). M2 fit analysis on PROMIS-57 as a whole, more clearly favors Graded response model. Absolute adherence to cutoffs are not needed when assessing model fit indices (Lai, 2016). Graded Response model (Samejima) has been recommended for PROMIS measures (PHO, 2014), and has better fit than the Rasch for the IRT and DIF analyses.

Some items have very high discrimination slopes (especially FAT, SOC and PAIN) and item misfit (only in SLP). Possible explanations are local independence violations, skewed or zero-inflated scores, and sample size. One of the methods shows LD, but not necessarily the domains and items with inflated discrimination. The sample may have too many "non-cases" and zero-inflation can inflate slopes (Reise, Rodriguez, Spritzer, & Hays, 2018), referring in part to (Wall, Park, & Moustaki, 2015). A recent simulation study (Smits, Öğreden, Garnier-Villarreal, Terwee, & Chalmers, 2020) suggests 1.5 to 2 points increased bias of discrimination with zero-inflation. IRT discrimination, LD and item fit needs to be examined in larger and more diverse samples, or else ignored as it is in 1PL and Rasch models. Two items showed minimal language DIF, however the amount of DIF found in these two items is small and of no consequence to the total score, judged by the visual representations. The sample is somewhat gender skewed, but a majority of women is also common in many patient populations. A strength of this study has been applying more advanced analysis methods, exposing the questionnaire to a closer scrutiny. Assessing seven PROMIS short forms at once has its advantages, as it allows for better comparison between domains, while validation of entire item banks would allow testing the PROMIS system for full theta range reliability, floor/ceiling effect, and full calibration of the scale in the new language.

Norwegian PROMIS-57 and PROMIS-29 and embedded short forms are sufficiently reliable and valid to be used in clinical care and research. Future studies should longitudinal reliability and responsivity in patient populations, as well as IRT calibration in a larger Norwegian sample.

References for the article


Ark, L. A. v. D. (2007). Mokken Scale Analysis in R. *J Stat Softw, 20*(11). doi:10.18637/jss.v020.i11

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist., 29*(4), 1165-1188. doi:10.1214/aos/1013699998

Bingham, C. e. a. (2019). P10. A real-world evidence-based assessment and intra-method correlative analysis of PROMIS-29, in PHO 2019 Conference Abstracts. *J Patient Rep Outcomes, 3*(1), 68. doi:10.1186/s41687-019-0157-7

Cella, D. (2015). *PROMIS 1 Wave 1* [survey data]. Retrieved from: https://doi.org/10.7910/DVN/0NGAKG

Cella, D. (2017). *PROMIS Profiles-HUI data*. Retrieved from: https://doi.org/10.7910/DVN/P7UKWR

Cella, D., Choi, S. W., Condon, D. M., Schalet, B., Hays, R. D., Rothrock, N. E., . . . Reeve, B. B. (2019). PROMIS((R)) Adult Health Profiles: Efficient Short-Form Measures of Seven Health Domains. *Value Health, 22*(5), 537-544. doi:10.1016/j.jval.2019.02.004

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med Care, 45*(5 Suppl 1), S3-s11. doi:10.1097/01.mlr.0000258615.42478.55

Chalmers, R., P. . (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *J Stat Softw, 48(6)*, 1-29. doi:doi: 10.18637/jss.v048.i06

Chen, W.-H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational Statistics, 22*(3), 265-289. doi:10.3102/10769986022003265

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. *J Stat Softw, 39*(8), 1-30.

Cook, K., Kallen, M., & Amtmann, D. (2009). Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation - Official Journal of the International Society of Quality of Life Research, 18*(4), 447-460. doi:10.1007/s11136-009-9464-4

Crins, M. H. P., Terwee, C. B., Ogreden, O., Schuller, W., Dekker, P., Flens, G., . . . Roorda, L. D. (2019). Differential item functioning of the PROMIS physical function, pain interference, and pain behavior item banks across patients with different musculoskeletal disorders and persons from general population. *Qual Life Res, 28*(5), 1231-1243. doi:10.1007/s11136-018-2087-x

Crins, M. H. P., van der Wees, P. J., Klausch, T., van Dulmen, S. A., Roorda, L. D., & Terwee, C. B. (2018). Psychometric properties of the PROMIS Physical Function item bank in patients receiving physical therapy. *PLoS One, 13*(2), e0192187. doi:10.1371/journal.pone.0192187

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *105*(3), 399-412. doi:10.1111/bjop.12046

Fischer, H. F., Wahl, I., Nolte, S., Liegl, G., Brahler, E., Lowe, B., & Rose, M. (2017). Language-related differential item functioning between English and German PROMIS Depression items is negligible. *Int J Methods Psychiatr Res, 26*(4). doi:10.1002/mpr.1530

Flynn, K. E., Dew, M. A., Lin, L., Fawzy, M., Graham, F. L., Hahn, E. A., . . . Weinfurt, K. P. (2015). Reliability and construct validity of PROMIS(R) measures for patients with heart failure who undergo heart transplant. *Qual Life Res, 24*(11), 2591-2599. doi:10.1007/s11136-015-1010-y

Hays, R. D., & Morales, L. S. (2001). The RAND-36 measure of health-related quality of life. *Ann Med, 33*(5), 350-357. doi:10.3109/07853890109002089

Hays, R. D., Sherbourne, C. D., & Mazel, R. M. (1993). The RAND 36-Item Health Survey 1.0. *Health Econ, 2*(3), 217-227. doi:10.1002/hec.4730020305

Hays, R. D., Spritzer, K. L., Schalet, B. D., & Cella, D. (2018). PROMIS(®)-29 v2.0 profile physical and mental health summary scores. *Qual Life Res, 27*(7), 1885-1891. doi:10.1007/s11136-018-1842-3

Hays, R. D. a. R., B. B . . (2008). Measurement and Modeling of Health-Related Quality

of Life. In e. Kris Heggenhougen and Stella Quah (Ed.), *International*

*Encyclopedia of Public Health* (Vol. 4, pp. 241-252 ). San Diego: Academic Press.

Healthmeasures Scoring service. NIH-PROMIS scoring service. Retrieved from https://www.assessmentcenter.net/ac_scoringservice

Hinchcliff, M., Beaumont, J. L., Thavarajah, K., Varga, J., Chung, A., Podlusky, S., . . . Cella, D. (2011). Validity of two new patient-reported outcome measures in systemic sclerosis: Patient-Reported Outcomes Measurement Information System 29-item Health Profile and Functional Assessment of Chronic Illness Therapy-Dyspnea short form. *Arthritis Care Res (Hoboken), 63*(11), 1620-1628. doi:10.1002/acr.20591

Holseter, C., Dalen, J. D., Krokstad, S., & Eikemo, T. A. (2015). Self-rated health and mortality in different occupational classes and income groups in Nord-Tröndelag County, Norway. *Selvrapportert helse og dödelighet i ulike yrkesklasser og inntektsgrupper i Nord-Tröndelag, 135*(5).

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. doi:10.1080/10705519909540118

Jensen, R. E., Potosky, A. L., Reeve, B. B., Hahn, E., Cella, D., Fries, J., . . . Moinpour, C. M. (2015). Validation of the PROMIS physical function measures in a diverse US population-based cohort of cancer patients. *Qual Life Res, 24*(10), 2333-2344. doi:10.1007/s11136-015-0992-9

Kang, T., & Chen, T. T. (2011). Performance of the generalized S-X2 item fit index for the graded response model. *Asia Pacific Education Review, 12*(1), 89-96. doi:10.1007/s12564-010-9082-4

Khanna, D., Maranian, P., Rothrock, N., Cella, D., Gershon, R., Khanna, P. P., . . . Hays, R. D. (2012). Feasibility and construct validity of PROMIS and "legacy" instruments in an academic scleroderma clinic. *Value Health, 15*(1), 128-134. doi:10.1016/j.jval.2011.08.006

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches. *Educ Psychol Meas, 75*(1), 22-56. doi:10.1177/0013164414529792

Lai, K. G., Samuel B. . (2016). The Problem with Having Two Watches: Assessment of Fit When RMSEA and CFI Disagree. *Multivariate Behavioral Research, 51:2-3*, 220-239. doi:10.1080/00273171.2015.1134306

Merriwether, E. N., Rakel, B. A., Zimmerman, M. B., Dailey, D. L., Vance, C. G. T., Darghosian, L., . . . Sluka, K. A. (2017). Reliability and Construct Validity of the Patient-Reported Outcomes Measurement Information System (PROMIS) Instruments in Women with Fibromyalgia. *Pain Med, 18*(8), 1485-1495. doi:10.1093/pm/pnw187

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research, 27*(5), 1171-1179. doi:10.1007/s11136-017-1765-4

PHO International, P. H. O. I. (2013). PROMIS®Instrument Development and Validation Scientific Standards Version 2.0. Retrieved from http://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf

PHO, P. H. O. I. (2014). *Minimum requirements for the release of PROMIS instruments after translation and recommendations for further psychometric evaluation* Retrieved from http://www.healthmeasures.net/images/PROMIS/Standards_for_release_of_PROMIS_instruments_after_translation_v8.pdf

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

Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care, 45*(5 Suppl 1), S22-31. doi:10.1097/01.mlr.0000250483.85507.04

Reise, S. P., Rodriguez, A., Spritzer, K. L., & Hays, R. D. (2018). Alternative Approaches to Addressing Non-Normal Distributions in the Application of IRT Models to Personality Measures. *J Pers Assess, 100*(4), 363-374. doi:10.1080/00223891.2017.1381969

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective. *Educ Psychol Meas, 73*(1), 5-26. doi:10.1177/0013164412449831

Riley, W. T., Rothrock, N., Bruce, B., Christodolou, C., Cook, K., Hahn, E. A., & Cella, D. (2010). Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: further evaluation of content validity in IRT-derived item banks. *Qual Life Res, 19*(9), 1311-1321. doi:10.1007/s11136-010-9694-5

Rose, A. J., Bayliss, E., Huang, W., Baseman, L., Butcher, E., Garcia, R. E., & Edelen, M. O. (2018). Evaluating the PROMIS-29 v2.0 for use among older adults with multiple chronic conditions. *Qual Life Res, 27*(11), 2935-2944. doi:10.1007/s11136-018-1958-5

Schalet, B., Revicki, D., Cook, K., Krishnan, E., Fries, J., & Cella, D. (2015). Establishing a Common Metric for Physical Function: Linking the HAQ-DI and SF-36 PF Subscale to PROMIS ® Physical Function. *J Gen Intern Med, 30*(10), 1517-1523. doi:10.1007/s11606-015-3360-0

Smits, N., Öğreden, O., Garnier-Villarreal, M., Terwee, C. B., & Chalmers, R. P. (2020). A study of alternative approaches to non-normal latent trait distributions in item response theory models used for health outcome measurement. *Statistical Methods in Medical Research*, 0962280220907625. doi:10.1177/0962280220907625

SSB. (2017). Utdanningsnivå i befolkningen. Retrieved from https://www.ssb.no/utdanning/artikler-og-publikasjoner/her-er-okningen-i-hoyere-utdanning-storst

Wall, M. M., Park, J. Y., & Moustaki, I. (2015). IRT Modeling in the Presence of Zero-Inflation With Application to Psychiatric Disorder Severity. *Appl Psychol Meas, 39*(8), 583-597. doi:10.1177/0146621615588184

# Appendix to Master thesis - supplementary material

## 1 Consent statement for online respondents

«Samtykke og instruksjoner», fra førstesiden nettskjema.no skjemaet

### Utprøvning av et nytt skjema for egenrapportert helse

Regional kompetansetjeneste for rehabilitering, Helse Sør-Øst, ber nå personer med og uten alvorlige helsetilstander, om å delta i utprøvningen av et nytt, internasjonalt måleskjema for egenrapportert helse. Undersøkelsen er anonym, og de som ønsker å delta, samtykker til deltakelse ved å gjennomføre og sende inn undersøkelsen. Resultatene vil både kunne bli benyttet i forbedringsarbeid i helsetjenesten og til forskning. Ingen besvarelser vil kunne spores tilbake til enkeltpersoner, og vil dermed heller ikke kunne kobles mot andre helsedata eller persondata.

I undersøkelsen vil man oppleve at flere spørsmål er like. Dette skyldes at vi her sammenligner et nytt skjema med et annet. Vi ber derfor om tålmodighet til å besvare hele undersøkelsen, som består av rundt 100 spørsmål, og tar mellom 15-20 minutter å fullføre i sin helhet. Det er viktig at du gjennomfører hele undersøkelsen, om du ønsker å delta.

Vi ber om at du deltar med kun 1 besvarelse. Ønsker du det, kan lenken til undersøkelsen deles med andre. Siden dette er en anonym undersøkelse, bruker alle som gjennomfører undersøkelsen den samme nettlenken. Resultatet fra undersøkelsen vil bli oppsummert og gjort tilgjengelig på våre websider (www.sunnaas.no/rkr) i løpet av 2019, i tillegg til publisering i internasjonale tidsskrift.

2

Example of T-score look-up table from www.healthmeasures.org,

this one for Depression short form.

Raw score summed from the 8 items as basis for looking up

T-scores and Standard error of measurement for an individual.

More information at http://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Ad

Alternative method for scoring multiple individuals:

http://www.healthmeasures.net/score-and-interpret/calculate-scores

**Adult v1.0 - Depression 8a**

*Short Form Conversion Table*

| Raw Summed Score | T-score | SE* |
|---|---|---|
| 8 | 38.2 | 5.7 |
| 9 | 44.7 | 3.3 |
| 10 | 47.5 | 2.7 |
| 11 | 49.4 | 2.3 |
| 12 | 50.9 | 2.0 |
| 13 | 52.1 | 1.9 |
| 14 | 53.2 | 1.8 |
| 15 | 54.1 | 1.8 |
| 16 | 55.1 | 1.7 |
| 17 | 55.9 | 1.7 |
| 18 | 56.8 | 1.7 |
| 19 | 57.7 | 1.7 |
| 20 | 58.5 | 1.7 |
| 21 | 59.4 | 1.7 |
| 22 | 60.3 | 1.7 |
| 23 | 61.2 | 1.7 |
| 24 | 62.1 | 1.8 |
| 25 | 63.0 | 1.8 |
| 26 | 63.9 | 1.8 |
| 27 | 64.9 | 1.8 |
| 28 | 65.8 | 1.8 |
| 29 | 66.8 | 1.8 |
| 30 | 67.7 | 1.8 |
| 31 | 68.7 | 1.8 |
| 32 | 69.7 | 1.8 |
| 33 | 70.7 | 1.8 |
| 34 | 71.7 | 1.8 |
| 35 | 72.8 | 1.8 |
| 36 | 73.9 | 1.8 |
| 37 | 75.0 | 1.9 |
| 38 | 76.4 | 2.0 |
| 39 | 78.2 | 2.4 |
| 40 | 81.3 | 3.4 |

*SE = Standard Error on T-score metric

# 3 PROMIS 57 (and 29) Items in Norwegian and English, with response options

| Item ID | PROMIS-57 Norwegian **domain names** and items | PROMIS-57 English **domain names** and items | **Response options** | Also included in |
|---|---|---|---|---|
| | **Fysisk funksjon** | **Physical function** | | **Promis-29** |
| PFA11 | Klarer du å utføre gjøremål som støvsuging eller hagearbeid? | Are you able to do housework such as vacuum cleaning or gardening? | Without any difficulty /With a little difficulty /With some difficulty /With much difficulty /Unable to do | Yes |
| PFA21 | Klarer du å gå opp og ned trapper i normalt tempo? | Are you able to climb and descend stairs in a normal tempo? | --- " --- | Yes |
| PFA23 | Klarer du å gå en tur på minst 15 minutter? | Are you able to go for a walk for at least 15 minutes? | --- " --- | Yes |
| PFA53 | Klarer du å gjøre ærend og gå i butikker? | Are you able to do errands and go shopping? | --- " --- | Yes |
| PFC12 | Begrenser helsen din deg nå i å utføre fysisk arbeid i to timer? | Does your current health status prevent you from performing physical work for two hours? | Not at all /A little / Some / Quite a lot/ Cannot manage | No |
| PFB1 | Begrenser helsen din deg nå i å gjøre enkelt husarbeid som å støvsuge eller feie gulv? | Does your current health status prevent you from performing moderate house work such as vacuum cleaning, sweeping or carrying groceries? | --- " --- | No |
| PFA5 | Begrenser helsen din deg nå i å løfte eller bære dagligvarer? | Does your current health status prevent you from lifting or carrying groceries? | --- " --- | No |
| PFA4 | Begrenser helsen din deg nå i å utføre tungt husarbeid som å vaske gulv, løfte eller flytte tunge møbler? | Does your current health status prevent you from performing hard housework such as mopping the floor, or lifting or moving heavy furniture? | --- " --- | No |
| | **Angst** | **Anxiety** | | |
| EDANX01 | Jeg følte meg redd | I felt fearful | Never /Rarely /Sometimes /Often /Always | Yes |
| EDANX40 | Det var vanskelig å fokusere på noe annet enn min angst | I found it hard to focus on anything other than my anxiety | --- " --- | Yes |
| EDANX41 | Bekymringene mine overveldet meg. | My worries overwhelmed me. | --- " --- | Yes |
| EDANX53 | Jeg følte meg urolig | I felt uneasy | --- " --- | Yes |
| EDANX46 | Jeg følte meg nervøs. | I felt nervous. | --- " --- | No |
| EDANX07 | Jeg følte at jeg trengte hjelp for min angst.. | I felt like I needed help for my anxiety | --- " --- | No |
| EDANX05 | Jeg følte meg engstelig | I felt anxious. | --- " --- | No |
| EDANX54 | Jeg følte meg anspent | I felt tense | --- " --- | No |
| | **Depresjon** | **Depression** | | |
| EDDEP04 | Jeg følte meg verdiløs | I felt worthless | Never /Rarely /Sometimes /Often /Always | Yes |
| EDDEP06 | Jeg følte meg hjelpeløs | I felt helpless | --- " --- | Yes |
| EDDEP29 | Jeg følte meg deprimert | I felt depressed | --- " --- | Yes |
| EDDEP41 | Jeg følte meg uten håp | I felt hopeless | --- " --- | Yes |
| EDDEP22 | Jeg følte meg mislykket | I felt like a failure | --- " --- | No |
| EDDEP36 | Jeg følte meg ulykkelig | I felt unhappy | --- " --- | No |
| EDDEP05 | Følte jeg at jeg ikke hadde noe å se frem til | I felt that I had nothing to look forward to. | --- " --- | No |
| EDDEP09 | Jeg følte at ingenting kunne muntre meg opp | I felt that nothing could cheer me up | --- " --- | No |

| | **Utmattelse** | **Fatigue** | | |
|---|---|---|---|---|
| HI7 | Jeg føler meg utmattet | I feel fatigued | Not at all /A little bit /Somewhat /Quite a bit /Very much | Yes |
| AN3 | Jeg har vanskelig for å begynne med ting fordi jeg er trett | I have trouble starting things because I am tired. | --- " --- | Yes |
| FATEXP41 | Hvor nedkjørt følte du deg i gjennomsnitt | How run-down did you feel on average? | --- " --- | Yes |
| FATEXP40 | Hvor utmattet var du i gjennomsnitt? | How fatigued were you on average? | --- " --- | Yes |
| FATEXP35 | Hvor mye var du gjennomsnittlig plaget av utmattelse? | How much were you bothered by your fatigue on average? | --- " --- | No |
| FATIMP49 | I hvilken grad har din utmattelse påvirket din fysiske funksjon? | To what degree did your fatigue interfere with your physical functioning? | --- " --- | No |
| FATIMP3 | Hvor ofte måtte du presse deg selv for å få ting gjort på grunn av utmattelse? | How often did you have to push yourself to get things done because of your fatigue? | Never /Rarely /Sometimes /Often /Always | No |
| FATIMP16 | Hvor ofte hadde du problemer med å fullføre ting på grunn av utmattelse? | How often did you have trouble finishing things because of your fatigue? | --- " --- | No |
| | **Søvnvansker** | **Sleep disturbance** | | |
| Sleep109 | Søvnkvaliteten min var | My sleep quality was | Very poor /Poor /Fair / Good / Very good | Yes |
| Sleep116 | Søvnen gjorde meg opplagt | My sleep was refreshing | Not at all /A little bit /Somewhat /Quite a bit /Very much | Yes |
| Sleep20 | Jeg hadde problemer med søvnen | I had a problem with my sleep | --- " --- | Yes |
| Sleep44 | Jeg hadde vanskelighet med å sovne | I had difficulty falling asleep | --- " --- | Yes |
| Sleep108 | Jeg sov urolig | My sleep was restless | --- " --- | No |
| Sleep72 | Jeg strevde med å sovne | I tried hard to get to sleep. | --- " --- | No |
| Sleep67 | Jeg bekymret meg for ikke å klare å sovne | I worried about not being able to fall asleep. | --- " --- | No |
| Sleep115 | Jeg var fornøyd med søvnen min | I was satisfied with my sleep | --- " --- | No |
| | **Evne til å delta i sosiale roller og aktiviteter** | **Ability to Participate in Social Roles and Activities** | | |
| SRPPER11_CaPS | Jeg har problemer med å utføre mine vanlige fritidsaktiviteter med andre | I have trouble doing all of my regular leisure activities with others | Never /Rarely /Sometimes /Often /Always | Yes |
| SRPPER18_CaPS | Jeg har problemer med å utføre alle de familieaktivitetene jeg ønsker å være med på | I have trouble doing all of the family activities that I want to do | --- " --- | Yes |
| SRPPER23_CaPS | Jeg har problemer med å utføre alt mitt vanlige arbeid (inkludert arbeid i hjemmet) | I have trouble doing all of my usual work (include work at home) | --- " --- | Yes |
| SRPPER46_CaPS | Jeg har problemer med å utføre alle aktiviteter med venner som jeg ønsker å gjøre | I have trouble doing all of the activities with friends that I want to do | --- " --- | Yes |
| SRPPER15_CaPS | Jeg må begrense de tingene jeg gjør for å ha det moro sammen med andre | I have to limit the things I do for fun with others | --- " --- | No |
| SRPPER28r1 | Jeg må begrense mine vanlige aktiviteter med venner | I have to limit my regular activities with friends | --- " --- | No |
| SRPPER14r1 | Jeg må begrense mine vanlige familieaktiviteter | I have to limit my regular family activities | --- " --- | No |
| SRPPER26_CaPS | Jeg har problemer med å gjøre alt arbeidet som er viktig for meg (inkludert arbeid i hjemmet) | I have trouble doing all of the work that is really important to me (include work at home)home) | --- " --- | No |

| | Pain Interference | Smertepåvirkning | | |
|---|---|---|---|---|
| PAININ9 | I hvor stor grad påvirket smerter dine daglige aktiviteter? | How much did pain interfere with your day to day activities? . | Not at all /A little bit /Somewhat /Quite a bit /Very much | Yes |
| PAININ22 | I hvor stor grad har smerter påvirket ditt arbeid i hjemmet? | How much did pain interfere with work around the home? . | --- " --- | Yes |
| PAININ31 | I hvor stor grad har smerter påvirket evnen din til å delta i sosiale aktiviteter? | How much did pain interfere with your ability to participate in social activities? . | --- " --- | Yes |
| PAININ34 | I hvor stor grad har smerter påvirket ditt husarbeid? | How much did pain interfere with your enjoyment of life? | --- " --- | Yes |
| PAININ12 | I hvor stor grad har smerter påvirket ting du vanligvis gjør for å ha det moro? | How much did pain interfere with the things you usually do for fun? | --- " --- | No |
| PAININ36 | I hvor stor grad har smerter påvirket gleden over å ta del i sosiale aktiviteter? | How much did pain interfere with your enjoyment of social activities? . | --- " --- | No |
| PAININ3 | I hvor stor grad har smerter påvirket livsgleden din? | How much did pain interfere with your household chores? . | --- " --- | No |
| PAININ13 | I hvor stor grad har smerter påvirket ditt familieliv? | How much did pain interfere with your family life? | --- " --- | No |
| Global07 | Hvordan vil du gradere smertene dine i gjennomsnitt? | How would you rate your pain on average | 0 1 2 3 4 5 6 7 8 9 10 | Yes |

4

## IRT PARAMETERS PER ITEM FOR PROMIS 57 GRADED RESPONSE MODEL

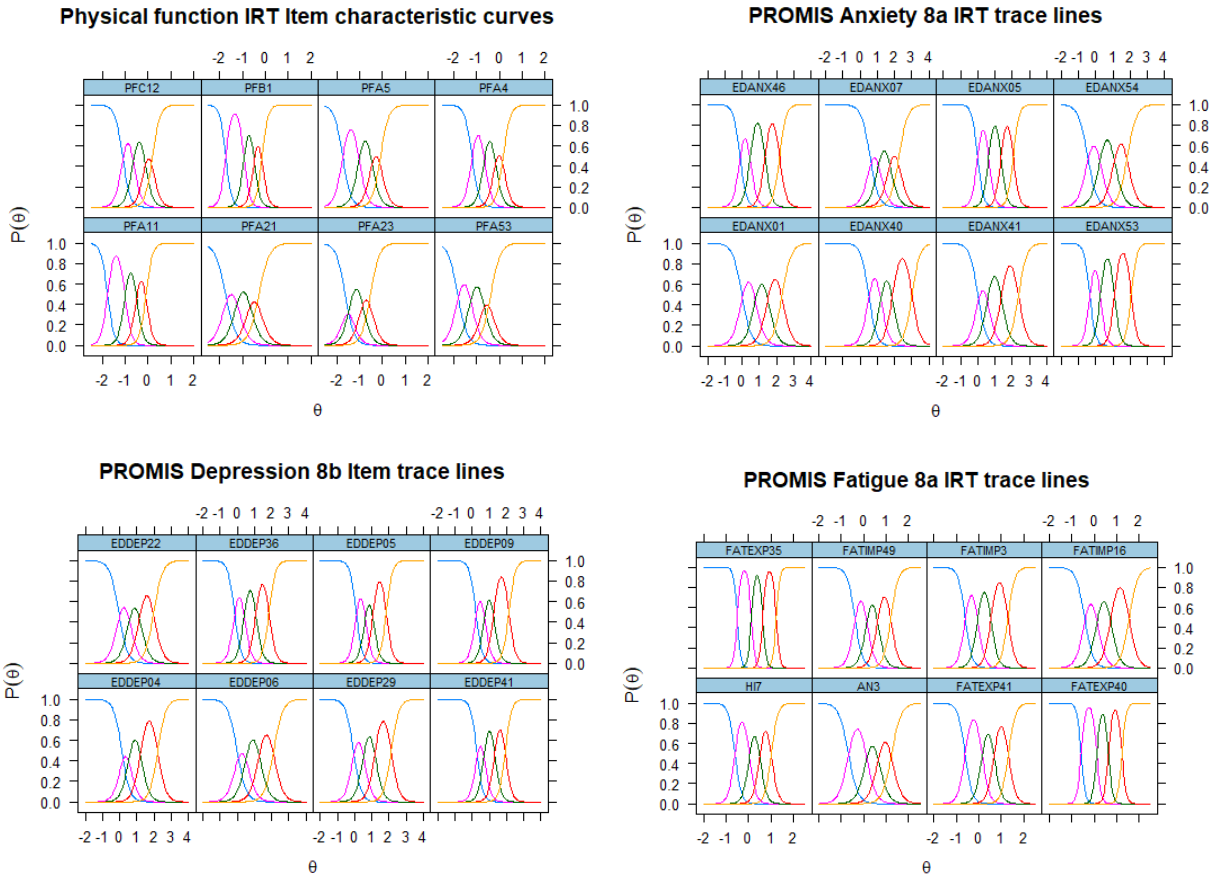
| Physical function: | a1 (=discrimination) | d1 | d2 | d3 | d4 | **Average difficulty** |
|---|---|---|---|---|---|---|
| PFA11 | 7,001 | -1,78 | -1,01 | -0,51 | -0,09 | **-0,85** |
| PFA21 | 4,314 | -1,73 | -1,23 | -0,69 | -0,27 | **-0,98** |
| PFA23 | 5,078 | -1,63 | -1,38 | -0,89 | -0,52 | **-1,10** |
| PFA53 | 4,826 | -1,83 | -1,26 | -0,72 | -0,38 | **-1,05** |
| PFC12 | 6,072 | -1,12 | -0,64 | -0,14 | 0,20 | **-0,43** |
| PFB1 | 7,771 | -1,73 | -0,93 | -0,49 | -0,13 | **-0,82** |
| PFA5 | 5,559 | -1,74 | -1,02 | -0,47 | -0,08 | **-0,83** |
| PFA4 | 6,421 | -1,21 | -0,66 | -0,19 | 0,16 | **-0,47** |
| Avg slope: | 5,88025 | | | | Avg of avg | **-0,82** |
| Anxiety: | a1 (=discrimination) | d1 | d2 | d3 | d4 | |
| EDANX01 | 3,76 | -0,01 | 0,77 | 1,50 | 2,33 | **1,15** |
| EDANX40 | 4,393 | 0,50 | 1,21 | 1,89 | 3,02 | **1,66** |
| EDANX41 | 4,157 | 0,00 | 0,58 | 1,37 | 2,37 | **1,08** |
| EDANX53 | 6,053 | -0,35 | 0,27 | 1,09 | 2,06 | **0,77** |
| EDANX46 | 5,347 | -0,13 | 0,47 | 1,33 | 2,18 | **0,96** |
| EDANX07 | 3,975 | 0,58 | 1,10 | 1,72 | 2,27 | **1,42** |
| EDANX05 | 6,018 | -0,02 | 0,63 | 1,35 | 2,06 | **1,00** |
| EDANX54 | 3,767 | -0,47 | 0,25 | 1,08 | 1,84 | **0,68** |
| Avg slope: | 4,68375 | | | | Avg of avg | **1,09** |
| Depression: | a1 (=discrimination) | d1 | d2 | d3 | d4 | |
| EDDEP04 | 4,239 | 0,12 | 0,57 | 1,22 | 2,22 | **1,04** |
| EDDEP06 | 3,737 | 0,00 | 0,55 | 1,29 | 2,12 | **0,99** |
| EDDEP29 | 4,456 | -0,07 | 0,52 | 1,19 | 2,13 | **0,94** |
| EDDEP41 | 5,556 | 0,27 | 0,70 | 1,31 | 1,93 | **1,05** |

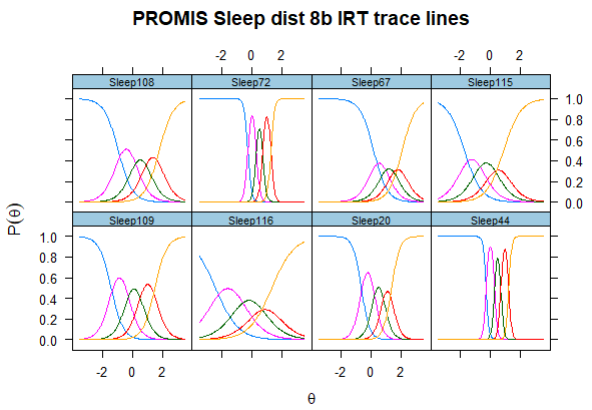
| | | | | | | |
|---|---|---|---|---|---|---|
| EDDEP22 | 3,889 | -0,06 | 0,57 | 1,18 | 1,99 | **0,92** |
| EDDEP36 | 5,25 | -0,16 | 0,42 | 1,09 | 1,86 | **0,80** |
| EDDEP05 | 5,594 | 0,09 | 0,61 | 1,07 | 1,84 | **0,90** |
| EDDEP09 | 5,446 | 0,21 | 0,72 | 1,25 | 2,14 | **1,08** |
| Avg slope: | 4,770875 | | | | Avg of avg | **0,97** |
| Fatigue: | a1 (=discrimination) | d1 | d2 | d3 | d4 | |
| HI7 | 7,01 | -0,61 | 0,04 | 0,51 | 1,02 | **0,24** |
| AN3 | 4,786 | -0,68 | 0,12 | 0,66 | 1,25 | **0,34** |
| FATEXP41 | 6,331 | -0,62 | 0,14 | 0,68 | 1,32 | **0,38** |
| FATEXP40 | 10,803 | -0,58 | 0,12 | 0,64 | 1,24 | **0,35** |
| FATEXP35 | 12,51 | -0,50 | 0,13 | 0,63 | 1,24 | **0,38** |
| FATIMP49 | 5,987 | -0,39 | 0,14 | 0,63 | 1,21 | **0,40** |
| FATIMP3 | 6,564 | -0,61 | -0,06 | 0,54 | 1,30 | **0,29** |
| FATIMP16 | 5,279 | -0,43 | 0,14 | 0,74 | 1,56 | **0,50** |
| Avg slope: | 7,40875 | | | | Avg of avg | **0,36** |
| Sleep interference: | a1 (=discrimination) | d1 | d2 | d3 | d4 | |
| Sleep109 | 2,501 | -1,414 | -0,311 | 0,546 | 1,504 | **0,08** |
| Sleep116 | 1,323 | -2,42 | -0,781 | 0,43 | 1,339 | **-0,36** |
| Sleep20 | 3,636 | -0,655 | 0,198 | 0,812 | 1,37 | **0,43** |
| Sleep44 | 10,211 | -0,291 | 0,273 | 0,693 | 1,222 | **0,47** |
| Sleep108 | 2,132 | -0,939 | 0,122 | 0,934 | 1,801 | **0,48** |
| Sleep72 | 8,391 | -0,262 | 0,307 | 0,725 | 1,282 | **0,51** |
| Sleep67 | 2,212 | 0,175 | 0,889 | 1,493 | 2,081 | **1,16** |
| Sleep115 | 1,763 | -1,741 | -0,745 | 0,159 | 0,885 | **-0,36** |
| Avg slope: | 4,021125 | | | | Avg of avg | **0,30** |
| Social: | a1 (=discrimination) | d1 | d2 | d3 | d4 | |
| SRPPER11_CaPS | 6,622 | -1,24 | -0,51 | 0,00 | 0,46 | **-0,32** |
| SRPPER18_CaPS | 7,017 | -1,36 | -0,62 | -0,02 | 0,45 | **-0,39** |
| SRPPER23_CaPS | 5,244 | -1,19 | -0,55 | 0,02 | 0,49 | **-0,31** |
| SRPPER46_CaPS | 9,2 | -1,17 | -0,40 | 0,14 | 0,56 | **-0,22** |
| SRPPER15_CaPS | 7,308 | -1,04 | -0,46 | 0,05 | 0,42 | **-0,26** |
| SRPPER28r1 | 10,023 | -1,08 | -0,46 | 0,00 | 0,48 | **-0,27** |
| SRPPER14r1 | 8,302 | -1,27 | -0,56 | -0,10 | 0,38 | **-0,39** |
| SRPPER26_CaPS | 6,472 | -1,11 | -0,48 | 0,04 | 0,48 | **-0,27** |
| Avg slope: | 7,5235 | | | | Avg of avg | **-0,30** |
| Pain interference: | a1 (=discrimination) | d1 | d2 | d3 | d4 | |
| PAININ9 | 8,926 | -0,37 | 0,26 | 0,72 | 1,17 | **0,45** |
| PAININ22 | 10,84 | -0,11 | 0,35 | 0,73 | 1,22 | **0,55** |

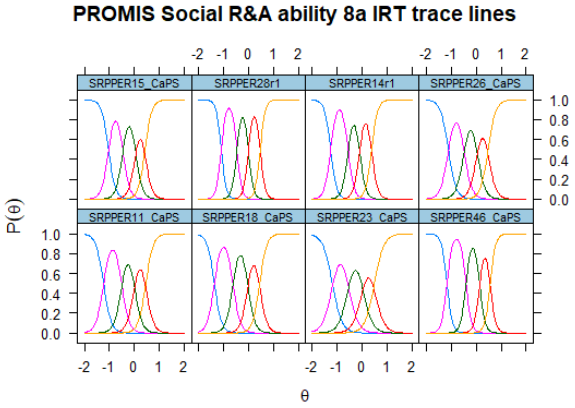
| | | | | | | |
|---|---|---|---|---|---|---|
| PAININ31 | 10,624 | 0,05 | 0,36 | 0,78 | 1,25 | **0,61** |
| PAININ34 | 8,302 | -0,08 | 0,43 | 0,77 | 1,25 | **0,59** |
| PAININ12 | 8,234 | -0,06 | 0,38 | 0,74 | 1,22 | **0,57** |
| PAININ36 | 7,732 | 0,00 | 0,46 | 0,78 | 1,25 | **0,62** |
| PAININ3 | 5,576 | -0,15 | 0,44 | 0,81 | 1,23 | **0,58** |
| PAININ13 | 7,488 | -0,02 | 0,46 | 0,84 | 1,34 | **0,66** |
| Avg slope: | 8,46525 | | | | Avg of avg | **0,58** |

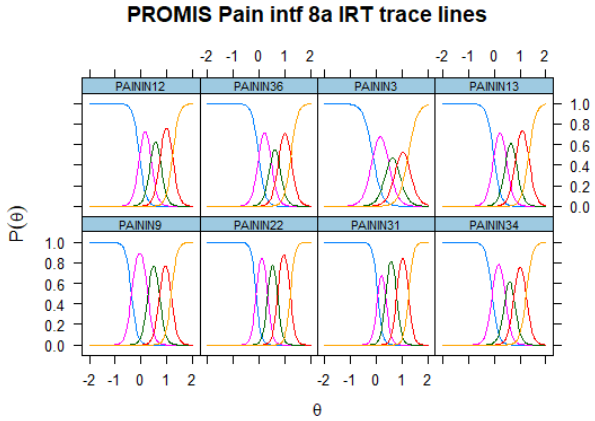
R-code per domain: PFgrmodel <- mirt (PFdata57, 1, rep("graded", 8), SE = TRUE)
coef(PFgrmodel, IRTpars=TRUE, simplify =TRUE)

# 5 PROMIS 57 IRT ICC plots:

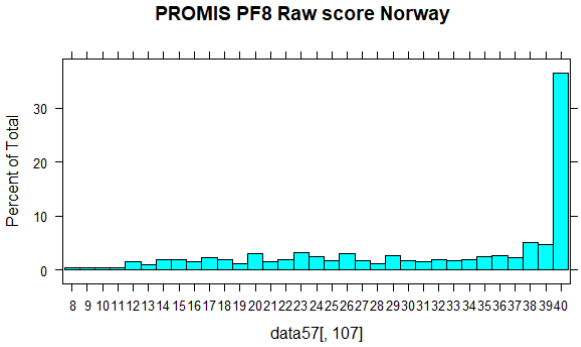


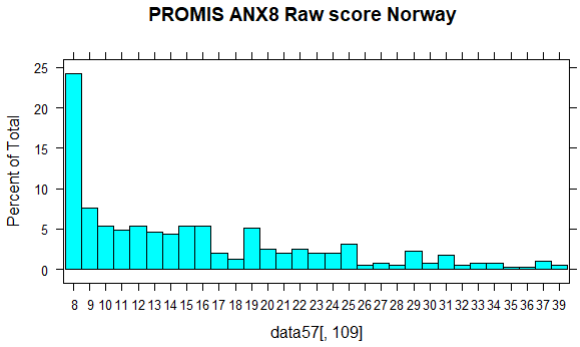


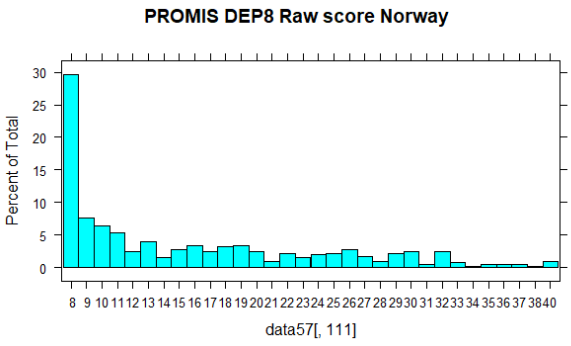



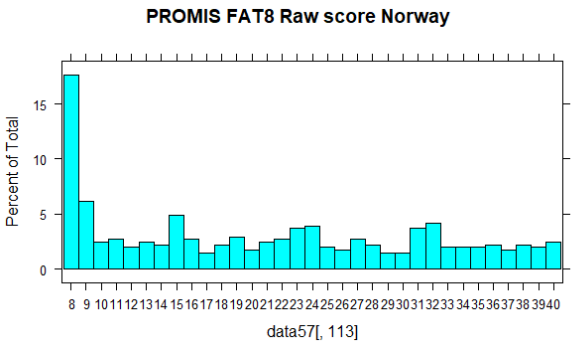



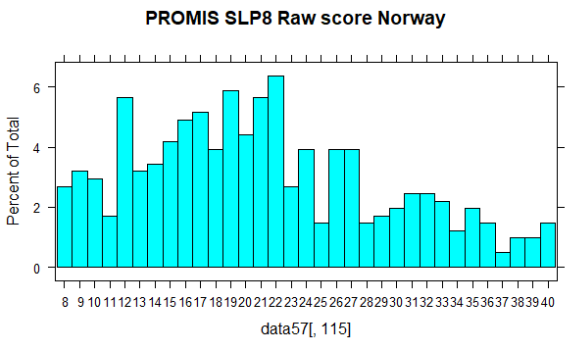

# 6 HISTOGRAMS of PROMIS-57 SCORE distributions

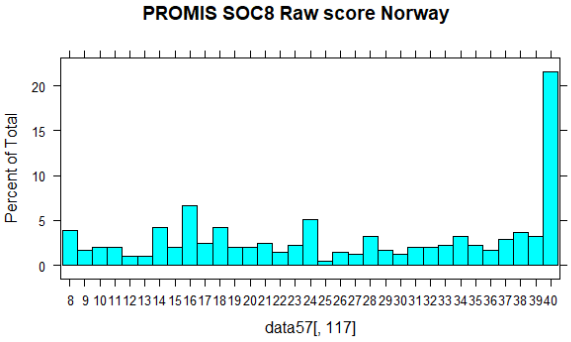


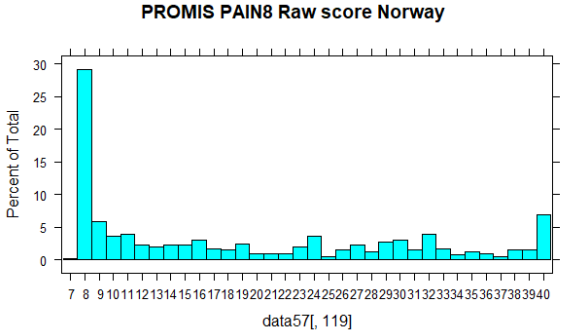



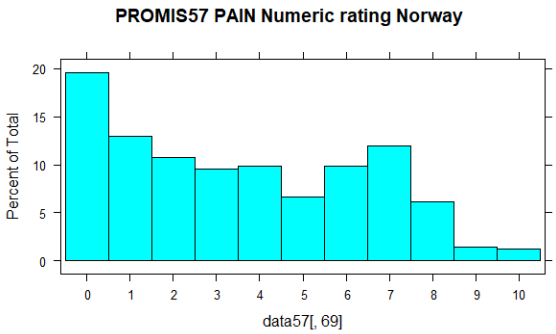



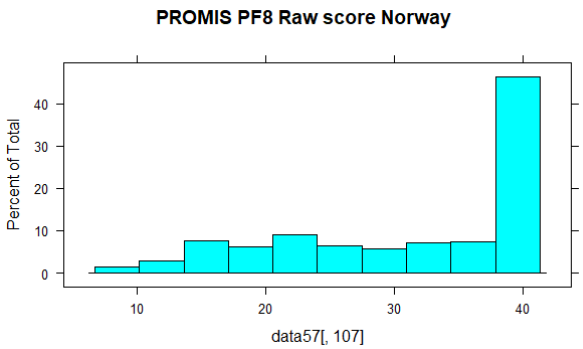



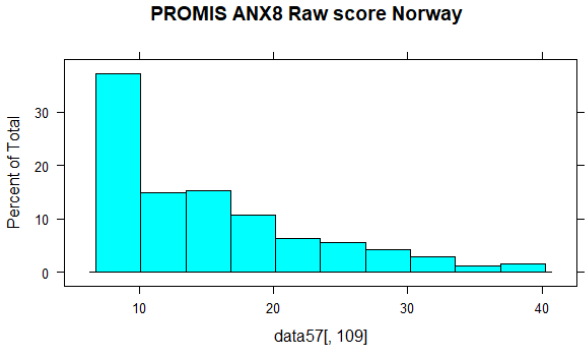

PROMIS SLP8 Raw score Norway
Percent of Total
data57[, 115]

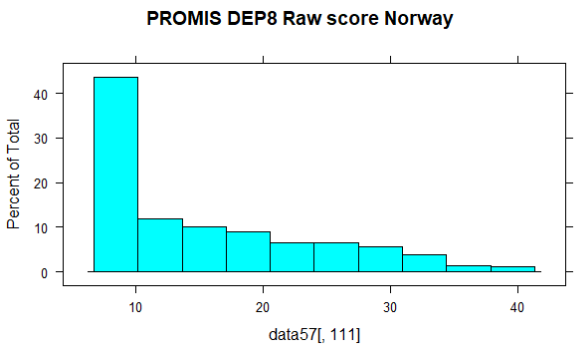

PROMIS SOC8 Raw score Norway
Percent of Total
data57[, 117]

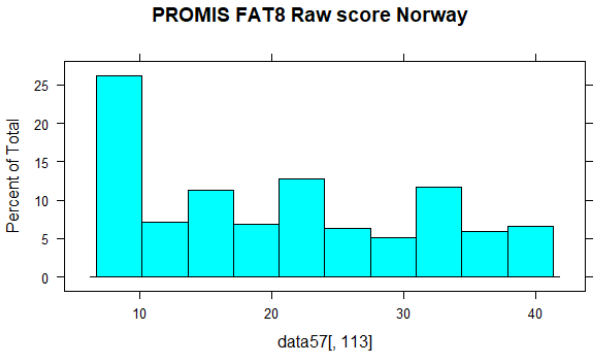

PROMIS PAIN8 Raw score Norway
Percent of Total
data57[, 119]

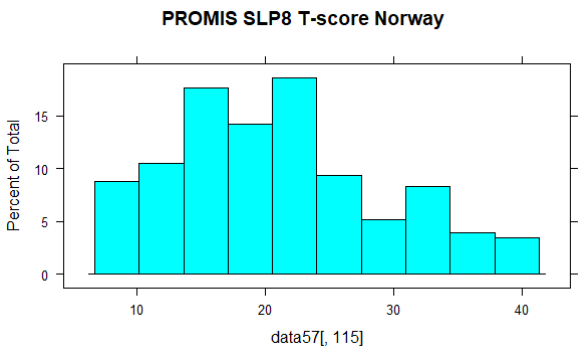

PROMIS57 PAIN Numeric rating Norway
Percent of Total
data57[, 69]

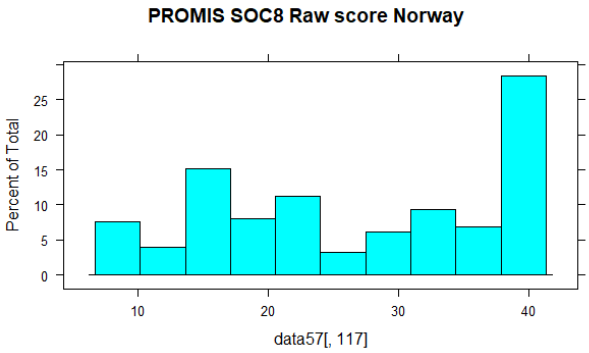

PROMIS PF8 Raw score Norway
Percent of Total
data57[, 107]

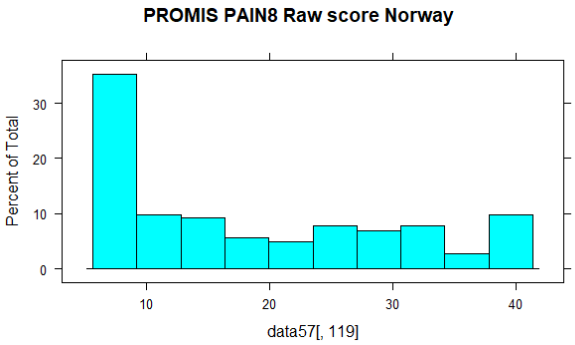

PROMIS ANX8 Raw score Norway
Percent of Total
data57[, 109]

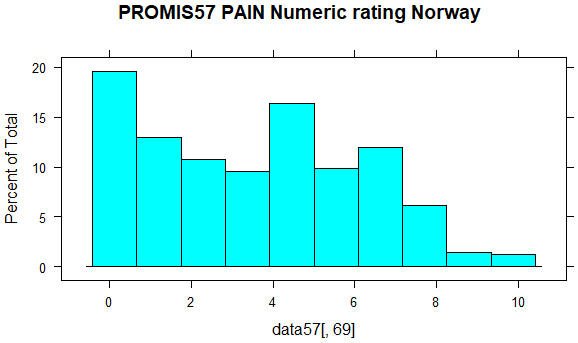

PROMIS DEP8 Raw score Norway
Percent of Total
data57[, 111]


PROMIS FAT8 Raw score Norway
Percent of Total
data57[, 113]


PROMIS SLP8 T-score Norway
Percent of Total
0
5
10
15
10
20
30
40
data57[, 115]


PROMIS SOC8 Raw score Norway
Percent of Total
0
5
10
15
20
25
10
20
30
40
data57[, 117]


PROMIS PAIN8 Raw score Norway
Percent of Total
0
10
20
30
10
20
30
40
data57[, 119]


PROMIS57 PAIN Numeric rating Norway
Percent of Total
0
5
10
15
20
0
2
4
6
8
10
data57[, 69]