# Assessment of depression in the adult general population using self-reported measures.

## Psychometric approaches for screening and severity appraisal

Gemma Vilagut Saiz

TESI DOCTORAL UPF / 2016

DIRECTORS DE LA TESI

Dr. Carlos García Forero

Dr. Jordi Alonso Caballero

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Universitat Pompeu Fabra Barcelona

*Al Miquel, la Clàudia i l'Ada,*
*al Cèsar i la Maite (allà on siguis)*

# Agraïments

Jordi, m'ha costat, però al final aquí està. Gràcies pel teu suport, confiança, ensenyaments i ajuda durant tot aquest temps, no només els anys de tesi, sinó en els 18 anys i mig (que es diuen ràpid!) que fa que estic a l'IMIM. Estic segura que sense alguna d'aquestes coses aquesta tesi no hagués tirat endavant, així que te la dec a tu. Carlos, amb tu no han estat tants anys, però si molt intensos. Ets un pou de coneixement. Moltes gràcies per la teva disponibilitat i ganes, per ajudar-me, escoltar-me i animar-me sempre, i per les llarguíssimes converses/discussions sempre interessant. Amb cafè o sense, i dinars (i sopars) pel mig, que no s'acabin!

Al llarg d'aquest temps hi ha hagut moltes altres persones que, potser sense ni adonar-se'n, han aportat petits grans de sorra, i fins i tot muntanyes senceres, a aquesta tesi, tant a nivell professional com personal. De l'IMIM, l'Antònia, per ser la primera en confiar en mi i estar-ne sempre pendent; i l'Olatz, perquè és la millor companya de despatx que es pot tenir, en tots els sentits (jo de gran vull ser com tu); la Puri i la Carme, que sempre estan en tot i a punt per ajudar; l'Àngels, que mai té un no per resposta, i fa uns massatges espectaculars ;-) ; la Montse, per la seguretat i els coneixements que sap transmetre tan bé; el Dave, per l'ajuda amb l'anglès i el suport tècnic amb el Refman i les cerques bibliogràfiques i els seus programets màgics, i per les seves fotos que ens alegren la vista al despatx; i la Gabi, pel seu carinyo i ànims infinits. Però també, molt importants, la MOMA, l'Albert, el *Pepito*, la Núria D, la Mercks, la

Sònia, la Chus, la Sandra, el Miki, la Sali, la Moni, la Yolanda, la Susa, la Itxaso, les Andreas, el Pere, el David, la Roser, l'Alison, la Nancy, el Fernando, el Ronny, els Rons,... I, en definitiva, TOTA la gent amb la que he col·laborat en algun moment i que m'han vist i ajudat a créixer, sobretot els investigadors d'ESEMeD i del World Mental Health.

A nivell personal, agrair al Miquel per acompanyar-me en el camí, animar-me, donar-me suport i aguantar-me (que a vegades no deu ser fàcil) especialment en aquests darrers mesos en els que combinar feina, tesi i vida no ha estat una tasca planera. A la Clàudia i l'Ada, per alegrar el camí posant-li música i floretes de colors cada dia. Al papa, al Jordi i a la Cari pel seu suport enorme i l'estima incondicional. I finalment, gràcies també a la resta de la família i als amics de Montcada, de la Plana i de l'escola, amb menció especial al Jordi V, l'Anna, el Toni, el Raúl, el Jordi i l'Aran, i els Xelus, i el Gabri i la Berta.

Aquesta tesi ha estat possible gràcies al finançament rebut per a cadascun dels projectes que n'han format part, i el finançament del Departament de Salut de la Generalitat de Catalunya (AGAUR 2014 SGR 748; AGAUR 2009 SGR 1095). Per a la impressió d'aquesta tesi doctoral he rebut l'ajuda de l'IMIM (Institut Hospital del Mar d'Investigacions Mèdiques).

## Summary

This thesis provides evidence on the validity and diagnostic accuracy of generic and specific self-reported measures, developed from different psychometric approaches, to assess depression in the general population.

First, we compare the reliability and diagnostic accuracy of the 12-item Short Form Health Survey (SF-12) traditional scoring with Multidimensional Item Response Theory (MIRT) scoring in an adult general population sample from 6 European Countries (n=21,425). Secondly, we conduct systematic literature review and meta-analysis of the diagnostic accuracy of the Center for Epidemiologic Studies Depression (CES-D) as a screener for major depression. Finally, we assess the psychometric properties of IRT-based Patient Reported Outcomes Measurement Information System (PROMIS) Depression measures in general population (n=1,503) and clinical (n=218) samples.

Our results indicate that: a) the MIRT SF-12 model is more reliable and has comparable diagnostic accuracy than other scoring methods; b) general and specific measures herein included yield good diagnostic accuracy (area under the curve values over 0.85) as depression screeners, with sensitivities and specificities generally over 80% at the selected cut-off points; c) PROMIS Depression meets IRT assumptions, its measures are highly reliable and show good construct validity and responsiveness to change; d) one PROMIS item showed signs of differential item functioning according to country (Spain and

US), but had negligible effects at test level, supporting measurement invariance.
We conclude that self-reported measures are adequate for assessing depression in the general population, and provide additional information beyond detection of pathological individuals. The IRT psychometric approach provide higher flexibility and precision in administering and scoring questionnaires in survey studies, also allowing direct comparisons between populations.

## Resum

Aquesta tesi proporciona evidencia sobre la validesa i la capacitat diagnòstica de mesures genèriques i específiques auto-reportades, construïdes des de diferents aproximacions psicomètriques per avaluar depressió en la població general.

Primer, comparem la fiabilitat i capacitat diagnòstica del qüestionari Short Form Health Survey (SF-12) utilitzant el càlcul de puntuació habitual, amb la puntuació obtinguda a partir de Teoria de Resposta a l'Ítem Multidimensional (TRIM), en una mostra representativa de la població general adulta de 6 països europeus (n=21.425). En segon lloc, fem una revisió sistemàtica de la literatura amb meta-anàlisi per avaluar la validesa diagnòstica del qüestionari Center for Epidemiologic Studies Depression (CES-D) per al cribratge de depressió major. Finalment, avaluem les propietats psicomètriques de les mesures de depressió del Patient Reported Outcomes

Measurement Information System (PROMIS de Depressió), basades en aproximació de Teoria de Resposta a l'Ítem (TRI) en una mostra de població general (n=1,503) i una mostra clínica de pacients (n=218).
Els resultats indiquen que: a) la puntuació basada en el model TRIM del SF-12 és més fiable però presenta una capacitat diagnòstica similar que la dels altres mètodes de puntuació; b) les mesures genèriques i especifiques que hem avaluat proporcionen una bona capacitat diagnòstica (àrea sota la corba superior a 0.85) per al cribratge de depressió en població general, amb sensibilitats i especificitats per sobre del 80% en els punts de tall seleccionats; c) el PROMIS de Depressió compleix totes les assumpcions de TRI, i les seves mesures presenten fiabilitats altes i bona validesa de constructe i sensibilitat al canvi; d) un dels ítems PROMIS presenta signes de funcionament diferencial de l'ítem segons el país (Espanya i US) però els efectes a nivell del test són negligibles, donant suport a la invariància de les mesures.

Concloem que les mesures auto-reportades estudiades són adequades per avaluar depressió en la població general, i proporcionen una informació valuosa que va més enllà la detecció dicotòmica d'individus amb patologia o sense. L'aproximació psicomètrica basada en TRI proporciona major flexibilitat en l'administració i puntuació dels qüestionaris i precisió més elevada, i facilita comparacions directes entre poblacions.

# Preface

Depressive disorders and especially major depressive disorder are commonly occurring, serious, recurrent disorders causing substantial disability and burden, both at the individual and societal level. They have been linked to diminished role functioning and quality of life, interference with education and work performance, elevated stigma and important morbidity and mortality rates, causing a high economic burden for society, related to both direct and indirect costs [1,2].

The incorporation of nonfatal disability resulting from health conditions into an overall measure of disease burden at the societal level was a key factor in bringing mental illness in general, and depression in particular, to global attention, as leading causes of burden, stressing the public health significance of these disorders [3–6].

To accomplish the public health goals of reducing the incidence, prevalence, severity, and economic impact of the disorders [7,8], information at population level is needed on the presence of depressive disorders, and associated impacts. However, obtaining and using reliable and valid measures of mental disorders is specially challenging, as a distinctive feature of mental disorders is that no physiological or laboratory test or radiological examination exists to definitively establish the diagnosis, and their identification generally depends on thoughts, behaviors and feelings.

Several self-reported questionnaires aiming at assessing mental health or psychological distress are available that allow the construct of depression to be evaluated from a continuous perspective of severity but also to use established cut-off points to identify individuals that may be at risk for depression.

Most of these measures were developed under the Classical Test Theory (CTT) approach. However, Item Response Theory (IRT) has recently been used for the development of specific instruments for the assessment of depression [9–12], producing outcomes with higher precision and a greater administration flexibility.

The global aim of this dissertation is to provide evidence regarding the construct validity and diagnostic accuracy of currently available specific and generic self-reported questionnaires developed according to different psychometric approaches (CTT and IRT) that are used to assess presence and severity of depression in the general population.

# Index

# 1. INTRODUCTION

## 1.1. Frequency and burden of depressive disorders

### a) Prevalence and individual impact of depressive disorders

Depressive disorders are commonly occurring disorders. Results from a comprehensive systematic review and meta-analysis of psychiatric epidemiological adult population surveys reported a pooled 12-month prevalence of depressive disorders of 5.4% (4.9%-6.0%) across 148 surveys, and a lifetime prevalence of 9.6% (8.5% to 10.7%) across 83 surveys [13].

Based on data from the WHO World Mental Health (WMH) surveys initiative, a major source of global information on the prevalence and correlates of common mental disorders, the median age of onset of major depression episode (MDE) is in the middle 20s, being similar across country groups (median 25.7 for high-income versus 24 for low and middle-income countries), with inter-quartile ranges between mid-late adolescence to the late 30s or early 40s [14]. This has an important impact on critical developmental stages of life, beginning early in life and influencing life opportunities through interference with performance in formative, work-related and personal activities. A clear example is the association between early onset of depression and school dropout found in several studies, with a 30% to 60% elevated odds of failure to complete secondary school, as well as subsequent unemployment and low income, especially in middle and high-income countries [15–18]. Some studies also suggest that depression is

disruptive to social relationships, with psychiatric disorders and depression in particular, found to be positively associated with lower likelihood of marriage and higher likelihood of divorce [19].

Depressive disorders are also significantly associated with medical comorbidity, both physical and mental, and mortality rates. In particular, data from clinical samples and community epidemiological surveys indicate that major depressive disorder is significantly associated with a wide variety of chronic physical disorders, including cardiovascular disease, diabetes, arthritis, asthma, cancer, or respiratory disease [20–32] and with other mental disorders [33]. Many of these studies show that Major Depressive Disorder (MDD) is a consistent predictor of other subsequent conditions, such as chronic heart failure, stroke or diabetes [22,24–26,29,31,32,34], and plays an important role in the etiology, course, and outcomes associated with some chronic diseases [35]. Several studies suggest that depression substantially increases the risk of death [25,36,37]. Eaton et al. found that depressive disorders raised the risk of all-cause mortality by about 70%, with an interquartile range of relative risks of 1.3 to 2.2 [2].

The WMH surveys show that mental disorders, depressive disorders in particular, are associated with high levels of disability that are often substantially higher than those associated with physical conditions. The functional domains most affected by mental disorders include social life and personal relationships, and, to a lesser extent, work and household functioning [38]. Still according to the WMH surveys, individuals with MDE have more than 34 days totally out of role in a year, being one of the highest disability impact levels associated with

either mental disorders or physical conditions. When estimates are adjusted by several factors, including comorbidities, the annual number of days totally out of role due to MDE is estimated to be 9. While this is a high impact, results suggest that a good part of the disability associated with MDE is due to the tendency of this disorder to coexist with other disorders and conditions [39].

Depressive disorders have also been strongly associated to experienced discrimination and stigma [40,41], which act as a barrier to social participation and successful vocational integration. Non-disclosure of depression (a consequence of self-stigma) is itself a further barrier to seeking help and to receiving effective treatment. Stigma is indeed a major barrier to the reduction of disability associated to depressive disorders [42].

## b) Global Burden of depressive disorders

The Global Burden of Disease (GBD) study provides a systematic international estimate of the leading causes of death and of disability through the estimation of disability adjusted life years (DALYs), which represent the loss of a healthy years of life as an aggregate of the years of life lived with disability (YLD) with the years of life lost due to premature mortality (YLL). In the first GBD publication with results from 1990 [3], unipolar depressive disorder emerged as the fourth leading contributor to the global burden of disease, with 3.7% of all DALYs, and was one of the leading causes of years of life lived with disability (YLD), accounting for 10.7% of total YLD. These results have since had a significant influence on prioritizing depressive disorders, and mental disorders as a group, in public health agendas;

particularly in promoting the addition of mental health interventions to health management plans [4].

Updated GBD estimates from 2010, based on a complete epidemiological re-assessment of diseases, a higher number of diseases and injuries and improved methodology, still ranked Major Depressive disorder as the 11th leading cause of global DALYs, accounting for 2.5% (1.9%–3.2%). They are also the second leading cause of disability, explaining 8.2% (5.9%–10.8%) of all YLDs, after low back pain [43]. Recent estimates of the GBD study in its 2013 edition showed that depression continues to be the second leading cause of YLD out of 301 acute and chronic diseases and injuries [44] and the 11th cause of disability adjusted life years [45]. It is worth noting that some authors suggest that the burden of mental disorders, and of depression, in the GBD study could be underestimated by 50%, since associated risk of suicide and premature mortality are not adequately taken into account [46].

The high prevalence of depression and the significant disability associated with the disorder involve elevated direct and indirect costs. The economic burden of depression in the United States, including workplace costs, direct costs and suicide- related costs, was estimated to be $210.5 billion in 2010 [47]. In Europe, the total cost of mental disorders in 2011 was estimated to be €461 billion, nearly 1000€ per inhabitant [48]. Affective disorders alone, and especially unipolar depression, accounted for more than half of the indirect costs. In Spain, the economic burden of mental disorders is estimated 46€ billion in 2010 with mood disorders accounting for more than 10€

billion [49]. The total figure of brain disorders costs in Spain was 84€ billion, which correspond to nearly 8% of its gross domestic product [49].

Despite the substantial associated impact, only a limited proportion of cases with common mental disorders is identified and receives appropriate treatment, leading to high rates of unmet needs [50]. Many individuals with depression do not have access to treatment or do not take advantage of services. Among patients presenting at healthcare settings, under-recognition of depression is common, with the rate of missed diagnosis of depression approaching 50% of the cases [51]. If not effectively treated, depression is likely to become a chronic disease. Just experiencing one episode of depression places an individual at a 50% risk for experiencing another episode, and further increases the chances of having more depression episodes in the future.

## 1.2. Population assessment of depressive disorders

In view of the magnitude of the burden caused by depressive disorders and other mental disorders, together with evidence of a much wider treatment gap (when compared with physical conditions), they must become a public health priority if overall population health is to be improved [6,52–54]. From the population point of view, assessment of the presence of the disorder and its associated impacts, and evaluation of need for and barriers to treatment with accurate data, should be the basis for health service and prevention planning.

This is essential to the public health goals of reducing the incidence, prevalence, severity, and economic impact of these disorders [7,8].

## a) Assessment approaches in the general population

Information about mental disorders at the general population level can be obtained either from the specific perspective of psychiatric epidemiology, or through population health surveillance systems, defined by the WHO as "*the continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice"* [55], that focus on a more integrated approach to health including both its physical and mental aspects.

Psychiatric epidemiology, the study of mental disorders in the population, has traditionally used specific data collection systems to gather information on the magnitude, impact and correlates of mental disorders. Current methods, sometimes described as the third generation of psychiatric epidemiology [56], are mainly based on representative general population surveys with a special focus on accurate assessment of specific mental disorders using standardized diagnostic criteria. Examples of successful and richly informative studies of this sort are: the National Comorbidity Survey (NCS) [57], carried out in the years 2000 - 2002 and providing prevalence estimates for a nationally representative sample of the US for the first time; and replicated about a decade later with the National Comorbidity Survey Replication (NCS-R) study [58]; the Netherlands Mental Health Survey and Incidence Study (NEMESIS) carried out in 1996 for assessment of prevalence and incidence in the Netherlands, with follow-up measurements in the same sample at 12 and 36 months

and a subsequent replication and expansion in 2007-2013 [59]; and the World Mental Health (WMH) Surveys initiative, a set of population-based epidemiological studies launched by the World Health Organization (WHO) in more than 30 countries, all of them using broadly the same interview and comparable methods [60].

On the other hand, population health surveillance systems implemented in different countries and regions are giving increasing importance to mental health [7,61–64], relying on an integrated evaluation approach of physical and mental health. One way this has been done is by incorporating common mental health measures in already existing general population health interview surveys [65], providing an opportunity for more regular monitoring than large and costly psychiatric epidemiology surveys [7,61,64]. Other sources of information used in health surveillance systems include medical and mortality registers [61,64], and surveys targeting medical health-care providers and insurers that are used to estimate outpatient visits and hospitalizations and reflect access to and use of health care by persons with mental illness [7]. However, neither routinely collected statistics on deaths related to mental health problems, nor hospital discharge data reflect the wider reality of mental health. These statistics contain no information on the large numbers of people who suffer from mental health problems but neither died nor are hospitalized as a result. This information can only be gathered through general population surveys.

Surveillance data, with accurate and timely information on the prevalence and effects of mental illness, can be useful to detect and

characterize trends in prevalence and severity, and identify barriers to care. It can have numerous applications, such as: a) serving as an early warning system for impending public health emergencies through the determination of the distribution and spread of the disease and estimation of its impact; b) developing and evaluating public health intervention, or track progress towards specified goals; and c) Monitoring and clarifying the epidemiology and history of a health condition or the impact of treatments on outcomes, to allow priorities to be set, and to inform public health policy and strategies on promotion, prevention and treatment programs; d) generating hypotheses and stimulating research [7,55].

## b) Instruments for the assessment of mental health in population-based studies

Measurement of mental disorders and mental health is often more complex than in other fields of medicine and health. A distinctive feature of mental disorders is that their identification exclusively depends on thoughts, behaviors and feelings. There are no physiological or laboratory tests or imaging techniques to definitively establish the diagnosis. Thus, one of the challenges in population-based studies is the development of reliable and valid measures of mental disorders.

The following broad types of assessment methods are available to help estimate the prevalence and incidence of mental disorders in large samples of individuals in a standardized way:

1.- Structured diagnostic interviews.

2.- Self-reported questionnaires.

***1.- Structured diagnostic interviews:***

Several studies on the stability of psychiatric diagnoses over time have found diagnoses obtained from examination conducted by a well-trained psychiatrist to have low inter-rater reliability [66], prompting the need for structured interviews following standardized diagnostic criteria, such as the Diagnostic and Statistical Manual of Mental Disorders (DSM) [67], or the WHO International Classification of Disease (ICD) [68]. Structured diagnostic interviews are developed to be as consistent with the selected standardized diagnostic criteria as possible, and their main purpose is to classify individuals into "healthy" and "ill". The interviews can be classified in two types:

a) *Semi-structured interviews*, interviews with an open and conversationally oriented flow, allowing new ideas to be brought up during the interview, but which include an interview guide or a list of questions or topics that need to be addressed, usually in a particular order. This kind of interview can only be performed by an experienced and reliable clinician (psychiatrist or psychologist).

Among the most frequently used semi-structured clinician administered examination interviews are the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) and the Structured Clinical Interview for DSM (SCID) [40]. This type of assessment tools has been rarely used in population-based epidemiologic studies [27,41],

primarily because of the expense of training and employing clinicians to undertake the assessments, and their main use has been restricted to clinical research on smaller samples of patients [42–48], and to validate the results of the less expensive or shorter techniques [49–53].

b) *Fully structured interviews*, consisting of verbatim interviews where the patient is asked pre-specified questions in fixed sequence and the responses are rated. Depending on the instrument, it may have been developed to be administered by a clinical psychiatrist or psychologist or a lay interviewer, all of them having been specifically trained for this purpose.

Fully structured research diagnostic interviews have experimented great improvements over the past 25 years, initiated with the development of the Diagnostic Interview Schedule (DIS) [69] for its use in the Epidemiological Catchment Area (ECA) in the 1980s [70], which is considered the first survey from the third generation of Psychiatric epidemiology surveys. The DIS was crafted concurrently with the creation of the DSM-III criteria [67] and adopted this diagnostic classification. With time, the DIS has evolved to remain consistent with changes in the DSM, with the current version producing DSM-IV diagnoses [71]. Other widely used fully-structured interviews are the Composite International Diagnostic Interview (CIDI) or the Mini-International Neuropsychiatric Interview (M.I.N.I.) [72], all of them including specific sections to evaluate depressive disorders.

The CIDI was first developed by the WHO as an expansion of the DIS, to incorporate ICD definitions of mental disorders, which was the international standard diagnostic system used and it was translated and field tested in several countries [73] and its first version followed DSM-III criteria. A subsequent review of the instrument included measures of risk factors, consequences, patterns and correlates of treatment and updated the instrument to DSM-IV diagnostic criteria [74].

The fact that fully-structured interviews allow administration by trained lay interviewers supposes a great advantage over clinician-administered interviews in terms of implementation and costs. Among them, the CIDI is growing in relevance as the standard measure of psychopathology in community epidemiological surveys on mental health [60]. In its variety of forms, the CIDI has been the instrument of choice in some of the most recent major nationally representative epidemiologic surveys to gather data on the prevalence and correlates of specific disorders, including the NCS and NCS-R, the NEMESIS and the WMH surveys described previously.

***2.- Self-reported questionnaires:***

While structured instruments provide specific psychiatric diagnoses using standardized criteria, they require highly trained interviewers or clinicians, and have complicated scoring algorithms. Additionally, in-depth assessment requires lengthy interviews, making them very time demanding. Because of such administration burden, their use in large scale surveys is specifically limited to psychiatric epidemiologic studies. Including them in other settings, such as general health interview

surveys where mental health is only one of a number of health indicators assessed would be unfeasible. Using lengthy interviews is also unattainable in clinical applications with stringent time-demands, and thus outside specialized mental health care services.

As an alternative, many short self-reported questionnaires or scales, also known as Patient Reported Outcome (PRO) measures, have been developed aiming at assessing mental health to be answered in a few minutes [76]. The instruments available differ in specificity of their content, from health related quality of life measures that may include one or several dimensions evaluating physical, social and emotional aspects of health; generic measures of mental health state or psychological distress; or specific questionnaires that evaluate particular mental disorders or syndromes. With regard to generic measures, there is difficulty in stablishing a conceptual definition of what is actually being measured, to distinguish for example between "distress" and "disorder", and between "psychological", "emotional", and "mental". Here, we have followed notation by McDowell and Newell [77], that refer to distress rather than disorder, and use the term "psychological" to connote a general level of problems, commonly referring to emotional problems, and at times to mental problems.

Table 1 includes a limited list of most widely used health related quality of life measures or generic questionnaires for the assessment of mental health, mainly exploring its affective spectrum.

***Table 1.*** *Health related Quality of Life and generic psychological distress self-reported questionnaires for mental health assessment*

| Questionnaire | Number of items (Short versions) | Concept measured (dimension of interest) | Type of measure | Year # |
|---|---|---|---|---|
| **Short Form Health survey (SF-36/SF-12)** [78] | 36 (12) | HRQoL* (mental health dimension; mental component summary) | Profile | 1993 |
| **World Health Organization Quality of Life Assessment (WHOQOL-BREF)** [79] | 24 | HRQoL (psychological domain) | Profile | 1998 |
| **EuroQoL 5D (EQ5D)**[80] | 5 | HRQoL (Anxiety/depression dimension; Index) | Index | 1990 |
| **General Health Questionnaire (GHQ)** [81] | 60 (30/28/20/12) | Psychological distress | Questionnaire | 1970 |
| **K10/K6 psychological Distress Scales** [75] | 10 (6) | Psychological distress | Scale | 2002 |

* HRQoL: Health Related Quality of Life
# Year of publication

Some of these generic measures are routinely included in large population health surveys. In a search carried out on the Health Interview Surveys (HIS) database on contents of Health Interview Surveys in Europe between 1998 and 2002 [82], it was found that out of a total of 64 surveys in 18 European countries, 21 included some

sort of questionnaire evaluating mental health. Most often, the surveys measured general mental health or psychological distress, and the most frequent tool was the GHQ-12 scale [82]. Health-related quality of life was evaluated in a similar percentage as general mental health, with the most common measures being some form of the Short Form Health Survey (SF) questionnaires [82].

The General Health Questionnaire (GHQ) [81] was developed in the late 1970s as a 60-item, multidimensional, self-reported screening instrument to detect current, diagnosable psychiatric disorder. It was designed for its use in general population surveys or in primary care, and several versions have been subsequently developed, with 60, 30, 28, 20 and 12 items. In its shortest version, the GHQ has produced results comparable to the longer 28-item version [83].

The Short Form Health Survey [78] is arguably one of the most widely used generic instruments for the assessment of self-perceived health related quality of life both in research and in clinical practice. Several versions of the questionnaire have been developed, the most popular ones being its original 36-item version (SF-36) and the 12-item short form (SF-12). These questionnaires measure health on 8 multi-item dimensions (Physical Functioning, Role Physical, Bodily Pain, General Health, Vitality, Social Functioning, Role Emotional and Mental Health) and two summary measures (Physical and Mental Component Summaries). The Mental Health dimension of the SF-36 is the 5-item Mental Health Inventory, which has shown good performance in tests of sensitivity (SN) and specificity (SP) relative to other screening tools for depression and other mental disorders [84].

Other more recent measures that have been developed for monitoring population prevalence and trends are the K10 and K6 psychological distress scales [75]. These scales have mainly been included in National Health surveys in the US, and in Australia [85,86].

The abbreviated version of the WHO Quality of Life Assessment (WHOQOL-BREF) questionnaire [79] and the EuroQoL (EQ-5D) [80] have also been used for evaluation of mental disorders but always in surveys that also included other generic or specific measures of mental health. The WHOQOL-BREF is a 24-item health related quality of life measure on 4 health domains (Physical Health, Psychological, Social Relationships, Environment) with special focus on its cross-cultural application. The EQ-5D, was originally developed for evaluative studies and policy research that measures generic health status as a single index score [80].

With regard to specific instruments for depression, several scales have been developed; the most relevant ones [76,87,88] are listed in table 2.

Among them, the Center for Epidemiologic Studies Depression (CES-D) [89] was specifically developed in 1977 to assess current level of depressive symptomatology in the general population and it is one of the most widespread brief scales for assessing depression. It consists of 20 items about symptoms that cover the components of depressed mood, feelings of guilt and worthlessness, feelings of helplessness and hopelessness, psychomotor retardation, loss of appetite, and sleep

disorders, occurring in the week prior to the interview, with 4-category response options on frequency.

***Table 2.*** *Specific self-reported questionnaires for depression assessment*

| Questionnaire | Number of items | Characteristics | Purpose | Year[#] |
|---|---|---|---|---|
| **Beck Depression Inventory (BDI-II)** [90] | 21 | DSM-IV depression symptoms | Severity in psychiatric patients/ Screening | 1996 |
| **Center for Epidemiological Studies Depression (CES-D)** [89] | 20 / 10 | Depressive symptomatology | Depression in general population | 1977 |
| **Patient Health Questionnaire (PHQ-9)** [91] | 9 | DSM-IV depression symptoms | Severity in psychiatric patients/ Diagnostic | 1999 |
| **PROMIS Depression** [92] | 28 | Depressive symptomatology | Mood Severity in general population | 2010 |
| **Self-rating Depression Scale Zung (SDS)** [93] | 20 | Depressive symptomatology | Severity in psychiatric patients/ Screening | 1965 |

[#] Year of publication

The Beck Depression Inventory (BDI) [94] and the Zung Self-Rating Depression Scale (SDS) [93] are also widespread instruments that were developed specifically to quantify severity of depression in psychiatric patients with the disorder. They include a similar number of questions

and use response formats that rely either on severity or frequency of the symptoms. The time frames of questions are "today" for the BDI, and "recently" for the SDS. The BDI was originally developed in 1961 and a revised version, BDI-II [90], was implemented in 1996 to match it more closely to the DSM-IV criteria.

The Patient Health Questionnaire (PHQ-9) [91,95] is a more recently developed depression scale that consists of 9 items that correspond to the symptoms identified in the DSM-IV. It was developed for diagnostic and severity measurement for major depressive disorders in clinical settings.

Lately, a self-reported measure of depression has been created as part of the National Institutes of Health-funded Patient Reported Outcomes Measures Information System (PROMIS) [92,96]. PROMIS instruments were developed using IRT, and include different domains from broad health areas (physical, mental, and social) known to be affected by several conditions or treatments, but avoiding items to carry condition-specific attributions. One of these domains is the PROMIS Depression domain that offers several advantages over the CTT approach used in developing the other instruments detailed here, such as the CES-D and the PHQ-9. In particular, the use of IRT models result in a common metric for different populations allowing for greater comparability across disorders and other PROMIS domains, a broader range of scores, greater precision in individual measures, and greater flexibility in test administration [97]. For example, items can be administered as static short forms or computerized adaptive tests (CATs), which select the best items to

sharpen the estimate of a person's status, based on prior responses to earlier questions. In general, experience with CAT suggests that the PROMIS Depression item bank provides excellent precision with as few as 4-6 items [98].

## 1.3. Using self-reported questionnaires for screening and severity assessment

Screening is of central importance in many health-related fields, including psychopathology [99]. Its main purpose is to identify in a large group of people, those who have an elevated probability of suffering from the disorder under study in order to target treatment and/or prevention. The US Preventive Services Task Force (USPSTF) has recently reaffirmed its recommendation to screen for depression in the general adult population [100], where there are adequate systems in place to ensure accurate diagnosis, effective treatment, and appropriate follow-up. The recommendation is based on sufficient evidence indicating that depression screening in the adult population [101] or pregnant and post-partum women [102], in combination with adequate support systems, improves clinical outcomes, and that the magnitude of harm of screening for depression in adults is small or non-existent.

At the population level, screening can help identify groups of individuals who may be at risk for depression, so that interventions aimed at primary prevention (i.e. the reduction of the onset of major depression) or at secondary prevention through the reduction of unmet need for help in the presence of the first symptoms, can be

implemented in those groups most likely to benefit from preventive interventions [103].

Most of the self-reported questionnaires described above were not designed for diagnostic purposes of clinical psychopathology, or depressive disorder, but to provide a numerical indication of the severity of the distress within a given period of time [104]. However, because of the economy and ease of administration, they have been used in many large community health surveys in preference to structured diagnostic instruments as dichotomous measures indicating whether a specific cut-off point value or threshold is reached, allowing to identify probable ill cases according to established categorical standards [82]. They are also commonly used as screeners in two-phase surveys where the self-reported questionnaire is administered to all respondents in the first phase and the second phase consists of an in-depth evaluation with a clinical interview to those screening positive and a random sample of negatives.

On the other hand, it is important to take into account that diagnostic interviews follow classificatory systems with a categorical classification approach, where a diagnosis is based on fulfillment of an arbitrary number of signs and symptoms agreed through expert consensus. According to this system, the specific symptom pattern is not as important as the number of symptoms met by the individual. Also, individuals are classified into discrete categories (healthy or ill) [105,106] assuming that health and sickness are two different entities. The latter contradicts the evidence that severity of the disorder is associated to its course [107–109].

Discrete boundaries within psychiatric diagnosis have several limitations in terms of validity [110]. When allowed to choose between definite depression, possible (subthreshold) depression, and unclear cases, clinicians rate over a third of their decisions as not definitive. Many errors are due to incorrect estimates of severity, for example when symptoms are diagnosed but judged clinically insignificant [111], proving the complexity of fitting the continuous variation in depression severity into a categorical diagnosis. By contrast, a dimensional system acknowledges that there may be clinically important individual differences among those who fall below and among those who fall above a categorical diagnostic threshold [112]. To take an extreme example, if the diagnostic criteria require duration of at least 28 days, the patient whose illness manifestations lasted 27 days is considered equivalent to someone who has never had that illness, but completely different from someone whose duration was 29 days. In turn, the latter patient is considered equivalent to someone who has experienced the signs and symptoms for years. In clinical application, this often translates into treating patients with minimal need, or denying treatment (or prevention) to patients who clearly need it [112].

The skip patterns implemented in most relevant structured diagnostic interviews, such as the CIDI or the MINI, that guide through the diagnostic criteria, further difficult the assessment or identification of subthreshold individuals that might benefit from specific prevention strategies. Moreover, it has been found that the burden of major depression follows the gradient of symptom severity, and assessing it

for subtypes of depression and according to this gradient may help prioritize treatment allocation [113]. All of this indicating the complementary value of a dimensional assessment of depression, as provided by self-reported measures, to the categorical diagnoses [112].

# 2. THESIS RATIONALE

Regardless of the relevance of mental disorders from the public health perspective they have been, until recently, neglected by the global health agenda [114]. WHO and other United Nations agencies, such as UNICEF [115–117] have developed standardized surveillance systems which produce a comparable set of indicators covering a number of major causes of disease burden. However, a similar systematic approach to measurement and data collection is still needed for mental disorders, for which standard indicators should be defined [114] in order to facilitate comparability over time, and between studies and countries. Such comparability enhances the usefulness of data. In particular, health policy will benefit from the understanding of the relative health status in one country compared to others. Currently, limited comparability across national health surveys is the norm, with a great variety of self-reported measures being used for the assessment of mental health in comprehensive national health surveys [82].

The most adequate measures should be identified based on the goals of assessment and on the metric properties of the instrument in relation to those goals. Measures need to be *reliable*, which refers the extent to which an instrument is free from random error; and *valid*, referring to the extent to which the instrument or score measures what it is supposed to measure. Different aspects of validity need to be considered: a) Content validity that assesses evidence about the extent to which items and domains are appropriate for their intended use; and b) Construct validity that evaluates theoretical implications associated with the construct, such as logical relations with other

measures or across groups. In particular, when used as screening measures, the performance of a scale is best determined by its ability to correctly identify those with (sensitivity) and those without (specificity) the target diagnosis according to the selected standardized criteria. This is known as diagnostic test accuracy in the clinical epidemiology field [118], although in other contexts it is also referred to as criterion-related validity [119,120]. In this thesis, we will use the term diagnostic accuracy when we refer to the ability to discriminate between pathological and non-pathological cases. The selection of an appropriate cut-off score is dependent upon which of these two characteristics is considered most important; that is, how the screening outcome will be used, and the consequences of correct and incorrect identification [121]. On the other hand, if the purpose is to evaluate treatment outcomes or population monitoring, we need the scale to be responsive to change, i.e. sensitive enough to detect changes in the clinical course of the pathology.

Among the available self-reported measures specific for depression, the CES-D is the only one that was explicitly developed for use in general population surveys. A simple search in bibliographic databases retrieves thousands of population-based and clinical studies using the CES-D in the last ten years. A number of these studies have evaluated its psychometric properties [122] and the accuracy of the instrument to detect major depression at the general population and primary care levels. The score of 16 was adopted as the cut-off point recommended for depression caseness shortly after the development of the scale in a somewhat arbitrary manner [123], and it has been extensively used ever since even though some other studies have claimed that this cut-

off point provides too low a specificity and that a higher threshold would be more appropriate [124]. In spite of its prominence, no work has been done to date to integrate the results on performance of the CES-D through meta-analysis nor to formally test the different cut-off points, which would provide precise and generalizable evidence about the performance of the CES-D and the interpretation of its results, and would allow establishing whether and how associated findings vary by particular subgroups.

As mentioned above, health related quality of life measures usually covering physical, emotional and social aspects of health, are also commonly included in health surveys and epidemiologic research to evaluate self-perceived health. Among them, the SF measures (SF-36 or SF-12) [78] are the most widely used and they have been extensively evaluated cross-culturally [125–129]. Previous studies have shown that the mental health dimensions of the SF-36 (i.e. the Mental Health Inventory, MHI) and the mental component summary, are both able to detect depression cases in community and patients samples with good results [78,130–133]. Three of the five items of the MHI are still included in the SF-12 and refer to symptoms related to the diagnostic criteria for common depressive and anxiety disorders. Furthermore, the questionnaire includes other items regarding functional impairment due to mental problems that are also related to experiences of distress or impairment. Thus, given its content and wide use, the mental component of the SF-12 could serve as a depression screener, or for monitoring prevalence and targeting treatment and prevention. However, the performance of the SF-12 for the assessment of mental disorders in the general population, has

hardly been studied [134]. This study is further complicated by the fact that standard scoring algorithms of the SF-12 are based on the prediction of SF-36 scores, and not in a psychometric model of its own. Given the extensive use of the SF-12, a psychometric model for analyzing its reliability and an analysis of threshold scores may provide both an alternative way of interpreting its score and information about the practical use of the scale as a screening instrument.

A new generation of health outcome instruments is being developed based on the principles of IRT, a general statistical approach for the design, analysis and scoring of questionnaires that model the probability of response of each examinee of a given ability to each item in the test. Provided IRT assumptions hold, this methodology offers many advantages over the CTT approach, used for the development of most self-reported measures so far: a) IRT modeling maximizes precision of the instruments; b) the resulting parameters are not sample- or test-dependent, thus comparability of scores across groups and across test forms is guaranteed. In relation to this, under IRT, the items themselves are characterized and thus, if the model fits, the items always measure the same trait at a specific level because item responses are linked to an ability level. This is not the case for CTT. This invariance property of IRT allow for an appealing characteristic of this approach: the fact that individual items from the item-banks can be extracted to create reliable, valid, comparable, and precise short forms of that specific domain, minimizing respondents' burden. The short forms can either be static, or created using computer adaptive testing (CAT), in which the presentation of the items is tailored individually to respondents using computer algorithms that maximize

information about the person's likely score [135,136]. The algorithm administers a test to a patient one question at a time. At each step the patient's prior responses determine whether to ask another question and, if so, which question to ask. The test is stopped when the patient's score has been estimated to a prescribed level of precision or to a pre-specified number of items. Hence, the computer adapts the test to use the fewest number of items required to assess a particular level of severity with the demanded precision.

Several item banks for CAT administration of depression have been developed [9,10,12]. Among them, the PROMIS system [96] includes a depression domain as part of the overall health profile; and it is also the only IRT-based depression measure available in Spanish and some other languages. Because of its item development, firmly based on content validity considerations, and application of IRT methodology, PROMIS stands out as a promising measure that could serve to evaluate prevalence but also to assess a wide range of severity levels, including subthreshold, which is difficult to assess from a diagnostic interview.

PROMIS Depression measures have shown to have good metric properties across US clinical samples [15]. Even though they have been adapted into different languages [18,19], no studies so far have assessed measurement equivalence of the adapted versions with respect to the US version nor their validity, including responsiveness to change and diagnostic accuracy, which is crucial to ensure comparability and gain evidence about the usefulness of PROMIS measures for screening and monitoring of depression.

# 3. OBJECTIVES

The global aim of this dissertation is to provide evidence regarding the construct validity, particularly diagnostic accuracy, of currently available specific and generic self-reported questionnaires to assess presence and severity of depression in the general population. This is done from different psychometric approaches, Item Response Theory and Classical Test Theory.

Specific objectives:

1. To evaluate the diagnostic accuracy of the Short Form 12 health survey (SF-12 v1) for the assessment of depression in the general population using different metric approaches. This is achieved in two sub-objectives:
    a. First, an instrumental objective to develop an Item Response Theory metric for version 1 of the SF-12.
    b. Secondly, a comparative objective where different SF-12 scoring systems, based on Item Response Theory and Classical Test Theory, are compared in a general population sample.

2. To comprehensively review evidence on the accuracy of the Center for Epidemiologic Studies Depression (CES-D) scale for assessment of depression in the general population and in primary care settings.

3. To evaluate the metric properties, particularly measurement equivalence between Spanish and English versions and diagnostic

accuracy of the PROMIS Depression measures for the assessment of depression in the general population of Spain.

4. To assess construct validity and responsiveness to change of PROMIS Depression measures in a clinical sample with mental health problems of Spain.

## 3.1. Hypotheses

1. Performance of the different scoring systems of the SF12v1 for screening depression caseness in the general population will be adequate (AUC over 0.80). A cut-off point will be provided with both sensitivity and specificity over 0.80.

2. Diagnostic accuracy of the CES-D as a screening measure for major depression in the general population will be adequate (AUC over 0.80), with the usually recommended cut-off point of 16 providing low specificity value in comparison with sensitivity.

3. PROMIS Depression measures will be valid and reliable, and will fulfill assumptions required for IRT calibration. Measurement equivalence will be found between Spanish and English versions of the instrument allowing the use of a common metric for both populations.

4. PROMIS Depression measures will show good diagnostic accuracy for detecting major depression in clinical settings, and high responsiveness to change.

# 4. PUBLICATIONS

This dissertation is presented as a compendium of five original publications which are briefly described below:

The first objective was accomplished in two different studies. We first developed a scoring method based on multidimensional IRT methodology for the SF-12 that was published in the *Journal of Clinical Epidemiology* (2013).

> Forero CG, Vilagut G, Adroher ND, Alonso J. Multidimensional item response theory models yielded good fit and reliable scores for the Short Form 12 questionnaire. J Clin Epidemiol. 2013;66(7):790–801.

Afterwards, we evaluated performance of different SF-12 scoring systems for screening purposes, using the CIDI as the gold standard. Results of this work were published in *Value in Health* in 2013.

> Vilagut G, Forero CG, Pinto-Meza A, Haro JM, De Graaf R, Bruffaerts R, Kovess V, de Girolamo G, Matschinger H, Ferrer M, Alonso J, on behalf of the ESEMeD Investigators. The mental component of the Short-Form 12 Health Survey (SF-12) as a measure of depressive disorders in the general population: Results with three alternative scoring methods. Value Heal. 2013;16(4):564–73.

Both articles have been included in first and second place of this dissertation, respectively.

To accomplish the second objective, we evaluated and provided pooled estimates of the screening properties of the Center for Epidemiologic Studies Depression (CES-D), which was specifically created for the evaluation of depressive symptomatology in the general population and is one of the most widely used depression questionnaires. To do so, a systematic review with meta-analysis of the CES-D was conducted and presented in the third of the articles that was published in *PLoS ONE (2016)*:

> Vilagut G, Forero CG, Barbaglia G, Alonso J. Screening for Depression in the General Population with the Center for Epidemiologic Studies Depression (CES-D): A Systematic Review with Meta-Analysis. PLoS One. 2016;11(5): e0155431

For the achievement of the third objective of this thesis, the performance of PROMIS Depression measures, which are based on IRT methodology, was evaluated in a general population sample. This work has been submitted to *Journal of Clinical Epidemiology* and is presented as the fourth article of this thesis.

> Vilagut G, Forero CG, Castro-Rodriguez JI, Olariu E, Barbaglia G, Astals M, Diez-Aja C, Gárriz M, Abellanas L, Alonso J, on behalf of the PROMIS.es Investigators. PROMIS Depression Item Bank Showed Measurement Equivalence between Spain and the US. J Clin Epidemiol (under review)

Finally, for the fourth objective of this dissertation, PROMIS Depression measures were assessed in a clinical sample, with special focus on its construct validity and responsiveness. This is the fifth

article of this dissertation and was published in the *Journal of Psychiatric Research* (2015):

> Vilagut G, Forero CG, Adroher ND, Olariu E, Cella D, Alonso J, on behalf of the INSAyD investigators. Testing the PROMIS® Depression measures for monitoring depression in a clinical sample outside the US. J Psychiatr Res. 2015; 68:140–50.

Each of these articles is subsequently presented.

## 4.1. Article 1: *"Multidimensional item response theory models yielded good fit and reliable scores for the Short Form-12 questionnaire"*

Forero CG, Vilagut G, Adroher ND, Alonso J. Multidimensional item response theory models yielded good fit and reliable scores for the Short Form 12 questionnaire. J Clin Epidemiol. 2013;66(7):790–801.

PMID: 23707080

## 4.2. Article 2: "*The Mental Component of the Short-Form 12 Health Survey (SF-12) as a Measure of Depressive Disorders in the General Population: Results with Three Alternative Scoring Methods*"

Vilagut G, Forero CG, Pinto-Meza A, Haro JM, de Graaf R, Bruffaerts R, et al. The Mental Component of the Short-Form 12 Health Survey (SF-12) as a Measure of Depressive Disorders in the General Population: Results with Three Alternative Scoring Methods. Value Heal. 2013 Jun;16(4):564–73. DOI: 10.1016/j.jval.2013.01.006

## 1.1. Article 3: *"Screening for Depression in the General Population with the Center for Epidemiologic Studies Depression (CES-D): A Systematic Review with Meta-Analysis"*

Vilagut G, Forero CG, Barbaglia G, Alonso J. Screening for Depression in the General Population with the Center for Epidemiologic Studies Depression (CES-D): A Systematic Review with Meta-Analysis. PLoS One. 2016;11(5): e0155431.

PMID: 27182821

Supplementary material for this article can be found in ANNEX 1 (page 177)

## 4.3. Article 4: *"PROMIS Depression Item Bank Showed Measurement Equivalence between Spain and the US"*

Vilagut G, Forero CG, Castro-Rodriguez JI, Olariu E, Barbaglia G, Astals M, Diez-Aja C, Gárriz M, Abellanas L, Alonso J, on behalf of the PROMIS.es Investigators. PROMIS Depression Item Bank Showed Measurement Equivalence between Spain and the US. J Clin Epidemiol (under review)

Supplementary material for this article can be found in ANNEX 2 (page 185)

# PROMIS Depression Item Bank Showed Measurement Equivalence between Spain and the US

Vilagut G[a,b,c], Forero CG[a,b,c], Castro-Rodriguez JI[d], Olariu E[a,$], Barbaglia G[a,#], Astals M[d], Diez-Aja C[d], Gárriz M[d], Abellanas L[d], Alonso J[a,b,c], on behalf of the PROMIS.es Investigators[1]

[a] Health Services Research Unit, IMIM (Institut Hospital del Mar d'Investigacions Mèdiques), Carrer del Doctor Aiguader, 88, Edifici PRBB, 08003, Barcelona, Spain.

[b] CIBER Epidemiología y Salud Pública (CIBERESP), Spain.

[c] Universitat Pompeu Fabra (UPF), Plaça de la Mercè, 10-12. 08002, Barcelona, Spain.

[d] Institute of Neuropsychiatry and Addictions (INAD), Parc de Salut Mar, Barcelona, Spain

[$]Current address: PHMR consulting, London, UK

#Current address: Assessment department, Agency for Health Quality and Assessment of Catalonia, Barcelona, Spain

[1] **PROMIS.es Investigators**: Gemma Vilagut, Carlos G. Forero, Jordi Alonso, Mònica Astals, Gabriela Barbaglia, Jose-Ignacio Castro-Rodriguez, Cristobal Diez-Aja, Miguel Gárriz, Elena Olariu, Adelina Abellanas, Jacobo Chamorro, Jose Manuel López-Santín, Carmen Sánchez-Gil

**Correspondence autor:**

Carlos G. Forero
Health Services Research Unit, IMIM (Institut Hospital del Mar d'Investigacions Mèdiques),
Carrer del Doctor Aiguader, 88, Edifici PRBB, 08003, Barcelona, Spain.
Phone: (+34) 933160 760; Fax: (+34) 933 160 797. E-mail: cgarcia@imim.es

**ABSTRACT**

**Objective:** Assessing measurement equivalence PROMIS Depression item bank in Spanish and the US, and validating it for its use in Spain.

**Study design and setting:** Cross-sectional study with Spanish adult general population sample (n=1,503). We tested measurement invariance checking IRT assumptions of unidimensionality (Confirmatory Factor Analysis goodness-of-fit with Comparative Fit Index, CFI>0.95, Root Mean Square Error of Approximation, RMSEA <0.08), local-independence (Principal Components and size of residual correlations), monotonicity and scalability (Mokken scaling). Differential Item Functioning (DIF) by age, sex, education and country were assessed using ordinal logistic regression (McFadden's pseudo $R^2$ change>0.02). We assessed reliability throughout the continuum based on test information (1-(1/Information)) and construct validity correlating PROMIS with legacy measures of depression, anxiety and disability.

**Results:** IRT assumptions held in Spain: unidimensionality (CFI=0.98, RMSEA=0.06), local independence (residuals<0.2), monotone homogeneity (coefficient H=0.7). One item showed DIF between the US and Spain, with minimal impact on the overall score. Score distribution using the US and Spanish-specific parameters were similar, with information values equivalent to reliabilities over 0.90 from -1 to +4 SDs around the population mean. Expected correlation pattern was found with legacy measures.

**Conclusions:** Measurement equivalence and good metric properties of the Spanish PROMIS Depression provide evidence of its adequacy for comparative research.

KEYWORDS: PROMIS Depression, measurement equivalence, differential item functioning, item response theory

## What is new?

Key Findings:

- PROMIS depression item bank has similar and excellent psychometric properties in the US and Spain. A single IRT parameter calibration set would suffice for efficiently and accurately measure depression in both populations.

What this adds to what is known:

- This is the first study that supports the metric invariance of the PROMIS item bank in two countries and languages.
- The robustness of the PROMIS Depression IRT model makes it an ideal instrument for comparative research
- PROMIS Depression show good accuracy for detecting depression

What is the implication, what should change now:

- There is need to develop substantive score interpretation strategies that avoid resorting to population anchors.

**[Word count: 4,386]**

## 1. Introduction

Patient-reported outcome (PRO) measures assess patients' views on their own health [1], an increasingly important perspective for research, clinical practice and policy-making [2,3]. To overcome caveats about the lack of comparability among PRO measures, the National Institutes of Health (NIH) called for psychometrically sound PRO measures, generalizable across populations (general or clinical) and conditions (chronic or acute).

In response to this request, the Patient-Reported Outcomes Measurement Information System (PROMIS®) [4] made an unprecedented effort to develop new PRO measures with emphasis on the precision and comparability of scores. This goal was achieved by applying a domain-specific approach on broad health areas (physical, mental, and social) that might be affected by conditions or treatments, but avoiding items to carry condition-specific attributions; and by using state-of- the-art item response theory (IRT) methods.

An especially important advantage of PROMIS, derived from the IRT methodology, is the measurement invariance property that allows direct comparisons among scores at group or individual level regardless of the population [5]. This has particular relevance in the case of international studies on comparative research [6,7], as verification that IRT assumptions are similarly met in different versions permit comparable item-bank calibrations [8]. Also, IRT facilitates evaluation of measurement invariance at the item level using differential item functioning (DIF) technics, which examine whether or not the likelihood of item (category) endorsement is equal across subgroups that are matched on the level of the trait being measured. Also, the impact of DIF at the test level (Differential Test Functioning, DTF) can be easily appraised. Another important advantage of PROMIS coming from IRT invariance property is the feasibility of a versatile administration that allows static short form versions or Computer Adaptive Test (CAT) administration yielding the same score for a given individual [9]. In spite of these advantages, existing quantitative approaches for testing measurement equivalence and

ruling out testing bias have not been extensively used in health measurement research in spite of being routinely used in the main cross-national educational studies [10,11]. PROMIS offers an excellent opportunity for obtaining internationally-equivalent PRO measures of health with IRT in different languages and cultures [12].

Depression is one of the most prominent domains of the 8 core domains composing the PROMIS health profile and therefore it is key for international comparisons. Depression impairs the course of many health problems and is a crucial aspect in assessing treatment effectiveness of these problems [13,14]. PROMIS Depression measures have proven good metric properties across US clinical samples [15] and comparable results to other measures such as the Patient health questionnaire (PHQ-9) or Beck Depression Inventory (BDI) [16,17]. They have been adapted into different languages [18,19], but no studies so far have assessed measurement equivalence of the adapted versions with the original US version.

The main objective of the current study was to evaluate cross-cultural validity and measurement equivalence of Spanish PROMIS depression measures in Spain as compared to original US item bank. Specifically, we aimed at: a) testing IRT assumptions and calibrating the Spanish version of PROMIS Depression item bank using Spanish general population data; b) evaluating differential item functioning (DIF) of the Spanish version by age, sex and education, and cross-cultural DIF comparing the Spanish version of PROMIS Depression with the original English version; c) assessing reliability and construct and criterion validity of the PROMIS Depression measures, including performance of the static short forms and simulated CATs.

## 2. Methods

An observational cross-sectional study using online Panel data was carried out in May 2015. Panelists were invited by a panel vendor company achieving a final sample with equivalent distribution to that of the Spanish general population in terms of age, sex and region groups. A total of 1,800 individuals completed the online questionnaire, administered using a block-design to minimize respondent burden. Each respondent was administered a maximum of 120 items,

maintaining at least 300 concurrent evaluations between PROMIS Depression item bank and any other instrument (see supplementary table 1). Socio-demographics were administered to all respondents. The study protocol was approved by the Institutional Review Board (IRB) at Parc de Salut Mar, Barcelona, Spain. Participants provided informed consent before entering the on-line survey, and confidentiality was ensured by avoiding delivery of personal information to the project investigators.

## 2.1. Study variables

### 2.1.1. PROMIS Depression item bank

PROMIS Depression item bank has 28 items focusing on negative mood, decrease in positive affect, information processing deficits, and negative self-image and social cognition [20]. Items have a 7-day recall period and 5-level frequency Likert type responses (1=Never; 2=rarely; 3 =Sometimes; 4= Often; 5=Always). Spanish versions of each item were obtained following the PROMIS standardized "universal" translation protocol [21], aiming at establishing just one language version for multiple countries instead of country–specific language versions [12]. The protocol includes expert reviews and cognitive debriefing interviews on target populations for each domain. A specific cognitive debriefing was conducted in Spain for guaranteeing language harmonization. The Spanish version was found conceptually equivalent for Spanish individuals. The PROMIS Depression item bank was applied along with the Anxiety (29 items) and Anger (22 items) Items banks, which had similar item logic in development and structure [20].

### 2.1.2. Other measures

*The Center for Epidemiologic Studies Depression (CES-D) scale*: a 20-item self-report scale that measures the level of depressive symptomatology in the general population [22] with 7-day recall period. Score ranges between 0 (best) to 60 (worst). A 16-point cut-off has been recommended for active depression with good properties for screening purposes [23].

*Patient Health Questionnaire 9-item (PHQ-9) scale*: a 9-item clinical symptom severity scale that assesses depression in the previous two weeks. Items have a 4-point Likert format that can

be summed up to obtain a severity score. A diagnostic algorithm for Major Depressive Syndrome (PHQ-9 MDS criteria) is also available: "Positive" cases are those individuals that have at least five items answered as "more than half the days" or "nearly every day", except for suicidal ideation which is counted as positive whenever the response is different than "never". At least one of the symptoms has to be either item 1 (Little interest or pleasure in doing things) or item 2 (Feeling down, depressed, or hopeless) [24].

*Generalized Anxiety Disorder 7-item scale (GAD-7)*: it is a 7-item scale for assessing the presence and severity of GAD. Items are summed to obtain a 0 to 21 severity score; higher values indicate higher anxiety levels. Cut-points of 5, 10, and 15 represent mild, moderate, and severe levels of anxiety [25].

*Beck Anxiety Inventory (BAI)*: A 21-item questionnaire for measuring anxiety severity over the previous week. The overall score ranges from 0 to 63, higher values representing higher anxiety [26].

*World Health Organization Disability Assessment Schedule 2.0, 12-item version (WHODAS 2.0):* it measures the degree of functional limitations in the previous 30 days, irrespective of medical diagnoses. Scores range from 0 (none) to 100 (highest) disability [27].

In addition, the following socio-demographic variables were assessed: age, sex, marital status, educational level and employment status.

## 2.2. Statistical methods

This study follows the analysis plan specified by PROMIS [8], assessing item bank metric properties using Item Response Theory (IRT) and classical test theory methods.

### 2.2.1. IRT Assumptions

IRT unidimensionality, local independence and monotonicity assumptions were evaluated before calibration. Unidimensionality was tested through Confirmatory Item Factor Analysis (CIFA) specifying a one-factor solution. Models were fitted on the polychoric correlations to take into account ordinal items, using Unweighted Least Squares estimator and Mean and

Variance corrections to obtain robust p-values and standard errors. Goodness of fit was examined with the Comparative Fit Index (CFI>0.95 for good fit), Tucker-Lewis index (TLI > 0.95 for good fit), and Root Mean Square Error of Approximation (RMSEA<0.06 for good fit, and 0.08 acceptable fit) [28].

Local independence assumption (i.e., absence of associations among items conditional on the latent trait), was assessed: a) checking for residual correlations over 0.2 in the one-factor CFA; b) visual inspection of the scree-plot of eigenvalues from a Principal Components Analysis (PCA) on the residual correlations [29]; and c) checking for the presence of PCA components with eigenvalues over 2.

Monotone homogeneity indicates the probability of endorsing more severe responses on an item follows an increasing monotonic function of the latent trait. This assumption was studied through visual evaluation of non-parametric IRT response curves, using Mokken scaling.

2.2.2. Item calibration

Item calibration was carried out using a logistic Graded Response Model (GRM) [30], obtaining IRT Expected a Posteriori (EAP) person theta scores. Goodness of fit was evaluated with the S-$X^2$ statistic, p-values below 0.001 were considered to have poor fit [31].

We also assessed PROMIS Depression short forms of 4 (version 4a), 6 (version 6a) and 8 items (versions 8a and 8b), deriving scores from item bank responses. Computer Adaptive Test (CAT) simulations were conducted using Expected a Posteriori (EAP) Theta estimation. Following same methods as first generation PROMIS CAT engine [32], we used Maximum Posterior Weighted Information item selection and the standard deviation of the posterior distribution as the standard error estimator. Two different stopping rules were used: a) standard error of measurement lower or equal to 0.32 or maximum number of items=12; and b) exactly 8 items.

2.2.3. DIF analysis and score metric

Measurement equivalence across groups was assessed as uniform (constant across theta) and non-uniform (varying across theta) differential item functioning (DIF) through ordinal logistic

regression models, conditioning on IRT Theta estimates [33]. Criteria for DIF were McFadden's pseudo R2 change>0.02 or relative change in beta >0.05. DIF was assessed for age (below and above the sample mean age of 48), sex and education groups (secondary or less versus the rest).

Central to this study was the evaluation of Cross-cultural DIF (US vs. Spain) using data in English from Wave 1 PROMIS calibration studies [34] in order to ensure the feasibility of using a common metric in Spain and US. As an additional comparison between metrics, we obtained two sets of scores: a) scores in Spain-specific metric (Spanish calibration parameters and T-score transformation with mean 50 and standard deviation of 10 in Spain); and b) scores in US metric (US item parameters [32] and T-score transformation with mean 50 and standard deviation of 10 in the US general population). The empirical distributions of both scores were plotted together in order to evaluate differences in distribution shapes.

In absence of DIF, the US metric for scores is used to analyse reliability and validity so that comparability is ensured.

2.2.4. Reliability

Reliability of PROMIS Depression measures and external instruments was assessed as internal consistency reliability using Guttman's Lambda 2 and Cronbach's alpha. Marginal reliability [35] was computed for the underlying PROMIS latent trait of the item bank. Test information was obtained for the item bank and short forms, and reliability throughout the latent trait level was estimated as 1-(1/information(θ)), where θ is the theta trait level.

2.2.5. Validity

Convergent and discriminant validity of the PROMIS Depression measures were evaluated using multitrait-multimethod matrix (MTMM) with external measures of depression, PHQ-9 and CES-D; and related traits of anxiety (BAI and GAD-7), and disability (WHODAS 2.0). From the expected MTMM structure, we hypothesized: PROMIS item bank would show the highest correlations with short forms and CATs, followed by external measures of the same trait (i.e. CES-D and PHQ-9). Lowest correlations would be yielded with different traits and non-

PROMIS questionnaires (i.e. BAI, GAD-7 and WHODAS-12). Intermediate correlation values are expected between different traits measured by the same method (i.e. PROMIS Depression, Anxiety and Anger; and between GAD-7 and PHQ-9). Concordance between different PROMIS Depression forms was estimated with mixed-effects Intraclass Correlation Coefficient (ICC) for absolute agreement [36].

Criterion validity of the PROMIS Depression was assessed using Receiver Operating Characteristic (ROC) curve analysis, using PHQ-9 diagnosis as gold-standard, non-parametric Area Under the Curve (AUC) and corresponding 95% confidence intervals were computed with values ranging from 0.5 (chance level performance) to 1.0 (perfect discrimination). The best screening cut-off point was selected on the basis of the empirical Youden Index, that is, the observed value that maximizes Sensitivity + Specificity – 1.

Statistical analyses were carried out with SAS™, version 9 [37] and R [38], including MIRT , lambda4 package Mokken and lordif packages. CIFA and IRT calibrations were estimated with Mplus version 7.0 [39]. CAT simulations were conducted with FireStar [40].

## 3. Results

The analyses are based on responses from the 1,503 individuals who were administered the PROMIS Depression item bank. Sample mean age was 48.5 (SD=14.8), and 52.6% of the participants were females, Over 75% of the sample had completed secondary education or more (see appendix table 2).

The unidimensional model showed adequate fit (CFI=0.98, TLI= 0.98) and RMSEA=0.065 (90% CI: 0.062-0.067). All residual correlations were lower than 0.15, and the PCA eigenvalues on the residual correlation matrix were lower than 1.53, with each component below 6% variance. Both results supported local independence (scree-plot in supplementary figure 1). Total scalability coefficient H was 0.7 (SE=0.012), within the range of monotone homogeneity and strong scalability (H>0.5). Item scalability coefficients were H over 0.3 (from 0.61 to 0.74)

and visual inspection of non-parametric item response curves also indicated monotonicity (results available upon request).

**Table 1** shows IRT parameters from Spanish calibrations. Item thresholds ranged from -0.48 to 3.82. Discrimination went from a low a=1.91 (EDDEP30 "*I had trouble making decisions*") to a high a=3.9 (EDDEP06. "*I felt helpless*"). No items showed misfit regarding S-$X^2$ criteria.

— INSERT TABLE 1, HERE —

In the assessment of DIF for language, only the item "*I felt disappointed in myself*" (EDDEP26) was flagged for uniform DIF, with $R^2$ change between model 1 and model 2 of 0.032 and $\beta_1$ relative change of 0.068 (see supplementary table 2). As shown in **figure 1**, accrued DIF had very little effect on the overall score, with score differences between initial and purified tests lower than 0.06, substantially smaller than the value 0.2, which would correspond to a small effect size. No DIF was observed for age, sex and education.

— INSERT FIGURE 1, HERE —

**Figure 2** depicts density functions of item bank scores obtained using US and Spanish-specific T-score metrics. Both density functions are similar, although US standard T-scores are slightly moved to the left by 1.2 points with a mean value of 48.8 as compared to a mean of 50.0 for the Spanish-specific scores. As previously mentioned, there is no signal of differential test functioning, indicate that displacement of distributions is due only to the use of different calibration samples. Given that IRT assumptions are met and calibration in both samples and homogeneous, a metric using a single set of parameters for both samples is justified. Standard US metric was then used, thus allowing for universal comparisons for subsequent analyses.

— INSERT FIGURE 2, HERE —

PROMIS item bank showed high internal consistency (Cronbach's alpha = 0.98 and Guttman's Lambda 2 = 0.98) and marginal reliability 0.92. Item bank information yielded reliabilities over 0.90 from theta -1 to almost 3.8 (see **figure 3**), with a maximum at $\theta$=0.60 (i.e., 56.0 in T-score: Information=76.8). This is equivalent to a standard error of measurement of 0.11 and reliability

0.99. Short form reliability went over 0.90 in z-score values around -0.5 to 3.0. Static short version 8a provided slightly higher information values throughout the continuum than version 8b, except at its lower end. Regarding the CAT administration, the stopping rule of a standard error of 0.32 required administering an average of 5.47 items (median=3). At a fixed stop to 8 items, the average standard error of measurement was 0.27.

— INSERT FIGURE 3, HERE —

As hypothesized by the MTMM, highest correlations were observed among PROMIS Depression forms, closely followed by those with scales measuring the same trait (PHQ-9 r=0.74; CES-D r= 0.73). PROMIS Depression forms showed moderate correlations with other traits, including anxiety (around 0.55 with BAI) and disability (around 0.65 with WHODAS-12). However, a substantial correlation was observed between PROMIS Depression and PROMIS Anxiety (r= 0.83), and GAD-7 (r=0.73). Similar results were obtained for legacy measures, with correlations between PHQ-9 and GAD-7 of 0.76 (see **table 2**). ICCs between PROMIS Depression item bank and other PROMIS Depression forms ranged from 0.90 with the 4-item short form to more than 0.95 with 8 items short forms and simulated CATs. When static short measures are compared graphically to the item bank, higher dispersion at the lower end (less severity) of the scale is observed (see supplementary figure 2). Importantly, the 4-item version presents a floor effect of 43%, while for the 8-item version is 25.8%. For CAT administrations, the floor effects were less than 17% and for the item bank it was 12.8%.

— INSERT TABLE 2, HERE —

**Figure 4** displays ROC Curves for PROMIS Depression measures for depression diagnosis (PHQ-9 MDS criteria). AUCs were 0.95 for the item bank and fixed short forms, and 0.94 for the CAT versions. The selected cut off point based on the Youden index for the PROMIS item bank was 57.8 (Sensitivity = 93.7 and Specificity = 89.0), slightly over the point of maximum information. The test information at this point was 76.1 (SEM =0.114).

— INSERT FIGURE 4, HERE —

## 4. Discussion

Since its public release, PROMIS developers encourage researchers to further validate the item banks across diverse populations in order to ensure its adequacy and comparability across settings [34]. In this study, based on comprehensive analyses of a sample with over 1,500 individuals from the general population in Spain, we have established that Spanish PROMIS Depression measures fulfill IRT assumptions, as well as good reliability and construct validity of all forms, complementing previous work on its construct validity and responsiveness of PROMIS in a Spanish clinical sample with common mental disorders [41]. Importantly, this is one of the first studies to assess cross-cultural and measurement equivalence of PROMIS Depression item bank, short forms and CAT in a population outside the US.

The availability of equivalent PRO measures in different languages is of special importance for cross-national comparisons or in healthcare settings that serve populations with diverse language preferences [42]. However, DIF testing is essential to study potential validity challenges according to ethnicity and language [43]. In this work, we found that only one item was flagged for language DIF, causing a negligible effect at the test level (with differences on the scores corresponding to an almost null effect size). This result represents original evidence of the quality and equivalence of the PROMIS "universal" translation approach. . Additionally, no evidence of DIF was observed according to other key variables within the Spanish sample, including age, sex, and education.

Our results should be interpreted taking into account the following limitations. First of all, DIF should be tested according to different conditions to ensure comparability of the instrument across conditions. Unfortunately, our sample did not allow this type of comparison: the number of depressive cases according to PHQ-9 was too small for an adequate evaluation of DIF. Moreover, no information was gathered on physical conditions. Although invariance of the English version has already been assessed across a number of conditions such as spinal cord injury, muscular dystrophy, or multiple sclerosis [16,44], further work is needed to assess DIF by conditions on PROMIS Depression items in other language versions. Second, to reduce

assessment burden, Major Depression was assessed with PHQ-9 diagnosis algorithm as gold standard instead of a structured or semi-structured interview with standard criteria. However, PHQ-9 has demonstrated excellent diagnostic accuracy for depressive disorder [45]. Finally, scores of the short forms and of the CATs were extracted from the responses to the full item bank. This might overestimate the performance of the different forms, as extraneous dimensions in actual responses add noise to the population IRT model. This was partially minimized by using person-parameter theta in the Short Forms and CATs from actual responses to the item bank, and not from simulated responses.

On the other hand, an important strength of this study is that it rigorously and comprehensively followed state-of-the-art methods used with the original PROMIS Depression item bank, allowing for an unbiased comparison.

The Spanish PROMIS Depression item bank accomplished the IRT assumptions of unidimensionality, local independence, monotonicity, and scalability. These results support the use and interpretation of an overall measure of depression based on IRT methodology. Previous studies evaluating PROMIS Depression, also found a single dimension [16,46] even if some of the times RMSEA values were over 0.08 suggesting sub-optimal fit.

Reliability of the item bank and short forms were high: all internal consistencies were over 0.90. This is in line with previously reported values for the general population and patient samples for the English version [32,47]. As reported in other studies [47,48], PROMIS internal consistency outperformed that of CES-D and PHQ-9. Test information of the item bank showed reliabilities over 0.90 from theta -1 to 4, covering 85% of the theoretical distribution of the trait. Short forms covered from theta -0.5 to 3.5 in the 8-item version with reliabilities over 0.90, a range that gets narrower as the number of items decreases. Reliability of 0.99 is achieved at the point of maximum information. This means that changes in z-scores equal or greater than 0.22 or 2.2 units in T-scores, corresponding to a small impact in Cohen's criteria for effect size, would be detected as significant at 5% nominal error. However, as it has been previously observed [32], information is substantially reduced at the lower end of the continuum, which corresponds to

low depression levels. Actually, only 4 of the items showed negative thresholds. Due to the low number of items and lower information gathered at the lower end of the scale, PROMIS Depression fixed short forms have presented substantial floor effect of more than 20% in patient samples with multiple sclerosis [49] or knee osteoarthritis[50]. This trend is also observed in our general population sample, especially in the short forms (43% scoring the minimum for the 4-items version and 26% for the 8-items). Floor effects of CATs were lower suggesting that CATs might be appropriate in situations where individuals with low levels of depression are to be expected. Nevertheless, the PROMIS 8-items short form has consistently shown to be at least as precise as other fixed short tests such as PHQ-9 and CES-D across the whole continuum [20,46,51].

PROMIS Depression measures perform well in terms of convergent validity, showing high correlations with legacy measures. PROMIS Depression discriminant validity was moderate with anxiety instruments, similar to previous studies [20,34,41]. This indicates more about the overlap between the constructs of depression and anxiety as currently operationalized rather than a flaw in the forms' validity. Similar relationships were found for depression and anxiety legacy measures in our study, and in other psychometric studies of depression and anxiety [52]as well as in clinical and more basic research [53]. We believe that future PROMIS work might in fact contribute to elucidate the similarities and differences between depression and anxiety from different angles [54,55].

Even though PROMIS measures were not developed for diagnostic purposes, discriminant validity results show very good performance of PROMIS Depression for the detection of depression in the general population. In our sample, discrimination results for the short forms are very similar to the item bank. This is probably due to the fact that cut-off points for depression are located in an area where the information is high across all forms of the instrument. Other studies have reported that PROMIS Depression is able to discriminate caseness as well as other instruments that are specifically used as screeners [41,48,56].

An important result of this study is that almost identical shapes of the sample score distributions were obtained using Spanish and US sets of parameters. Only a minimal difference was observed in the point where the reference value is set. This highlights a frequently overlooked property of IRT scores: scale metric is arbitrary provided that there is measurement invariance, meaning that the same results –up to a linear transformation of estimated item parameters- will be obtained regardless of the population where they have been calibrated. In PROMIS, scoring standards are "norm based" in the sense that person scores represent the location with respect to the average of a reference population arbitrarily decided, in this case the US general population. If we had decided to fix metrics in the Spanish general population, population average in Spain would have been T=50, and US and Spanish sample-estimated item parameters would be a linearly equivalent. Using different parameters for each specific population has the undesirable consequence of masking real differences across populations, misleadingly indicating all averages are equal. This is the same as comparing differences in means across groups using standardized scores within those groups. A unique common metric is the only way we could assess the variability in scores due to differences in groups.

PROMIS has chosen to use a metric based on the US general population. This may not the most interpretable metric in other populations, such as Spain. However, interpretability of the scores could be improved by developing country specific norms based on this common metric, obtaining cut scores for clinically meaningful category intervals, as it have already been done for other PROMIS domains [57,58], or identifying score differences that represent a minimal important change. On the other hand, as reference values themselves can be arbitrarily chosen, other options would be to obtain scoring standards referred to content-related interpretations, mapping items into the distribution of person performance with standard-setting methods such as Angoff's or Bookmark methods [59]. This strategy is successfully used in international education research as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), where individual and group scores have substantive interpretations in reference to content-related standards [10,11].

In any case, applying content-related interpretations in the health context has great potential for yielding new information on health constructs, creating disorder-specific cut-points, or minimal important differences, and also comparative research.

## 5. Conclusions

Spanish PROMIS Depression item bank and short forms have shown for the first time measurement equivalence of a different language, with excellent metric properties and good cross-cultural validity. Our study provides evidence of the adequacy of PROMIS Depression measures for monitoring depression in the general population, both in Spain and across countries. It also suggests that they are also as good instruments as legacy measures for detection of depression. When individuals with low levels of depression are to be identified, CAT administrations or tailored static short forms that include items at the lower end of the continuum might be more appropriate than the originally proposed short forms.

## Tables and figures titles and footnotes

Table 1. Parameters from the Graded Response from Calibrations obtained using the Spanish PROMIS Depression Panel data

Table 2. Multi-trait multi-methods correlation matrix of PROMIS depression measures with other scales.

Footnotes:

†Abbreviations: CAT (Computer Adaptive Test); CES-D (Center for Epidemiologic Studies Depression Scale): PHQ-9 (Patient Health Questionnaire 9 item); GAD-7 (Generalized Anxiety Disorder 7 item); BAI (Beck Anxiety Inventory); WHODAS (World Health Organization Disability Assessment Schedule)

ǂ Diagonal terms within brackets are Cronbach's alpha internal consistency coefficient

$ Mixed-effects Intraclass Correlation coefficient for absolute agreement among PROMIS depression measures

& Not available because both scales were not administered to the same individuals

Figure 1. Comparison of initial theta versus purified theta taking into account language specific parameters for item 12, highlighted for differential item functioning.

Figure 2. Kernel density functions of scores based on US calibrations and Spanish calibrations

Figure 3. Test Information Curve for the PROMIS depression full item and short versions

Figure 4. Roc curves for diagnostic accuracy of PROMIS depression measures for depression diagnostic (PHQ-9 Major Depressive Syndrome)

## Acknowledgements

This work was supported by grants from Instituto de Salud Carlos III FEDER: Fondo Europeo de Desarrollo Regional» (PI13/00506) and the Department of Health of the Generalitat de Catalunya, Spain (AGAUR 2014 SGR 748; AGAUR 2009 SGR 1095). G.V. was supported by "Fondo de Investigación Sanitaria, Instituto de Salud Carlos III (ISCIII)" (ECA07/059). C.G.F. was supported by a ''Juan de la Cierva'' fellowship to from Ministerio de Ciencia e Innovación FSE (JCI-2009-05486).

We thank Helena Correia for her assistance on cognitive debriefing of PROMIS measures in Spain.

Research did not involve human or animal experimentation. The authors declare that they have no conflict of interest.

**References**


[1] Black N. Patient reported outcome measures could help transform healthcare. Br Med J 2013;346:f167.

[2] Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, et al. Reporting of Patient-Reported Outcomes in Randomized Trials. JAMA 2013;309:814. doi:10.1001/jama.2013.879.

[3] Devlin NJ, Appleby J. Getting the most out of PROMs: Putting health outcomes at the heart of NHS decision-making. Health Econ 2010:1–92.

[4] Ader DN. Developing the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45:S1–2. doi:10.1097/01.mlr.0000260537.45076.74.

[5] Van der Linden W, Hambleton R. Handbook of Modem Item Response Theory. New York, NY: Springer New York; 1998. doi:10.1007/978-1-4757-2691-6.

[6] Hambleton RK, Patsula L. Adapting Tests for Use in Multiple Languages and Cultures. Soc Indic Res 1998;45:153–71. doi:10.1023/A:1006941729637.

[7] Milfont TL, Fischer R. Testing measurement invariance across groups: Applications in crosscultural research. Int J Psychol Res 2010;3:111–21.

[8] Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45:S22–31.

[9] Van der Linden WJ, Glas CAW. Computerized adaptive testing: Theory and practice. Boston, MA: Kluwer; 2000.

[10] OECD. PISA 2012 technical report. Paris: 2014.

[11] Martin MO, Mullis IVS, Foy P, Arora A. Creating and Interpreting the TIMSS and PIRLS 2011 Context Questionnaire Scales Reporting Context Questionnaire Scales in. TIMSS 2011 Assess Fram 2011:1–11.

[12] Alonso J, Bartlett SJ, Rose M, Aaronson NK, Chaplin JE, Efficace F, et al. The case for an international patient-reported outcomes measurement information system (PROMIS(R)) initiative. Heal Qual Life Outcomes 2013;11:210.

[13] Turk DC, Dworkin RH, Allen RR, Bellamy N, Brandenburg N, Carr DB, et al. Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. Pain 2003;106:337–45.

[14] Paap MCS, Bode C, Lenferink LIM, Terwee CB, van der Palen J. Identifying key domains of health-related quality of life for patients with chronic obstructive pulmonary disease: interviews with healthcare professionals. Qual Life Res 2015;24:1351–67. doi:10.1007/s11136-014-0860-z.

[15] Schalet BD, Pilkonis PA, Yu L, Dodds N, Johnston KL, Yount S, et al. Clinical validity of PROMIS Depression, Anxiety, and Anger across diverse clinical samples. J Clin Epidemiol 2016;73:119–27. doi:10.1016/j.jclinepi.2015.08.036.

[16] Cook KF, Kallen MA, Bombardier C, Bamer AM, Choi SW, Kim J, et al. Do measures of depressive symptoms function differently in people with spinal cord injury versus primary care patients: the CES-D, PHQ-9, and PROMIS®-D. Qual Life Res 2016:1–10. doi:10.1007/s11136-016-1363-x.

[17] Kim J, Chung H, Askew RL, Park R, Jones SMW, Cook KF, et al. Translating CESD-20 and PHQ-9 Scores to PROMIS Depression. Assessment 2015. doi:10.1177/1073191115607042.

[18] PROMIS. PROMIS Translations available (web page) 2016. http://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-

promis/available-translations (accessed July 1, 2016).
[19] Terwee CB, Roorda LD, De Vet HCW, Dekker J, Westhovens R, Van Leeuwen J, et al. Dutch-Flemish translation of 17 item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS). Qual Life Res 2014;23:1733–41. doi:10.1007/s11136-013-0611-6.
[20] Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS(R)): depression, anxiety, and anger. Assessment 2011;18:263–83. doi:10.1177/1073191111411667.Item.
[21] Correia H. PROMIS Instrument Development and Validation Scientific Standards Version 2.0. Appendix 14. Translation and Cultural Adaptation. NIH PROMIS Web Page 2013.
[22] Radloff LS. The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. Appl Psychol Meas 1977;1:385–401.
[23] Vilagut G, Forero CG, Barbaglia G, Alonso J. Screening for Depression in the General Population with the Center for Epidemiologic Studies Depression (CES-D): A Systematic Review with Meta-Analysis. PLoS One 2016;11:e0155431. doi:10.1371/journal.pone.0155431.
[24] Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann 2002;32:509–15. doi:170553651.
[25] Spitzer RL, Kroenke K, Williams JBW, Löwe B. A Brief Measure for Assessing Generalized Anxiety Disorder. Arch Intern Med 2006;166:1092. doi:10.1001/archinte.166.10.1092.
[26] Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: psychometric properties. J Consult Clin Psychol 1988;56:893–7.
[27] Ustun TB, Kostanjsek N, Chatterji S, Rehm J. Measuring health and disability: manual for WHO Disability Assessment Schedule (WHODAS 2.0). Geneva: World Health Organization; 2010.
[28] Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Model A Multidiscip J 1999;6:1–55.
[29] Linacre JM. Local independence and residual covariance: A study of olympic figure skating ratings. J Appl Meas 2009;10:157–69.
[30] Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika 1970;35:139. doi:10.1007/BF02290599.
[31] McKinley RL, Mills CN. A Comparison of Several Goodness-of-Fit Statistics. Appl Psychol Meas 1985;9:49–57. doi:10.1177/014662168500900105.
[32] Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Qual Life Res 2010;19:125–36. doi:10.1007/s11136-009-9560-5.
[33] Swaminathan H, Rogers HJ. Detecting Differential Item Functioning Using Logistic Regression Procedures. J Educ Meas 1990;27:361–70. doi:10.1111/j.1745-3984.1990.tb00754.x.
[34] Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. 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 2010;63:1179–94. doi:10.1016/j.jclinepi.2010.04.011.
[35] Lord, Frederic M; Novick MR. Statistical Theories of Mental Test Scores. Reading, Mass: Addison-Wesley; 1968.
[36] McGraw KO, Wong SP. Forming inferences about some intraclass correlation

coefficients. Psychol Methods 1996;1:30–46.
[37] S A S Institute Inc. SAS/STAT software, version 9.1 for Windows. Cary, NC: SAS Institute Inc.; 2002.
[38] R Core Team. R: A language and environment for statistical computing. R Found Stat Comput 2014.
[39] Muthén LK, Muthén BO. Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén; 2012.
[40] Choi SW. Firestar : Computerized Adaptive Testing Simulation Program Response Theory Models. Appl Psychol Meas 2009;33:644–5. doi:10.1177/0146621608329892.
[41] Vilagut G, Forero CG, Adroher ND, Olariu E, Cella D, Alonso J, et al. Testing the PROMIS® Depression measures for monitoring depression in a clinical sample outside the US. J Psychiatr Res 2015;68:140–50. doi:10.1016/j.jpsychires.2015.06.009.
[42] Dawson J, Doll H, Fitzpatrick R, Jenkinson C, Carr AJ. The routine use of patient reported outcome measures in healthcare settings. BMJ 2010;340:c186. doi:10.1136/bmj.c186.
[43] Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN, et al. 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 2009;51:148–80. doi:10.1111/j.1365-2958.2010.07165.x.Characterization.
[44] Chung H, Kim J, Askew RL, Jones SMW, Cook KF, Amtmann D. Assessing measurement invariance of three depression scales between neurologic samples and community samples. Qual Life Res 2015;24:1829–34. doi:10.1007/s11136-015-0927-5.
[45] Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. Can Med Assoc J 2012;184:E191–6. doi:10.1503/cmaj.110829.
[46] Schalet BD, Cook KF, Choi SW, Cella D. Establishing a Common Metric for Depressive Symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. Psychol Assess 2014;26:88-513–27. doi:10.1016/j.janxdis.2013.11.006.
[47] Pilkonis PA, Yu L, Dodds NE, Johnston KL, Maihoefer CC, Lawrence SM. Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study. J Psychiatr Res 2014;56:112–9. doi:10.1016/j.jpsychires.2014.05.010.
[48] Kroenke K, Yu Z, Wu J, Kean J, Monahan PO. Operating Characteristics of PROMIS Four-Item Depression and Anxiety Scales in Primary Care Patients with Chronic Pain. Pain Med 2014;15:1892–901. doi:10.1111/pme.12537.
[49] Amtmann D, Kim J, Chung H, Bamer AM, Askew RL, Wu S, et al. Comparing CESD-10, PHQ-9, and PROMIS depression instruments in individuals with multiple sclerosis. Rehabil Psychol 2014;59:220–9. doi:10.1037/a0035919.
[50] Driban JB, Morgan N, Price LL, Cook KF, Wang C. Patient-Reported Outcomes Measurement Information System (PROMIS) instruments among individuals with symptomatic knee osteoarthritis: a cross-sectional study of floor/ceiling effects and construct validity. BMC Musculoskelet Disord 2015;16:253. doi:10.1186/s12891-015-0715-y.
[51] Gibbons LE, Feldman BJ, Crane HM, Mugavero M, Willig JH, Patrick D, et al. Migrating from a legacy fixed-format measure to CAT administration: calibrating

the PHQ-9 to the PROMIS depression measures. Qual Life Res Qual Life Res 2011;20:1349–57. doi:10.1007/s11136-011-9882-y.
[52] Den Hollander-Gijsman ME, Wardenaar KJ, De Beurs E, Van Der Wee NJA, Mooijaart A, Van Buuren S, et al. Distinguishing symptom dimensions of depression and anxiety: An integrative approach. J Affect Disord 2012;136:693–701. doi:10.1016/j.jad.2011.10.005.
[53] Craske MG. Transdiagnostic treatment for anxiety and depression. Depress Anxiety 2012;29:749–53. doi:10.1002/da.21992.
[54] Sunderland M, Mewton L, Slade T, Baillie AJ. Investigating differential symptom profiles in major depressive episode with and without generalized anxiety disorder: true co-morbidity or symptom similarity? Psychol Med 2010;40:1113–23. doi:10.1017/S0033291709991590.
[55] Simms LJ, Grös DF, Watson D, O'Hara MW. Parsing the general and specific components of depression and anxiety with bifactor modeling. Depress Anxiety 2008;25. doi:10.1002/da.20432.
[56] Fischer HF, Klug C, Roeper K, Blozik E, Edelmann F, Eisele M, et al. Screening for mental disorders in heart failure patients using computer-adaptive tests. Qual Life Res 2014;23:1609–18. doi:10.1007/s11136-013-0599-y.
[57] Cella D, Choi S, Garcia S, Cook KF, Rosenbloom S, Lai J-S, et al. Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. Qual Life Res 2014;23:2651–61. doi:10.1007/s11136-014-0732-6.
[58] Cook KF, Victorson DE, Cella D, Schalet BD, Miller D. Creating meaningful cut-scores for Neuro-QOL measures of fatigue, physical functioning, and sleep disturbance using standard setting with patients and providers. Qual Life Res 2015;24:575–89. doi:10.1007/s11136-014-0790-9.
[59] Hambleton RK. Setting performance standards on educational assessments and criteria for evaluating the process. In: Cizek GJ, editor. Setting Perform. Stand. Concepts, methods Perspect., Mahwah, NJ: Lawrence Erlbaum Associates; 2001, p. 89–116.

**Table 1. Parameters from the Graded Response from Calibrations obtained using the Spanish PROMIS Depression Panel data**

| | IRT parameters | | | | | Item fit statistics | |
|---|---|---|---|---|---|---|---|
| **Item** | **Slope** | **Thresholds** | | | | **S-X² (DF)** | **p-value** |
| | **a** | **b1** | **b2** | **b3** | **b4** | | **(S-X²)** |
| EDDEP04. I felt worthless | 3.47 | 0.49 | 1.14 | 2.15 | 2.89 | 91.5 (91) | 0.465 |
| EDDEP05. I felt that I had nothing to look forward to | 3.36 | 0.12 | 0.86 | 1.84 | 2.88 | 94.9 (96) | 0.512 |
| EDDEP06. I felt helpless | 3.9 | 0.34 | 1 | 1.81 | 2.91 | 126.6 (91) | 0.008 |
| EDDEP07. I withdrew from other people | 2.55 | 0.16 | 0.96 | 1.98 | 3.37 | 113 (115) | 0.535 |
| EDDEP09. I felt that nothing could cheer me up | 4.0 | 0.3 | 0.94 | 1.9 | 2.83 | 76 (84) | 0.722 |
| EDDEP14. I felt that I was not as good as other people | 3.01 | 0.28 | 0.97 | 1.88 | 2.59 | 115.8 (110) | 0.335 |
| EDDEP17. I felt sad | 3.06 | -0.48 | 0.37 | 1.51 | 2.51 | 94.4 (94) | 0.468 |
| EDDEP19. I felt that I wanted to give up on everything | 3.74 | 0.45 | 1.13 | 1.96 | 2.46 | 84.7 (88) | 0.581 |
| EDDEP21. I felt that I was to blame for things | 2.81 | 0.26 | 1 | 2.04 | 2.71 | 121 (111) | 0.242 |
| EDDEP22. I felt like a failure | 3.69 | 0.39 | 0.97 | 1.82 | 2.62 | 87.3 (98) | 0.771 |
| EDDEP23. I had trouble feeling close to people | 2.69 | 0.32 | 1.15 | 1.94 | 2.96 | 125.8 (114) | 0.212 |
| EDDEP26. I felt disappointed in myself | 3.56 | 0.24 | 0.95 | 1.83 | 2.54 | 81.1 (96) | 0.862 |
| EDDEP27. I felt that I was not needed | 2.95 | 0.43 | 1.08 | 1.97 | 2.95 | 121.3 (106) | 0.146 |
| EDDEP28. I felt lonely | 2.8 | 0.18 | 0.91 | 1.7 | 2.45 | 132.9 (118) | 0.164 |
| EDDEP29. I felt depressed | 4.07 | 0.1 | 0.75 | 1.64 | 2.43 | 94.4 (88) | 0.302 |
| EDDEP30. I had trouble making decisions | 1.91 | -0.29 | 0.96 | 2.44 | 3.82 | 107.2 (103) | 0.369 |
| EDDEP31. I felt discouraged about the future | 2.55 | -0.37 | 0.52 | 1.64 | 2.59 | 118.9 (113) | 0.333 |
| EDDEP35. I found that things in my life were overwhelming | 2.82 | 0.06 | 0.85 | 1.85 | 2.78 | 145.4 (110) | 0.013 |
| EDDEP36. I felt unhappy | 2.92 | -0.42 | 0.52 | 1.7 | 2.65 | 92.6 (95) | 0.55 |

| | IRT parameters | | | | | Item fit statistics | |
|---|---|---|---|---|---|---|---|
| **Item** | **Slope** | **Thresholds** | | | | **S-X² (DF)** | **p-value** |
| EDDEP39. I felt I had no reason for living | 3.31 | 0.98 | 1.49 | 2.28 | 2.9 | 98.9 (80) | 0.075 |
| EDDEP41. I felt hopeless | 3.78 | 0.41 | 1.03 | 2.04 | 2.77 | 76.8 (92) | 0.872 |
| EDDEP42. I felt ignored by people | 2.5 | 0.28 | 1.15 | 2.13 | 3.03 | 102.7 (112) | 0.725 |
| EDDEP44. I felt upset for no reason | 2.96 | 0.39 | 1.17 | 2.08 | 2.88 | 141.5 (100) | 0.004 |
| EDDEP45. I felt that nothing was interesting | 3.81 | 0.52 | 1.13 | 2.04 | 2.83 | 84.7 (89) | 0.61 |
| EDDEP46. I felt pessimistic | 2.95 | 0.01 | 0.79 | 1.78 | 2.41 | 104.5 (109) | 0.603 |
| EDDEP48. I felt that my life was empty | 3.88 | 0.57 | 1.19 | 1.89 | 2.56 | 98.4 (87) | 0.19 |
| EDDEP50. I felt guilty | 2.56 | 0.4 | 1.16 | 2.2 | 2.77 | 127.8 (106) | 0.074 |
| EDDEP54. I felt emotionally exhausted | 3.06 | 0.12 | 0.85 | 1.64 | 2.42 | 160 (111) | 0.002 |

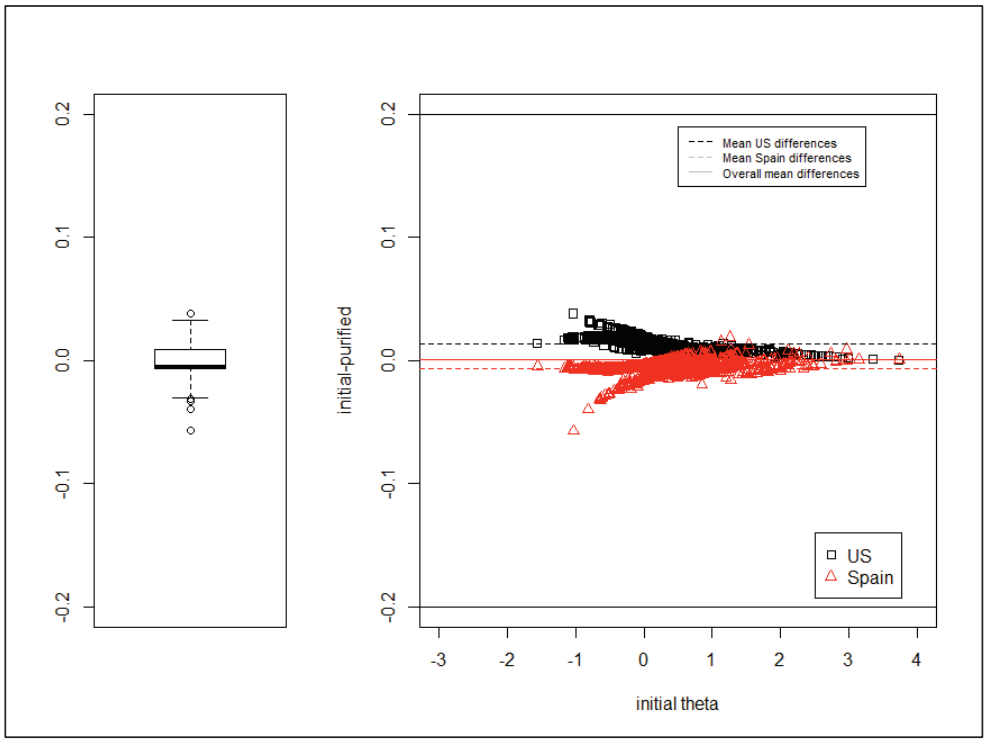



**Figure 1. Comparison of initial theta versus purified theta taking into account language specific parameters for item 12, highlighted for differential item functioning**

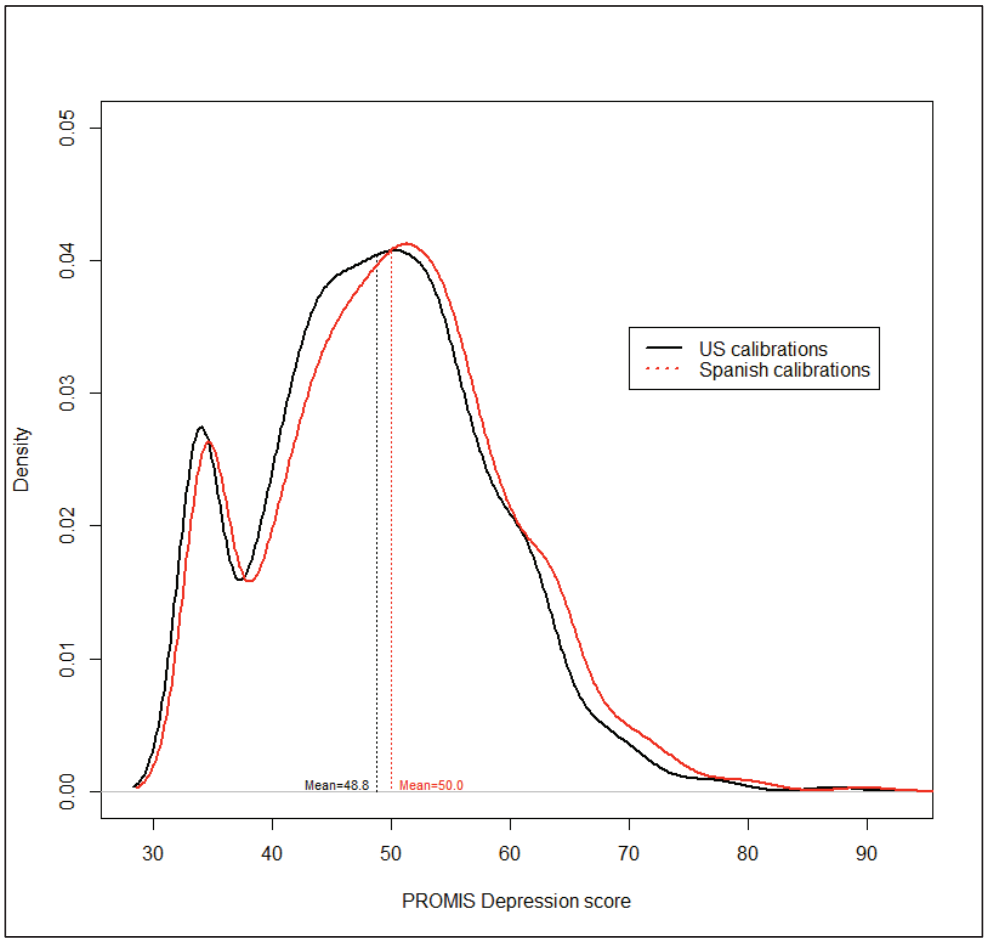



**Figure 2. Kernel density functions of scores based on US calibrations and Spanish calibrations**

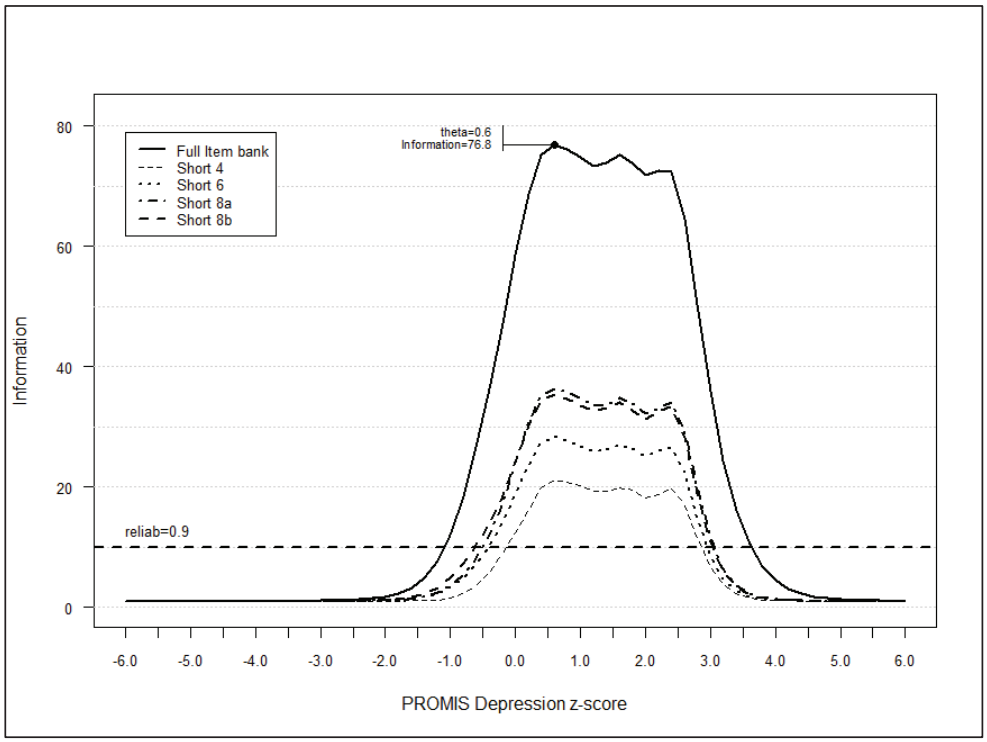



**Figure 3. Test Information Curve for the PROMIS depression full item and short versions**

**Table 2. Multi-trait multi-methods correlation matrix of PROMIS depression measures with other scales.**

| | PROMIS Depression | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Full Item bank** | **Short 4a** | **Short 6a** | **Short 8a** | **Short 8b** | **CAT (SE≤0.32)** | **CAT (8 items)** | **PROMIS Anxiety** | **PROMIS Anger** | **CES-D** | **PHQ-9** | **GAD-7** | **BAI** | **WHODAS** |
| PROMIS Depression - Full bank | *[0.98*]* | | | | | | | | | | | | | |
| - Short 4a | 0.91[$] | *[0.90]* | | | | | | | | | | | | |
| - Short 6a | 0.95[$] | 0.96[$] | *[0.93]* | | | | | | | | | | | |
| - Short 8a | 0.96[$] | 0.95[$] | 0.99[$] | *[0.95]* | | | | | | | | | | |
| - Short 8b | 0.96[$] | 0.94[$] | 0.98[$] | 0.99[$] | *[0.95]* | | | | | | | | | |
| - CAT (SE≤0.32) | 0.96[$] | 0.89[$] | 0.94[$] | 0.94[$] | 0.95[$] | NA | | | | | | | | |
| - CAT (8 items) | 0.97[$] | 0.91[$] | 0.96[$] | 0.96[$] | 0.97[$] | 0.97[$] | NA | | | | | | | |
| PROMIS Anxiety | 0.83 | 0.78 | 0.80 | 0.79 | 0.80 | 0.80 | 0.81 | *[0.97]* | | | | | | |
| PROMIS Anger | 0.70 | 0.63 | 0.67 | 0.67 | 0.68 | 0.68 | 0.68 | 0.73 | *[0.95]* | | | | | |
| CES-D | 0.73 | 0.74 | 0.74 | 0.74 | 0.73 | 0.72 | 0.73 | 0.71 | 0.57 | *[0.89]* | | | | |
| PHQ-9 | 0.74 | 0.74 | 0.73 | 0.74 | 0.73 | 0.71 | 0.73 | 0.69 | 0.56 | -[&] | *[0.88]* | | | |
| GAD-7 | 0.73 | 0.72 | 0.72 | 0.71 | 0.72 | 0.71 | 0.72 | 0.74 | 0.62 | -[&] | 0.76 | *[0.90]* | | |
| BAI | 0.54 | 0.56 | 0.55 | 0.54 | 0.54 | 0.54 | 0.53 | 0.70 | 0.52 | 0.66 | -[&] | -[&] | *[0.93]* | |
| WHODAS | 0.66 | 0.67 | 0.65 | 0.65 | 0.63 | 0.61 | 0.63 | 0.60 | 0.52 | -[&] | 0.64 | 0.59 | -[&] | *[0.93]* |

†Abbreviations: CAT (Computer Adaptive Test); CES-D (Center for Epidemiologic Studies Depression Scale): PHQ-9 (Patient Health Questionnaire 9 item); GAD-7 (Generalized Anxiety Disorder 7 item); BAI (Beck Anxiety Inventory); WHODAS (World Health Organization Disability Assessment Schedule)

‡ Diagonal terms within brackets are Cronbach's alpha internal consistency coefficient

[$] Mixed-effects Intraclass Correlation coefficient for absolute agreement among PROMIS depression measures

[&] Not available because both scales were not administered to the same individuals

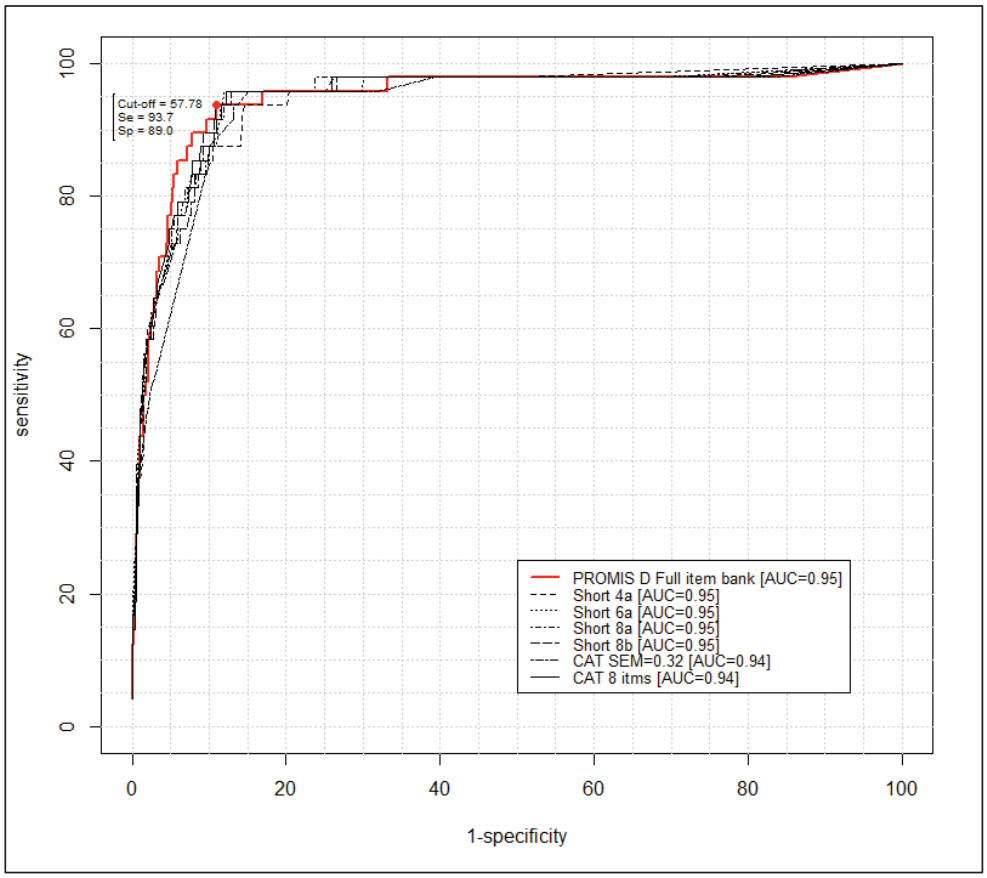



Figure 4. Roc curves for diagnostic accuracy of PROMIS depression measures for depression diagnostic (PHQ-9 Major Depressive Syndrome)

## 4.4. Article 5: *"Testing the PROMIS® Depression measures for monitoring depression in a clinical sample outside the US"*

Vilagut G, Forero CG, Adroher ND, Olariu E, Cella D, Alonso J, on behalf of the INSAyD investigators. Testing the PROMIS® Depression measures for monitoring depression in a clinical sample outside the US. J Psychiatr Res. 2015; 68:140–50.

PMID: 26228413

Supplementary material for this article can be found in ANNEX 3 (page 193)

# 5. SUMMARY OF FINDINGS AND DISCUSSION

This dissertation provides evidence on the validity and diagnostic accuracy of available self-reported measures for the assessment of depression, following different approaches. On the basis of a dimensional stance of assessment these studies compare, on one side, generic versus specific measures of depression; and on the other side, CTT versus IRT approaches.

Attention has focused on three of these measures. First, the mental component of the SF-12, one of the most widely used health related quality of life measures, regularly included in a number of health interview surveys around the world, that allow obtaining summary estimates of self-perceived physical and mental health. Importantly, both the CESD and the SF-12 were developed under the CTT approach. However, an alternative scoring method for the SF-12 using item weights obtained from Rasch analysis was proposed after its development and, as part of our work for this dissertation, we have developed a multidimensional IRT model specific for the SF-12. second, the CES-D, a 20-item questionnaire that was specifically developed for assessing depressive symptomatology in the general population and has been extensively used for this purpose. Finally, we evaluate performance of the PROMIS Depression measures originally conceived from the IRT approach, which confer desirable properties to the measures, including higher precision, higher comparability of scores and greater flexibility of administration than CTT approach.

In this chapter, we will separately summarize and discuss the results regarding the different objectives of this thesis, each one referring to one of these questionnaires. An overall discussion evaluating the advantages of using self-reported measures in general, and comparing the different approaches is provided afterwards.

## 5.1. Findings and discussion by objective

### a) Objective 1: Performance of the SF-12v1 for the assessment of depression according to different scoring systems

To respond to the first objective, we first developed an IRT model for the SF-12 that was not subject to SF-36 scores, as existing scoring methods for the SF-12 are. To do so, as described in the second article of this dissertation [137], two different multidimensional structures and scoring algorithms for the SF-12v1 were proposed using IRT multidimensional graded response models: a) Unidimensional response process (URP): Two subsets of 6 items, each loading on one of two correlated latent dimensions; and b) Bidimensional response process (BRP): two uncorrelated latent dimensions, physical and mental, with both dimensions participating in all individual items. The BRP was the best fitting model with good results according to the goodness of fit statistics, and provided greatest information throughout the continuum and had highest model based reliability. These results suggested that the BRP was better suited for obtaining individual scores, and thus it was the one used in a subsequent article to evaluate the accuracy of the SF-12 to screen for depression. Data from the European Study for Epidemiology of Mental Disorders

(ESEMeD) project was used, consisting of representative samples from the adult general population of 6 European countries (n= 21,425), with 12-month and 30-day major depression diagnosis based on the CIDI 3.0 as gold standard. Results according to three scoring methods were compared: a) the Mental Component Summary (MCS-12) scoring method proposed by Ware [138]; b) the RAND-12 scoring proposed by Hays [139]; and c) our BRP scoring algorithm based on a multidimensional IRT.

***Summary of findings:***

Results showed AUCs around 0.92 for 30-days disorder, and slightly lower for the 12-month diagnostic (AUC=0.85), and no statistically significant differences were found across the three methods in terms of accuracy. The cut-off point that provided the best trade-off between sensitivity and specificity for the mental component of the SF-12 to screen for 30-day depressive disorders in Europe was 45.6, with sensitivity of 0.86 and specificity of 0.88. This threshold value would imply that when administered to the general European population, the SF-12 will miss only 14% of the true cases, while 12% of those identified as depressed would not comply with DSM-IV diagnostic criteria.

***Discussion:***

Based on our results, discriminant capacity of the mental component of the SF-12v1 to detect 30-days depressive disorder is high. This is in agreement with the only previous study [134] that had assessed the accuracy of the SF-12 by the time our article was published in Value in Health. In that study, conducted in Australia, Gill et al [134] reported

AUCs of 0.92 for both MCS-12 and RAND-12 for the assessment of 30-days depression using ICD-10 as gold standard. At the recommended cut-off point of 45, sensitivity was 0.87 and specificity was over 0.80, indicating similar sensitivity but lower specificity at a slightly lower cut-off point that the value of 45.6 recommended by us. Of note, AUCs obtained for depressive disorders are still high, showing that the mental component of the questionnaire is sensitive to the presence of a recent disorder or subthreshold symptomatology. A more recent study, also conducted in Australia, provides further evidence on the accuracy of the SF-12 for 30-day depressive disorder, showing slightly poorer but still high concordance with its target diagnosis, with an AUC of 0.85 for the MCS-12 scoring method and 0.87 for the RAND-12 [140].

The different scoring methods evaluated provided comparable results in terms of accuracy, although the recommended cut-off point differed from one method to the other: the cut-off point suggested for the RAND-12 MHC was 44.5, while for the BRP-12MHS it was 40.2. Nevertheless, it is important to stress that these cut-off points are not directly comparable, because the MCS-12 and RAND-12 MHC are norm- based scores with a mean of 50 in the US general population, and mean scores in our sample are higher (mean MCS-12 = 53.7 and mean RAND-12 MHC = 52.9), while BRP-12 MHS are T-scored to have a mean of 50 in our sample. To be able to compare them, specific methods of scale linking based on equipercentile curves or linear linking should be applied.

A limitation of the study that needs to be born in mind when interpreting the results is the fact that measurement invariance, that would indicate that a given measure can be interpreted in a conceptually similar manner across groups, has not been evaluated in the study. However, model fit for the BRP-12 is adequate. Even though noninvariance in all possible subgroups cannot be entirely ruled out by an isolated fit index, it would be more likely if fit indexes were so poor as to suggest that a different kind of model, affecting a subsample of substantial size, could be underlying the full sample.

## b) Objective 2: Exhaustive review of the evidence on the accuracy of the CES-D

To respond to the second objective of this thesis, we conducted a systematic review of the available evidence evaluating the accuracy of the CES-D for the detection of major depression, and provided pooled estimates of performance for different cut off points obtained with meta-analysis.

***Summary of findings:***

Our results, based on 28 studies evaluating performance of the CES-D for detecting major depression, indicate adequate diagnostic accuracy of the questionnaire, with an AUC estimate of 0.87. At the recommended threshold level of 16, sensitivity is close to 0.90, at the expense of a moderate specificity of 0.70. A higher cut-off value of 20 yields a better trade-off between sensitivity and specificity (sensitivity = 0.83 and specificity = 0.78).

No evidence is found supporting differential performance depending on whether the questionnaire is used in general population surveys or primary care settings; no statistically significant difference is observed in terms of other variables assessed, such as language (English versus other), prevalence of the disorder, or gold standard used. However, age group is only marginally non-significant and, although the nominal alpha level ($p=0.053$) is not reached, graphical results stratified by age seem to indicate lower accuracy among younger age individuals.

***Discussion:***

To our knowledge this is the first study providing meta-analytic evaluation of the performance of the CES-D questionnaire. Our results are within the range of other usual questionnaires. In particular, a review of the GHQ-12 reported similar results with median sensitivity of 0.84 and specificity of 0.76 for any current mental disorder [83]. However the GHQ-12 assesses less specific mental distress symptoms that are not exclusive of depression and other studies have reported slightly worse results when performance of the GHQ-12 to detect depression is specifically evaluated [141]. When our results on the CES-D are compared with studies published on the PHQ-9, the CES-D performs worse in terms of sensitivity and specificity [142,143]. However, most of the PHQ-9 results were based on primary care and clinical samples. While the little available evidence on the PHQ 9 for its use in general population indicates good properties [140,144,145], only one of these studies included diagnostic assessment with structured interviews allowing to evaluate screening performance and obtained good accuracy results with an AUC of 0.88 [140].

With regard to the selection of the cut-off point, according to the values of sensitivity and specificity obtained and assuming a 10% prevalence of depression, if 10,000 individuals were screened, 1,000 of them would be expected to have depression. Using the cut-off point of 16, 870 cases would be detected, whereas the cut-off of 20 would detect 830 individuals (4.6% less). Concurrently, the number of individuals that would be wrongly classified as probably depressed would decrease from 2,700 for the cut-point 16 to 1,980 for the cut-point 20 (a 27% reduction). If the CES-D is used as a case-finding instrument for identifying patients requiring in-depth evaluation, maximizing sensitivity should be prioritized so that missed cases are reduced. However, when medical and psychiatric resources are limited, it is also important to minimize the false-positive rate to reduce the burden of subsequent in-depth assessment. Thus, considering the important reduction in false positives, the cut-off point of 20 would be recommended in primary care settings. When the intended use is for epidemiological studies to evaluate the relationships between depressive symptomatology and other variables across population subgroups, this consideration may not be so important, as non-cases that are located over the upper boundary will probably be subthreshold individuals, especially given the controversy around categorical classifications [110,112].

For the interpretation of these results, it is important to take into account that accuracy results for different cut-off points were based on different studies, and most of them presented results only for the usually recommended threshold of 16. Thus, evidence for other cut-

off points is more limited, and the incremental validity across different thresholds would be more adequately assessed in study designs obtaining accuracy results for a complete set of cut-off points within the same sample. Moreover, some of the studies presented results only for the optimal threshold within that study, which has been found to overestimate test performance [146].

### c) Objective 3: Metric properties and accuracy of PROMIS Depression in the general population in Spain

To appraise the third objective of this dissertation, we carried out a cross-sectional on-line survey on a sample provided by a panel vendor company. The total sample size was 1,503 individuals and its distribution in terms of age, sex and region groups was equivalent to that of the Spanish general population.

***Summary of findings:***

Results show that Spanish PROMIS Depression measures fulfill the IRT assumption of unidimensionality, local independence and monotonicity, and high reliability and construct validity is found for all forms assessed. Only one item was flagged for differential item functioning according to language, although the effect at the test level was negligible, with score differences between initial and purified tests lower than 0.06, substantially smaller than the value 0.2 corresponding to a small effect size.

***Discussion:***

IRT assumptions have been confirmed previously for the original version of the instrument [98,147,148] and our results provide further support for the Spanish version for obtaining an overall measure of depression based on IRT methodology. The fact that no important differential functioning at the test level exists in terms of language lends support to evidence of the equivalence of the PROMIS "universal" translation approach. Consistently with previous results [149], our results also reveal no DIF in terms of age, sex or education, ensuring measurement equivalence between these groups.

Internal consistency reliabilities are over 0.90 for all forms, outperforming those from the legacy measures CES-D and PHQ-9, as already found in other studies [150,151]. Importantly, the test information function of the item bank showed reliabilities over 0.90 from theta scores -1 to 4, covering 85% of the entire theoretical distribution of the trait, with maximum information at theta=0.6 (T-score=56.0). Information at that point was 76.8 (i.e. standard error of measurement of 0.11). This implies that changes in z-scores greater than or equal to 0.22 (or 2.2 units in T-scores), corresponding to a small impact in Cohen's criteria for effect size, would be detected as statistically significant at the 5% nominal error level. Short forms provided reliabilities over 0.90 from theta around -0.5 or lower to approximately 3.5. Therefore, in all cases, information is substantially reduced at the lower end of the continuum (low depression levels), as already observed previously [98], due to the small number of items in this area. This causes substantial floor effects ranging from 43% in the 4-item short version to approximately 21% in the 8-item version in

our general population, and around 20% in other patient samples [148,152]. Floor effects of the item banks and CAT administrations are substantially lower, thus being more appropriate than short forms in situations where individuals with low levels of depression are to be expected. PROMIS Depression 8-items short form has consistently shown to be at least as precise as other fixed short tests such as PHQ-9 and CES-D across the whole continuum. This indicates that PROMIS measures have the potential to evaluate severity levels with greater precision at a wider range of the continuum than the other measures, even with a lower number of items [92,150,153–155].

All forms of PROMIS Depression have shown excellent discrimination [156] with AUC values over 0.94, and sensitivity of 0.93 and specificity of 0.89 at the recommended cut-off point of 57.8 for the item bank. Other studies have reported that PROMIS Depression is good at discriminating between cases and non-cases [150,157,158].

The limitations of this particular study need to be considered for the interpretation of results. First of all, DIF testing according to disease conditions was not carried out. It is important to do so in order to ensure comparability of scores among conditions. Unfortunately, our sample did not allow this type of comparison, as the number of depressive cases based on PHQ-9 was too small for an adequate evaluation and no information was gathered on physical conditions. Measurement invariance of the English version of PROMIS Depression measures has already been assessed across a number of conditions such as spinal cord injury, muscular dystrophy, or multiple sclerosis [16,45], but further work is needed to evaluate whether

equivalence across conditions is maintained for other language versions. Second, the gold-standard used for major depression was the PHQ-9 diagnosis. Even though it has demonstrated excellent diagnostic accuracy for depressive disorder [46], a structured or semi-structured interview based on standardized criteria are a more acceptable standard.

## d) Objective 4: Construct validity and responsiveness of PROMIS Depression in a clinical sample from Spain

To assess the fourth objective, data from the Inventory of Depression and Anxiety Symptoms (INSAyD) project was used [159], a prospective study of a sample of 218 patients from primary care and specialized mental health care seeking help for active symptoms of mood or anxiety.

***Summary of findings:***

Results show high sample-based internal consistency reliability for both PROMIS Depression item bank and 8-item short form, with Cronbach's alpha and lambda 2 over 0.94, and precision representing reliabilities of 0.90 or higher through a substantial range of the continuum for the item bank (from 1 standard deviation (SD) below the mean to around 1.7 SD above). PROMIS Depression measures show good known groups validity, with effect sizes over 1.6 among disorder groups expected to differ; adequate convergent and discriminant validity, with patterns of correlations with other measures following established *a priory* hypothesis; and high criterion validity, with area under the curve of 0.89. The measures detected large

differences from baseline to 3-month follow up for patients who recovered, and small differences for stable patients, indicating good responsiveness to change.

***Discussion:***

Analogous to the results observed on the general population discussed previously, PROMIS Depression measures assessed also show excellent metric properties when applied to a clinical sample of individuals with common mental disorders or subthreshold individuals with emotional problems but without an active DSM diagnosis. High correlations were found between PROMIS measures and other legacy measures for depression indicating high convergent validity. On the other hand, correlations with anxiety measures were moderate to high, consistently with previous literature [92,148,151,160,161]. This trend is also observed, and to an even greater extent, between the legacy measures, PHQ-9 and Hamilton Depression, and anxiety. This may be indicative of the important overlap between depression and anxiety as already highlighted by previous studies evaluating depression and anxiety constructs [162–165], as well as in clinical and more basic research [166–170]. Despite these high correlations with anxiety constructs, PROMIS Depression is able to discriminate individuals with depression from those with anxiety and subthreshold individuals, with large effect sizes for individual with depression or comorbid depression and anxiety as compared to individuals without an active diagnosis, and small effect sizes for individuals with anxiety without depression, supporting high discriminant validity of the instrument. The large area under the curve of PROMIS Depression, with a value close to 0.90 and substantially higher than that for anxiety disorders

(AUC=0.64), also provides support for the good discrimination ability of the instrument.

Another important result that supports the use of PROMIS Depression measures for the purposes of monitoring of severity levels of depression is the high responsiveness to change observed. Mean change of PROMIS Depression measures among recovered patients at 3 months had an effect size of 0.98 for the item bank, comparable to that of the PHQ-9 (Cohen's effect size d=1.05). It is important to emphasize the ability of PROMIS Depression measures to detect individual change. In particular, around the point that maximizes overall diagnostic effectiveness PROMIS scales are able to identify differences of 3 points or higher in T-scale as statistically significant: corresponding to a small effect size of 0.3.

The main limitation of this particular study is that the MINI-International Neuropsychiatric Interview 5.0 (MINI) was used as the gold standard. While other structured or semi-structured interviews such as the CIDI or the SCID may be more recognized references, they are too lengthy and thus their use was not feasible in our study, and the MINI has shown good diagnostic accuracy for major depressive disorder [72].

## 5.2. Overall discussion

Besides the limitations already highlighted for each of the specific objectives of this dissertation there are additional limitations that need to be taken into account for a global interpretation of the results. First, the selection of self-reported measures to be evaluated has not been exhaustive. As explained in chapter 1, many other questionnaires have been used to assess depression. In particular, the PHQ-9 was originally developed based on CTT approach to evaluate severity of depression in psychiatric patients, and has been recently highlighted as one of the measures with best diagnostic accuracy and good psychometric properties in clinical studies [142,143]. Even though there is little available evidence on the PHQ-9 for its use in general population and further evaluation is needed in this sense, some studies report high floor effects in samples of patients with chronic conditions, suggesting that it may not discriminate well among persons with low levels of depression [148]. Additionally, other studies have reported the lowest test information (highest error) throughout the continuum for the PHQ-9 as compared to PROMIS Depression or CES-D when all these tests are calibrated in a single metric. This might be related to the fact that the PHQ-9 is exclusively composed of the 9 symptoms from the symptoms criteria of depression. Importantly, the last question of the PHQ-9 assesses suicidal or self-injurious thoughts, and it may be problematic to include it in population surveys when no adequate interventions can be carried out on positive respondents to this item. Second, the gold standard applied for diagnostic accuracy evaluation was not common across studies. This might have affected comparability of the results obtained across measures. However, in the

systematic review carried out for the CES-D, the gold standard did not arise as a statistically significant source of heterogeneity. Third, the objectives of the papers included in this thesis complement each other, but each one uses a different sample and tackles a slightly different (although related) question. Because of this, this thesis analyzes the five articles to gain insight of different aspects of the adequacy of different self-reported measures for the assessment of depression in the general population, and it draws valid conclusions successfully mapping the question of interest, rather than providing definitive answers.

Despite these limitations, results from this dissertation show that the three measures herein assessed provide good diagnostic accuracy for depression screening, all of them with AUC values over 0.85, and with selected cut-off points providing sensitivity and specificity values well over 0.80 for most of the cases. Given that healthcare planning is mainly based on clinically relevant cases, it is of great importance to obtain adequate thresholds reflecting clinical cases from epidemiological measures [114]. However, self-reported measures also allow for a dimensional assessment of depression or psychological distress over a continuous spectrum. Thus, it is not necessary to restrict interpretations to a dichotomous classification based on the cut-off point, as the interpretation of scores along the continuum also provides valuable information. A continuous measure, for instance, allows identification of different severity levels [65] and it may also identify subthreshold individuals, which may benefit from preventive or treatment strategies specifically designed for their severity level. In diagnostic interviews, items to be assessed are compulsory and fixed,

based on clinical consensus rather than on empirical evidence [171]. Changes are typically imposed in criteria in order to increase consensus and interrater reliability (not validity), and even a small change in one criterion might have important effects on the definition of the clinical construct [172]. In contrast, item content of self-reported measures is based on their metric properties, taking both reliability and validity into account, and changes in individual items do not greatly influence the validity of the construct. The dimensional approach of self-reported measures avoids difficulties related to changes in the consensus criteria of categorical nosology.

Among the self-reported questionnaires available, a generic measure such as the SF-12 may provide good diagnostic accuracy for depression screening, as the results of our work have actually shown. Given its good metric properties proven in a large number of studies, the SF-12 would also be recommended for conducting broad evaluations of self-perceived physical and mental health from a continuous perspective. One limitation though is that the content of this scale does not allow obtaining additional information about the distribution of the depression construct in the population. For this purpose, specific questionnaires whose items are focused on the depression construct, such as the CES-D or PROMIS Depression, would be more adequate.

As of today, most self-reported health questionnaires, and specifically those on mental health, either general psychological distress or disorder specific, are based on the CTT approach. From CTT, questionnaires are closed tools where all items of the test are

administered to all respondents. Under these circumstances, the need to measure patients with varying levels of severity of disorder necessarily lengthens the test. Additionally, failure to include items about symptoms reflecting a wide range of severity may result in an instrument with floor or ceiling effect. Therefore, a precise and wide ranging instrument should include several questions at each relevant level of severity. In contrast, the flexibility of administration supported by IRT methods allows developing item banks with a wide spectrum of symptoms without increasing administration burden, as not all the items need to be administered to each subject. If required technical resources are available, one may choose to apply adaptive technology to construct an optimal test for each subject. This would be done by adaptively selecting further items from the item bank that are appropriate to the individual's trait level ($\theta$) after each item has been answered [136].

Alternatively, fixed short forms could be created by selecting the most informative items throughout the continuum or within a specific range of interest. Therefore, the same item bank could supply items to be administered to a general population sample in order to screen for probable cases, for instance, and to a clinical sample for severity evaluation. Although the items administered to the different samples might differ (with items administered to the general population sample probably representing lower levels of severity), they will all be scored using the same metric, facilitating harmonization and comparability across different forms of the same questionnaire.

Another important characteristic of IRT methodology as compared to CTT is the fact that, given a certain collection of items, IRT models provide a set of weights that optimize measurement precision [173]. This is observed in the IRT-based scores that we obtained for the SF-12, which outperformed the original scoring system in terms of reliability. Thus, with IRT methodology, efficiency of psychometric measurement can be optimized with little or no cost in precision. In particular, our results show that the PROMIS depression item bank, which is composed of 28 items, provide great precision throughout an important range of the continuum, especially over the areas representing highest levels of depression. Our CAT simulations show that with a stopping rule of a standard error of measurement (SEM) of 0.32 (equivalent to a reliability of 0.90) an average of 5.5 items and a median of 3 items are to be administered; and if a fixed stopping rule is set at 8 items, the average SEM is reduced to 0.27, corresponding to a reliability of almost 0.93. Similarly, with static short versions of 4 to 8 items, high information values around 20 (SEM=0.23) or more are still obtained in a relevant area of the continuum representing moderate to severe levels of depression. And yet, intraclass correlations between static short versions and CATs with the full item bank are over 0.90.

Of relevance, similarly to what has been found in previous research [98], our results indicate that measurement precision of PROMIS Depression at the healthier end of the continuum is relatively poor, and only 4 of the items present negative thresholds. This translates to substantial floor effects in the static short versions originally proposed, and may be an indication of the need to include additional items in the

lower end of the continuum. Mental health constructs such as depression are inherently multidimensional while PROMIS measures are based on unidimensional IRT. Given that application of unidimensional models to multidimensional data can result in biased trait estimates [174], a sufficiently unidimensional construct had to be obtained for the PROMIS Depression item bank. Therefore, items that did not comply with the IRT assumption of unidimensionality were discarded from the bank, and the final number of items included in the item bank was reduced. A CAT measure for Depression has been proposed based on Multidimensional IRT with 389 items, a substantially larger pool of items than PROMIS [12]. However, additional validation of this measure has not been reported so far, nor is there any information available on the adaptation of this item bank into other languages. Multidimensional CAT is an exciting area for future development, but can also be very computer intensive and the interpretation of scores may become even more complex [175]. Further research is needed on the performance of CAT administration in Multidimensional IRT methodology.

Another relevant point to be considered in the comparison between CTT and IRT approaches is that CTT estimates are dependent on both the test content and the sample. True scores are dependent on the test and item parameters, such as item difficulty (proportion correct) and discrimination (point biserial correlations) and test statistics such as reliability depend on the sample. Such dependencies can limit comparability of scores. The examinee's ability level in relation to the construct being measured is an intrinsic property of the individual, independent of the test. From CTT, a test can be conceived

as a closed sample of behaviors directly caused by the construct being assessed. The CTT test score for an individual is the number of behaviors in the test which are fulfilled by that individual. Therefore, test score (i.e. true score estimates) will always depend on the specific sample of behaviors that comprise the test. For instance, even though his intrinsic ability level is constant, an examinee will have a lower true score if the test is composed of infrequent behaviors and higher true score when it is composed of common behaviors.

In contrast, under the IRT approach, item and person parameters are sample independent, provided that the model fits the data. In IRT, item statistics are independent of the groups from which they are estimated, and scores describing subject proficiency do not depend on the test difficulty. Additionally, IRT models facilitate evaluation of measurement invariance at the item level using differential item functioning (DIF) technics, which examine whether or not the likelihood of item (category) endorsement is equal across subgroups, matched on the level of the trait being measured. Lack of measurement equivalence renders group comparisons ambiguous. Therefore, assessment of DIF provides valuable information on whether differences in the epidemiology of depression according to specific groups, such as cross-national differences, are real or are an artifact of aspects of the measurement process, such as differing interpretations of questions by members of the different groups. The relevant consequence is that IRT measures will better facilitate comparability of scores as compared to CTT.

Overall, the results of this thesis support the use of the dimensional approach using questionnaires for assessing depression in the general population, both for screening and severity assessment purposes. This approach is not opposed to the categorical evaluation used in diagnostic interviews, but rather may complement diagnosis by locating the individual in a severity spectrum, where the diagnostic decision is associated with a specific threshold value.

The decision to choose among questionnaires should depend on the substantive objectives of the survey. It should not be motivated by practical aspects such as burden of administration in terms of length or interview time, as practical challenges can be overcome by the use of the most adequate technical approach for administering and analyzing the questionnaire. And IRT is the optimal approach that allows the shortest administration while providing the maximum information.

# 6. CONCLUSIONS

Specific conclusions by questionnaire:

a) Multidimensional IRT provides a specific measurement model for the SF-12 that is detached from the SF-36 scores and has good metric properties.

b) The three scoring methods for the mental component of the SF-12 have good diagnostic accuracy, with similar results in terms of sensitivity and specificity.

c) The SF-12 could be used for the assessment of depression in epidemiologic studies. However, the SF-12 is not recommended when the study objectives are related to the dimensional assessment of the depression construct and/or its distribution in the population.

d) The CES-D has good accuracy for use in general population health surveys but it is not recommended as an isolated measure for individual diagnostic purposes.

e) Using a diagnostic cut-off point for the CES-D of 20 instead of the conventional value of 16 should be recommended. It reduces the false positive rates, which in turn diminishes the burden of in-depth assessment of potential cases without a great loss of sensitivity.

f) IRT based PROMIS Depression measures show excellent metric properties, especially diagnostic accuracy and responsiveness to change, supporting its adequacy for screening and monitoring of depression in the general population.

g) The PROMIS Depression measurement model shows excellent fit and cross-cultural measurement invariance, supporting direct comparability between populations.

General conclusions:

h) Unlike diagnostic interviews, self-reported measures allow for an adequate dimensional assessment of the depression construct providing relevant severity information beyond the categorical classification of individuals.

i) The decision about the selection of a questionnaire in general population surveys should be guided by substantive considerations, and not mostly based on practical issues (such as administration burden). Administration burden and other practical challenges can be solved by selecting the most adequate technical approach for administering and analyzing the instrument.

j) The IRT psychometric approach provides a flexible and precise method for administering and scoring questionnaires, allowing for direct comparability between populations.

## References


[1] Kessler RC, Bromet EJ. The epidemiology of depression across cultures. Annu Rev Public Health 2013;34:119–38. *doi:10.1146/annurev-publhealth-031912-114409*.

[2] Eaton WW, Martins SS, Nestadt G, Bienvenu OJ, Clarke D, Alexandre P. The Burden of Mental Disorders. Epidemiol Rev 2008;30:1–14. *doi:10.1093/epirev/mxn011*.

[3] Murray C] L. Burden of Disease A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020 Cambridge, MA: Harvard School of Public Health on behalf of the World Helalth Organization and the World Bank; 1996.

[4] Prince M, Patel V, Saxena S, Maj M, Maselko J, Phillips MR, et al. No health without mental health. Lancet 2007;370:859–77.

[5] Becker AE, Kleinman A. Mental Health and the Global Agenda. N Engl J Med 2013;369:66–73. *doi:10.1056/NEJMra1110827*.

[6] Thornicroft G, Votruba N, Whiteford H, Degenhardt L, Rehm J, al. et, et al. Does the United Nations care about mental health? The Lancet Psychiatry 2016;3:599–600. *doi:10.1016/S2215-0366(16)30079-7*.

[7] Centers for Disease Control and Prevention. Mental illness surveillance among adults in the United States. MMWR Suppl 2011;60:1–29.

[8] Eaton WW, Johns Hopkins Bloomberg School of Public Health. Department of Mental Health. Public mental health. Oxford

University Press; 2012.

[9] Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. Development of a computer-adaptive test for depression (D-CAT). Qual Life Res 2005;14:2277–91.

[10] Forkmann T, Kroehne U, Wirtz M, Norra C, Baumeister H, Gauggel S, et al. Adaptive screening for depression--recalibration of an item bank for the assessment of depression in persons with mental and somatic diseases and evaluation in a simulated computer-adaptive test environment. J Psychosom Res 2013; 75:437–43.

[11] Gardner W, Shear K, Kelleher KJ, Pajer KA, Mammen O, Buysse D, et al. Computerized adaptive measurement of depression: a simulation study. BMC Psychiatry 2004;4:13.

[12] Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kim JB, et al. Development of a computerized adaptive test for depression. Arch Gen Psychiatry 2012;69:1104–12.

[13] Steel Z, Marnane C, Iranpour C, Chey T, Jackson JW, Patel V, et al. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. Int J Epidemiol 2014:476–93. *doi:10.1093/ije/dyu038.*

[14] Bromet E, Andrade LH, Hwang I, Sampson NA, Alonso J, Girolamo G de, et al. Cross-national epidemiology of DSM-IV major depressive episode. BMC Med 2011;9:90. *doi:10.1186/1741-7015-9-90.*

[15] Lee S, Tsang A, Breslau J, Aguilar-Gaxiola S, Angermeyer M, Borges G, et al. Mental disorders and termination of education in high-income and low- and middle-income countries:

epidemiological study. Br J Psychiatry 2009;194:411–7. *doi:10.1192/bjp.bp.108.054841*.

[16] Vaughn MG, Wexler J, Beaver KM, Perron BE, Roberts G, Fu Q. Psychiatric correlates of behavioral indicators of school disengagement in the United States. Psychiatr Q 2011;82:191–206. *doi:10.1007/s11126-010-9160-0*.

[17] Fergusson DM, Boden JM, Horwood LJ. Recurrence of major depression in adolescence and early adulthood, and later mental health, educational and economic outcomes. Br J Psychiatry 2007;191:335–42. *doi:10.1192/bjp.bp.107.036079.*

[18] Kawakami N, Abdulghani EA, Alonso J, Bromet EJ, Bruffaerts R, Caldas-de-Almeida JM, et al. Early-life mental disorders and adult household income in the World Mental Health Surveys. Biol Psychiatry 2012;72:228–37. *doi:10.1016/j.biopsych.2012.03.009.*

[19] Breslau J, Miller E, Jin R, Sampson NA, Alonso J, Andrade LH, et al. A multinational study of mental disorders, marriage, and divorce. Acta Psychiatr Scand 2011;124:474–86. *doi:10.1111/j.1600-0447.2011.01712.x.*

[20] Derogatis LR, Morrow GR, Fetting J, Penman D, Piasetsky S, Schmale AM, et al. The prevalence of psychiatric disorders among cancer patients. JAMA 1983;249:751–7.

[21] Van der Kooy K, van Hout H, Marwijk H, Marten H, Stehouwer C, Beekman A. Depression and the risk for cardiovascular diseases: systematic review and meta analysis. Int J Geriatr Psychiatry 2007;22:613–26. *doi:10.1002/gps.1723.*

[22] Ohira T, Iso H, Satoh S, Sankai T, Tanigawa T, Ogawa Y, et al. Prospective study of depressive symptoms and risk of stroke among japanese. Stroke 2001;32:903–8.

[23] Pratt LA, Ford DE, Crum RM, Armenian HK, Gallo JJ, Eaton WW. Depression, psychotropic medication, and risk of myocardial infarction. Prospective data from the Baltimore ECA follow-up. Circulation 1996;94:3123–9.

[24] Gross AL, Gallo JJ, Eaton WW. Depression and cancer risk: 24 years of follow-up of the Baltimore Epidemiologic Catchment Area sample. Cancer Causes Control 2010;21:191–9. *doi:10.1007/s10552-009-9449-1.*

[25] Carney RM, Freedland KE, Miller GE, Jaffe AS. Depression as a risk factor for cardiac mortality and morbidity: a review of potential mechanisms. J Psychosom Res 2002;53:897–902.

[26] Scott KM, de Jonge P, Alonso J, Viana MC, Liu Z, O'Neill S, et al. Associations between DSM-IV mental disorders and subsequent heart disease onset: beyond depression. Int J Cardiol 2013;168:5293–9. *doi:10.1016/j.ijcard.2013.08.012.*

[27] Alonso J, de Jonge P, Lim CCW, Aguilar-Gaxiola S, Bruffaerts R, Caldas-de-Almeida JM, et al. Association between mental disorders and subsequent adult onset asthma. J Psychiatr Res 2014;59:179–88. *doi:10.1016/j.jpsychires.2014.09.007.*

[28] Anderson R, Freedland K, Clouse R, Lustman P. The prevalence of comorbid depression in adults with diabetes. Diabetes Care 2001;24:1069–78. *doi:10.1111/j.1464-5491.2009.02903.x.*

[29] de Jonge P, Alonso J, Stein DJ, Kiejna A, Aguilar-Gaxiola S, Viana MC, et al. Associations between DSM-IV mental disorders and diabetes mellitus: a role for impulse control disorders and depression. Diabetologia 2014;57:699–709. *doi:10.1007/s00125-013-3157-9.*

[30] Scott KM, de Jonge P, Alonso J, Viana MC, Liu Z, O'Neill S, et al. Associations between DSM-IV mental disorders and subsequent heart disease onset: beyond depression. Int J Cardiol 2013;168:5293–9. *doi:10.1016/j.ijcard.2013.08.012.*

[31] Aguilar-Gaxiola S, Loera G, Geraghty EM, Ton H, Lim CCW, de Jonge P, et al. Associations between DSM-IV mental disorders and subsequent onset of arthritis. J Psychosom Res 2016;82:11–6. *doi:10.1016/j.jpsychores.2016.01.006.*

[32] Bruffaerts R, Demyttenaere K, Kessler RC, Tachimori H, Bunting B, Hu C, et al. The associations between preexisting mental disorders and subsequent onset of chronic headaches: a worldwide epidemiologic perspective. J Pain 2015;16:42–52. *doi:10.1016/j.jpain.2014.10.002.*

[33] Alonso J, Angermeyer MC, Bernert S, Bruffaerts R, Brugha TS, Bryson H, et al. 12-Month comorbidity patterns and associated factors in Europe: Results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. Acta Psychiatr Scand Suppl 2004;109:28–37.

[34] Scott KM, Bruffaerts R, Tsang A, Ormel J, Alonso J, Angermeyer MC, et al. Depression-anxiety relationships with chronic physical conditions: Results from the World Mental Health surveys. J Affect Disord 2007;103:113–20.

*doi:10.1016/j.jad.2007.01.015.*

[35] Chapman DP, Perry GS, Strine TW. The vital link between chronic disease and depressive disorders. Prev Chronic Dis 2005;2:A14.

[36] Cuijpers P, Schoevers RA. Increased mortality in depressive disorders: a review. Curr Psychiatry Rep 2004;6:430–7.

[37] Wulsin LR, Vaillant GE, Wells VE. A systematic review of the mortality of depression. Psychosom Med 1999;61:6–17.

[38] Ormel J, Petukhova M, Chatterji S, Aguilar-Gaxiola S, Alonso J, Angermeyer MC, et al. Disability and treatment of specific mental and physical disorders across the world. Br J Psychiatry 2008;192:368–75. *doi:10.1192/bjp.bp.107.039107.*

[39] Alonso J, Petukhova M, Vilagut G, Chatterji S, Heeringa S, Üstün TB, et al. Days out of role due to common physical and mental conditions: Results from the WHO World Mental Health surveys. Mol Psychiatry 2011;16:1234–46. *doi:10.1038/mp.2010.101.*

[40] Lasalvia A, Zoppei S, Van Bortel T, Bonetto C, Cristofalo D, Wahlbeck K, et al. Global pattern of experienced and anticipated discrimination reported by people with major depressive disorder: A cross-sectional survey. Lancet 2013;381:55–62. *doi:10.1016/S0140-6736(12)61379-8.*

[41] Alonso J, Buron A, Bruffaerts R, He Y, Posada-Villa J, Lepine J-P, et al. Association of perceived stigma and mood and anxiety disorders: results from the World Mental Health Surveys. Acta Psychiatr Scand 2008;118:305–14. *doi:10.1111/j.1600-0447.2008.01241.x.*

[42] National Academies of Sciences E and M. Ending Discrimination Against People with Mental and Substance Use Disorders. Washington, D.C.: National Academies Press; 2016. *doi:10.17226/23442*.

[43] Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJL, et al. Burden of Depressive Disorders by Country, Sex, Age, and Year: Findings from the Global Burden of Disease Study 2010. PLoS Med 2013;10:e1001547. *doi:10.1371/journal.pmed.1001547*.

[44] Vos T, Barber RM, Bell B, Bertozzi-Villa A, Biryukov S, Bolliger I, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: A systematic analysis for the Global Burden of Disease Study 2013. Lancet 2015;386:743–800. *doi:10.1016/S0140-6736(15)60692-4*.

[45] Murray CJL, Barber RM, Foreman KJ, Ozgoren AA, Abd-Allah F, Abera SF, et al. Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990–2013: quantifying the epidemiological transition. Lancet 2015;386:2145–91. *doi:10.1016/S0140-6736(15)61340-X*.

[46] Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. The Lancet Psychiatry 2016;3:171–8. *doi:10.1016/S2215-0366(15)00505-2*.

[47] Greenberg PE, Fournier A-A, Sisitsky T, Pike CT, Kessler RC. The Economic Burden of Adults With Major Depressive Disorder in the United States (2005 and 2010). J Clin Psychiatry 2015;76:155–62.

*doi:10.4088/JCP.14m09298.*

[48] Gustavsson A, Svensson M, Jacobi F, Allgulander C, Alonso J, Beghi E, et al. Cost of disorders of the brain in Europe 2010. Eur Neuropsychopharmacol 2011;21:718–79. *doi:10.1016/j.euroneuro.2011.08.008.*

[49] Parés-Badell O, Barbaglia G, Jerinic P, Gustavsson A, Salvador-Carulla L, Alonso J, et al. Cost of Disorders of the Brain in Spain. PLoS One 2014;9:e105471. *doi:10.1371/journal.pone.0105471*.

[50] Weissman MM, Myers JK, Thompson WD. Depression and its treatment in a US urban community-1975-1976. Arch Gen Psychiatry 1981;38:417–21. *doi:10.1001/archpsyc.1981.01780290051005.*

[51] Lecrubier Y, Nutt, Kirmayer, Davidson, Ono, Lin, et al. Prescribing patterns for depression and anxiety worldwide. J Clin Psychiatry 2001;62:31–8.

[52] Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: Findings from the Global Burden of Disease Study 2010. Lancet 2013;382:1575–86. *doi:10.1016/S0140-6736(13)61611-6.*

[53] Patel V, Chisholm D, Parikh R, Charlson FJ, Degenhardt L, Dua T, et al. Global Priorities for Addressing the Burden of Mental, Neurological, and Substance Use Disorders. Dis. Control Priorities, Third Ed. (Volume 4) Ment. Neurol. Subst. Use Disord., 2016, p. 1–27. *doi:10.1596/978-1-4648-0426-7_ch1*.

[54] Rudenstine S, Galea S. Preventing brain disorders: a framework

for action. Soc Psychiatry Psychiatr Epidemiol 2015;50:833–41. *doi:10.1007/s00127-015-1007-4*.

[55] WHO. WHO | Public Health. WHO 2010. http://www.who.int/topics/public_health_surveillance/en/ (accessed June 25, 2016).

[56] Dohrenwend BP. A psychosocial perspective on the past and future of psychiatric epidemiology. Am J Epidemiol 1998;147:222–31.

[57] Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, et al. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. Arch Gen Psychiatry 1994;51:8–19. *doi:10.1001/archpsyc.1994.03950010008002*.

[58] Kessler RC, Merikangas KR. The National Comorbidity Survey Replication (NCS-R): background and aims. Int J Methods Psychiatr Res 2004;13:60–8.

[59] Bijl R V, van Zessen G, Ravelli A, de Rijk C, Langendoen Y. The Netherlands Mental Health Survey and Incidence Study (NEMESIS): objectives and design. Soc Psychiatry Psychiatr Epidemiol 1998;33:581–6.

[60] Kessler RC, Haro JM, Heeringa SG, Pennell B-E, ÃœstÃ¼n TB. The World Health Organization World Mental Health Survey Initiative. Epidemiol Psichiatr Soc 2006;15:161–6.

[61] de Pedro Cuesta J, Saiz Ruiz J, Roca M, Noguer I. Salud mental y salud pública en España: vigilancia epidemiológica y prevención. Psiquiatr Biológica 2016;23:67–73. *doi:10.1016/j.psiq.2016.03.001*.

[62] Goldenberg JS, Contreras Escudero L. Diseño y puesta en marcha de un sistema de vigilancia epidemiológica en salud mental. Rev Panam Salud Pública 2002;11:83–92. *doi:10.1590/S1020-49892002000200004.*

[63] Institut National de Santé Publique du Québec. Chronic Disease Surveillance. Surveillance of Mental Disorders in Québec : Prevalence , Mortality and Service Utilization Profile Introduction. Québec: 2010.

[64] Zhou W, Xiao S. Existing public health surveillance systems for mental health in China. Int J Ment Health Syst 2015;9:3. *doi:10.1186/1752-4458-9-3.*

[65] Freeman EJ, Colpe LJ, Strine TW, Dhingra S, McGuire LC, Elam-Evans LD, et al. Public health surveillance for mental health. Prev Chronic Dis 2010;7:A17.

[66] Duckworth GS, Kedward HB. Man or machine in psychiatric diagnosis. Am J Psychiatry 1978;135:64–8. *doi:10.1176/ajp.135.1.64.*

[67] American Psychiatric Association. Diagnostic and statistical manual of mental disorders (3rd Edition). 1980. *doi:10.1016/B978-1-4377-2242-0.00016-X.*

[68] WHO. ICD-10 Classification of Mental and Behavioral Disorders. Diagnostic Criteria for Research. World Health Organization Geneva 1993.

[69] Robins LN, Helzer JE, Croughan JL, Ratcliff KS. National Institute of Mental Health diagnostic interview schedule: Its history, characteristics, and validity. Arch Gen Psychiatry 1981;38:381–9. *doi:10.1001/archpsyc.1981.01780290015001.*

[70] Regier DA, Myers JK, Kramer M, Robins LN, Blazer DG, Hough RL, et al. The NIMH Epidemiologic Catchment Area program. Historical context, major objectives, and study population characteristics. Arch Gen Psychiatry 1984;41:934–41.

[71] American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). Fourth Edi. Washington, DC: American Psychiatric Association; 2000.

[72] Sheehan D V, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 2:22–33.

[73] Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, et al. The Composite International Diagnostic Interview. An epidemiologic Instrument suitable for use in conjunction with different diagnostic systems and in different cultures. Arch Gen Psychiatry 1988;45:1069–77.

[74] Kessler RC, Üstün TB. The World Mental Health (WMH) Survey Initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). Int J Methods Psychiatr Res 2004;13:93–121. *doi:10.1002/mpr.168.*

[75] Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SL, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002;32:959–76.

[76] Williams JWJ, Pignone M, Ramirez G, Perez Stellato C. Identifying depression in primary care: a literature synthesis of

case-finding instruments. Gen Hosp Psychiatry 2002;24:225–37.

[77] McDowell I. Measuring Health: A guide to rating scales and questionnaires, 3rd. Ed. New York: Oxford University Press; 2006.

[78] Ware, John E.; Snow, Kristin K; Kosinski, Mark; Gandek B. SF-36 Health Survey Manual and Interpretation Guide. Boston, MA: The Health Institute, New England Medical Center; 1993.

[79] WHOQOL Group. Development of the World Health Organization WHOQOL-BREF quality of life assessment. Psychol Med 1998;28:551–8.

[80] The EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. Health Policy (New York) 1990;16:199–208. *doi:10109801*.

[81] Goldberg DP, Blackwell B. Psychiatric illness in general practice. A detailed study using a new method of case identification. Br Med J 1970;1:439–43.

[82] Lehto-Järnstedt U-S, Aromaa A. Mental health measurement in comprehensive national health surveys. Report.

[83] Goldberg DP, Gater R, Sartorius N, Ustun TB, Piccinelli M, Gureje O, et al. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. Psychol Med 1997;27:191–7.

[84] Berwick DM, Murphy JM, Goldman PA, Ware JE, Barsky AJ, Weinstein MC. Performance of a Five-Item Mental Health Screening Test. Med Care 1991;29:169–76. *doi:10.1097/00005650-199102000-00008.*

[85] Furukawa TA, Kessler RC, Slade T, Andrews G. The performance

of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being. Psychol Med 2003;33:357–62. doi:10.1017/S0033291702006700.

[86] Pratt LA. Serious Psychological Distress, as Measured by the K6, and Mortality. Ann Epidemiol 2009;19:202–9. *doi:10.1016/j.annepidem.2008.12.005.*

[87] Williams JWJ, Noel PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? J Am Med Assoc 2002; 287:1160–70.

[88] Pignone MP, Gaynes BN, Rushton JL, Burchell CM, Orleans CT, Mulrow CD, et al. Screening for depression in adults: a summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 2002;136:765–76.

[89] Radloff LS. The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. Appl Psychol Meas 1977;1:385–401.

[90] Beck AT, Steer RA, Brown GK. Manual for the Beck depression inventory-II. San Antonio, TX Psychol Corp 1996:1–82.

[91] Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA 1999;282:1737–44.

[92] Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS(R)): depression, anxiety, and anger. Assessment 2011;18:263–83. *doi:10.1177/1073191111411667.Item.*

[93] Zung W. A self-rating depression scale. Arch Gen Psychiatry 1965;12:63–70. *doi:10.1001/archpsyc.1965.01720310065008.*

[94] Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561–71. *doi:10.1001/archpsyc.1961.01710120031004.*

[95] Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16:606–13.

[96] Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007;45:S3–11.

[97] Van der Linden W, Hambleton R. Handbook of Modem Item Response Theory. New York, NY: Springer New York; 1998. *doi:10.1007/978-1-4757-2691-6.*

[98] Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Qual Life Res 2010;19:125–36. *doi:10.1007/s11136-009-9560-5.*

[99] Morrison AS. Screening in Chronic disease (Monographs in Epidemiology and Biostatistics). Oxford University Press; 1992.

[100] Siu AL, US Preventive Services Task Force. Screening for depression in adults. JAMA 2016;315:380–7. *doi:10.1001/jama.2015.18392.*

[101] O'Connor EA, Whitlock EP, Gaynes B, Beil TL. Screening for

depression in adults and older adults in primary care: An updated systematic review. Evid Synth 2009;75:167.

[102] O 'connor E, Rossom RC, Henninger M, Groom HC, Burda BU, Henderson JT, et al. Evidence Synthesis. Screening for Depression in Adults: An Updated Systematic Evidence Review for the U.S. Preventive Services Task Force. 2016.

[103] Le HN, Boyd RC. Prevention of major depression: Early detection and early intervention in the general population. Clin Neuropsychiatry 2006;3:6–22.

[104] Mcquaid JR, Stein MB, McCahill M, Laffaye C, Ramel W. Use of brief psychiatric screening measures in a primary care sample. Depress Anxiety 2000;12:21–9.

[105] Myers JK, Weissman MM. Use of a self-report symptom scale to detect depression in a community sample. Am J Psychiatry 1980;137:1081–4. *doi:10.1176/ajp.137.9.1081*.

[106] Weissman MM, Sholomskas D, Pottenger M, Prusoff BA, Locke BZ. Assessing depressive symptoms in five psychiatric populations: a validation study. Am J Epidemiol 1977;106:203–14.

[107] Williams LM, Rush AJ, Koslow SH, Wisniewski SR, Cooper NJ, Nemeroff CB, et al. International Study to Predict Optimized Treatment for Depression (iSPOT-D), a randomized clinical trial: rationale and protocol. Trials 2011;12:4. *doi:10.1186/1745-6215-12-4.*

[108] Saveanu R, Etkin A, Duchemin A-M, Goldstein-Piekarski A, Gyurak A, Debattista C, et al. The international Study to Predict Optimized Treatment in Depression (iSPOT-D): outcomes from the acute phase of antidepressant treatment. J Psychiatr Res

2015;61:1–12.
*doi:10.1016/j.jpsychires.2014.12.018.*

[109] Arnow BA, Blasey C, Williams LM, Palmer DM, Rekshan W, Schatzberg AF, et al. Depression subtypes in predicting antidepressant response: A report from the iSPOT-D trial. Am J Psychiatry 2015;172:743–50.
*doi:10.1176/appi.ajp.2015.14020181*.

[110] Kendell R, Jablensky A. Distinguishing between the validity and utility of psychiatric diagnoses. Am J Psychiatry 2003;160:4–12.
*doi:10.1176/appi.ajp.160.1.4.*

[111] Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. Lancet 2009;374:609–19.
*doi:10.1016/S0140-6736(09)60879-5.*

[112] Helzer JE, Kraemer HC, Krueger RF. The feasibility and need for dimensional psychiatric diagnoses. Psychol Med 2006;36:1671–80.
*doi:10.1017/S003329170600821X.*

[113] Biesheuvel-Leliefeld KEM, Kok GD, Bockting CLH, de Graaf R, ten Have M, van der Horst HE, et al. Non-fatal disease burden for subtypes of depressive disorder: population-based epidemiological study. BMC Psychiatry 2016;16:139.
*doi:10.1186/s12888-016-0843-4.*

[114] Baxter AJ, Patton G, Scott KM, Degenhardt L, Whiteford HA. Global Epidemiology of Mental Disorders: What Are We Missing? PLoS One 2013;8:e65514.
*doi:10.1371/journal.pone.0065514.*

[115] Pullum TW. An assessment of the quality of data on health and nutrition in the DHS surveys, 1993-2003 . DHS Methodol Reports

No 6 2008.

[116] Patton GC, Coffey C, Cappa C, Currie D, Riley L, Gore F, et al. Health of the world's adolescents: a synthesis of internationally comparable data. Lancet 2012;379:1665–75. *doi:10.1016/S0140-6736(12)60203-7.*

[117] Gottlieb CA, Maenner MJ, Cappa C, Durkin MS, Countries I of MC on NSD in D, Engle P, et al. Child disability screening, nutrition, and early learning in 18 countries with low and middle incomes: data from the third round of UNICEF's Multiple Indicator Cluster Survey (2005–06). Lancet 2009;374:1831–9. *doi:10.1016/S0140-6736(09)61871-7.*

[118] Sackett DL, Haynes B, Guyatt G. Clinical epidemiology : a basic science for clinical medicine. 2nd Ed. Boston: Little, Brown and Company; 1991.

[119] Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002;11:193–205.

[120] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol n.d.;63:737–45. *doi:10.1016/j.jclinepi.2010.02.006.*

[121] Zarin DA, Earls F. Diagnostic decision making in psychiatry. Am J Psychiatry 1993;150:197–206. *doi:10.1176/ajp.150.2.197.*

[122] van Ballegooijen W, Riper H, Cuijpers P, van Oppen P, Smit JH. Validation of online psychometric instruments for common mental

health disorders: a systematic review. BMC Psychiatry 2016;16:45. *doi:10.1186/s12888-016-0735-7.*

[123] Boyd JH, Weissman MM, Thompson WD, Myers JK. Screening for depression in a community sample. Understanding the discrepancies between depression symptom and diagnostic scales. Arch Gen Psychiatry 1982;39:1195–200.

[124] Cuijpers P, Boluijt P, van Straten A. Screening of depression in adolescents through the Internet : sensitivity and specificity of two screening questionnaires. Eur Child Adolesc Psychiatry 2008;17:32–8.

[125] Gandek B, Ware JE. Methods for validating and norming translations of health status questionnaires: The IQOLA Project approach. J Clin Epidemiol 1998;51:953–9. *doi:10.1016/S0895-4356(98)00086-9.*

[126] Gandek B, Ware JE, Aaronson NK, Alonso J, Apolone G, Bjorner J, et al. Tests of data quality, scaling assumptions, and reliability of the SF- 36 in eleven countries: Results from the IQOLA Project. J Clin Epidemiol 1998;51:1149–58. *doi:10.1016/S0895-4356(98)00106-1.*

[127] Wagner AK, Gandek B, Aaronson NK, Acquadro C, Alonso J, Apolone G, et al. Cross-cultural comparisons of the content of SF-36 translations across 10 countries: Results from the IQOLA Project. J Clin Epidemiol 1998;51:925–32. *doi:10.1016/S0895-4356(98)00083-3.*

[128] Ware JE, Gandek B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. J Clin Epidemiol 1998;51:903–12. doi:10.1016/S0895-4356(98)00081-X.

[129] Vilagut G, Ferrer M, Rajmil L, Rebollo P, Permanyer-Miralda G, Quintana JM, et al. El Cuestionario de Salud SF-36 español: una década de experiencia y nuevos desarrollos. Gac Sanit 2005;19:135–50.

[130] Ware JE, Kosinski M, Keller SD. SF-36 Physical and Mental Health Summary Scales : A User's Manual. Boston, MA Heal Institute, New Engl Med Center 1994.

[131] Friedman B, Heisel M, Delavan R. Validity of the SF-36 five-item Mental Health Index for major depression in functionally impaired, community-dwelling elderly patients 2005;53:1978–85. *doi:10.1111/j.1532-5415.2005.00469.x*.

[132] Rumpf H-J, Meyer C, Hapke U, John U. Screening for mental health: validity of the MHI-5 using DSM-IV Axis I psychiatric disorders as gold standard. Psychiatry Res 2001;105:243–53. *doi:10.1016/S0165-1781(01)00329-8*.

[133] Berwick DM, Murphy JM, Goldman PA, Ware JE, Barsky AJ, Weinstein MC. Performance of a five-item mental health screening test. Med Care 1991;29:169–76.

[134] Gill SC, Butterworth P, Rodgers B, Mackinnon A. Validity of the mental health component scale of the 12-item Short-Form Health Survey (MCS-12) as measure of common mental disorders in the general population. Psychiatry Res 2007;152:63–71. *doi:10.1016/j.psychres.2006.11.005*.

[135] Choi SW, Swartz RJ. Comparison of CAT Item Selection Criteria for Polytomous Items. Appl Psychol Meas 2009;33:419–40. *doi:10.1177/0146621608327801*.

[136] Van der Linden WJ, Glas CAW. Computerized adaptive testing: Theory and practice. Boston, MA: Kluwer; 2000.

[137] Forero CG, Vilagut G, Adroher ND, Alonso J, ESEMeD/MHEDEA Investigators. Multidimensional item response theory models yielded good fit and reliable scores for the Short Form-12 questionnaire. J Clin Epidemiol 2013;66:790–801. *doi:10.1016/j.jclinepi.2013.02.007*.

[138] Ware J, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996;34:220–33. *doi:10.2307/3766749.*

[139] Hays RD, Morales LS. The RAND-36 measure of health-related quality of life. Ann Med 2001;33:350–7. *doi:10.3109/07853890109002089.*

[140] Kiely KM, Butterworth P. Validation of four measures of mental health against depression and generalized anxiety in a community based sample. Psychiatry Res 2015;225:291–8. *doi:10.1016/j.psychres.2014.12.023*.

[141] Meader N, Mitchell AJ, Chew-Graham C, Goldberg D, Rizzo M, Bird V, et al. Case identification of depression in patients with chronic physical health problems: a diagnostic accuracy meta-analysis of 113 studies. Br J Gen Pract 2011;61:e808-20. *doi:10.3399/bjgp11X613151.*

[142] Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. J Gen Intern Med 2007;22:1596–602. *doi:10.1007/s11606-007-0333-y.*

[143] Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire

(PHQ-9): a meta-analysis. Can Med Assoc J 2012;184:E191–6. *doi:10.1503/cmaj.110829.*

[144] Martin A, Rief W, Klaiberg A, Braehler E. Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. Gen Hosp Psychiatry 2006;28:71–7. *doi:10.1016/j.genhosppsych.2005.07.003.*

[145] Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009;114:163–73. *doi:10.1016/j.jad.2008.06.026.*

[146] Leeflang MMG, Moons KGM, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clin Chem 2008;54:729–37. *doi:10.1373/clinchem.2007.096032.*

[147] Cook KF, Kallen MA, Bombardier C, Bamer AM, Choi SW, Kim J, et al. Do measures of depressive symptoms function differently in people with spinal cord injury versus primary care patients: the CES-D, PHQ-9, and PROMIS®-D. Qual Life Res 2016:1–10. *doi:10.1007/s11136-016-1363-x.*

[148] Amtmann D, Kim J, Chung H, Bamer AM, Askew RL, Wu S, et al. Comparing CESD-10, PHQ-9, and PROMIS depression instruments in individuals with multiple sclerosis. Rehabil Psychol 2014;59:220–9. *doi:10.1037/a0035919.*

[149] Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN, et al. 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 2009;51:148–80. *doi:10.1111/j.1365-2958.2010.07165.x.Characterization.*

[150] Kroenke K, Yu Z, Wu J, Kean J, Monahan PO. Operating Characteristics of PROMIS Four-Item Depression and Anxiety Scales in Primary Care Patients with Chronic Pain. Pain Med 2014;15:1892–901. *doi:10.1111/pme.12537.*

[151] Pilkonis PA, Yu L, Dodds NE, Johnston KL, Maihoefer CC, Lawrence SM. Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study. J Psychiatr Res 2014;56:112–9. *doi:10.1016/j.jpsychires.2014.05.010.*

[152] Driban JB, Morgan N, Price LL, Cook KF, Wang C. Patient-Reported Outcomes Measurement Information System (PROMIS) instruments among individuals with symptomatic knee osteoarthritis: a cross-sectional study of floor/ceiling effects and construct validity. BMC Musculoskelet Disord 2015;16:253. *doi:10.1186/s12891-015-0715-y.*

[153] Olino TM, Yu L, McMakin DL, Forbes EE, Seeley JR, Lewinsohn PM, et al. Comparisons across depression assessment instruments in adolescence and young adulthood: An item response theory study using two linking methods. J Abnorm Child Psychol 2013;41:1267–77. *doi:10.1007/s10802-013-9756-6.*

[154] Schalet BD, Cook KF, Choi SW, Cella D. Establishing a Common Metric for Depressive Symptoms: Linking the BDI-II, CES-D, and

PHQ-9 to PROMIS Depression. Psychol Assess 2014;26:88-513–27.
*doi:10.1016/j.janxdis.2013.11.006.*

[155] Gibbons LE, Feldman BJ, Crane HM, Mugavero M, Willig JH, Patrick D, et al. Migrating from a legacy fixed-format measure to CAT administration: calibrating the PHQ-9 to the PROMIS depression measures. Qual Life Res Qual Life Res 2011;20:1349–57.
*doi:10.1007/s11136-011-9882-y.*

[156] Akobeng AK. Understanding diagnostic tests 3: receiver operating characteristic curves. Acta Paediatr 2007;96:644–7.
*doi:10.1111/j.1651-2227.2006.00178.x.*

[157] Fischer HF, Klug C, Roeper K, Blozik E, Edelmann F, Eisele M, et al. Screening for mental disorders in heart failure patients using computer-adaptive tests. Qual Life Res 2014;23:1609–18.
*doi:10.1007/s11136-013-0599-y.*

[158] Vilagut G, Forero CG, Adroher ND, Olariu E, Cella D, Alonso J, et al. Testing the PROMIS® Depression measures for monitoring depression in a clinical sample outside the US. J Psychiatr Res 2015;68:140–50.
*doi:10.1016/j.jpsychires.2015.06.009.*

[159] Olariu E, Castro-Rodriguez J-I, Álvarez P, Garnier C, Reinoso M, Martín-López LM, et al. Validation of clinical symptom IRT scores for diagnosis and severity assessment of common mental disorders. Qual Life Res 2015;24:979–92.
*doi:10.1007/s11136-014-0814-5.*

[160] Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. 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 2010;63:1179–94. *doi:10.1016/j.jclinepi.2010.04.011.*

[161] Bajaj JS, Thacker LR, Wade JB, Sanyal a J, Heuman DM, Sterling RK, et al. PROMIS computerised adaptive tests are dynamic instruments to measure health-related quality of life in patients with cirrhosis. Aliment Pharmacol Ther 2011;34:1123–32. *doi:10.1111/j.1365-2036.2011.04842.x.*

[162] Brown T a, Chorpita BF, Barlow DH. Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. J Abnorm Psychol 1998;107:179–92. *doi:10.1037/0021-843X.107.2.179.*

[163] Simms LJ, Grös DF, Watson D, O'Hara MW. Parsing the general and specific components of depression and anxiety with bifactor modeling. Depress Anxiety 2008;25:E34-46. *doi:10.1002/da.20432.*

[164] Gomez R, McLaren S. The Center for Epidemiologic Studies Depression Scale: Support for a Bifactor Model With a Dominant General Factor and a Specific Factor for Positive Affect. Assessment 2015;22:351–60. *doi:10.1177/1073191114545357.*

[165] Den Hollander-Gijsman ME, Wardenaar KJ, De Beurs E, Van Der Wee NJA, Mooijaart A, Van Buuren S, et al. Distinguishing symptom dimensions of depression and anxiety: An integrative approach. J Affect Disord 2012;136:693–701. *doi:10.1016/j.jad.2011.10.005.*

[166] Nitschke JB, Heller W, Imig JC, McDonald RP, Miller GA. Distinguishing Dimensions of Anxiety and Depression. Cognit Ther Res 2001;25:1–22. *doi:10.1023/A:1026485530405.*

[167] Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: Review and meta-analysis. Am J Psychiatry 2000;157:1552–62. *doi:10.1176/appi.ajp.157.10.1552.*

[168] Kalueff A V., Tuohimaa P. Experimental modeling of anxiety and depression. Acta Neurobiol Exp (Wars) 2004;64:439–48.

[169] Keller J, Nitschke JB, Bhargava T, Deldin PJ, Gergen J a, Miller G a, et al. Neuropsychological differentiation of depression and anxiety. J Abnorm Psychol 2000;109:3–10. *doi:10.1037/0021-843X.109.1.3.*

[170] Craske MG. Transdiagnostic treatment for anxiety and depression. Depress Anxiety 2012;29:749–53. *doi:10.1002/da.21992.*

[171] Widiger TA, Clark LA. Toward DSM V and the classification of psychopathology. Psychol Bull 2000;126:946–63. *doi:10.1037/0033-2909.126.6.946.*

[172] Pereda N, Forero CG. Contribution of Criterion A2 to PTSD Screening in the Presence of Traumatic Events. J Trauma Stress 2012;25:587–91. *doi:10.1002/jts.21736.*

[173] Lord, Frederic M; Novick MR. Statistical Theories of Mental Test Scores. Reading, Mass: Addison-Wesley; 1968.

[174] Gibbons RD, Immekus JC, Darrell Bock R, Gibbons Director RD.

Multi-dimensional and hierarchical modeling monograph 1 The Added Value of Multidimensional IRT Models Multi-dimensional and hierarchical modeling monograph 2 2007.

[175] Bue J, Ae B, Chang C-H, Thissen D, Reeve BB. Developing tailored instruments: item banking and computerized adaptive assessment 2007.
*doi:10.1007/s11136-007-9168-6.*

# 7. ANNEX 1. Supplementary material for article 3

Supplementary material for article:

Vilagut G, Forero CG, Barbaglia G, Alonso J. Screening for Depression in the General Population with the Center for Epidemiologic Studies Depression (CES-D): A Systematic Review with Meta-Analysis. PLoS One. 2016;11(5): e0155431.

# 8. ANNEX 2: Supplementary material for article 4

Supplementary material for article:

Vilagut G, Forero CG, Castro-Rodriguez JI, Olariu E, Barbaglia G, Astals M, Diez-Aja C, Gárriz M, Abellanas L, Alonso J, on behalf of the PROMIS.es Investigators. PROMIS Depression Item Bank Showed Measurement Equivalence between Spain and the US. J Clin Epidemiol (under review)

**Supplementary material**

**Supplementary table 1. Design of the Interviews**

| | | Questionnaire Sections | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **PROMIS Depression** | **PROMIS Anxiety** | **PROMIS Anger** | **CES-D** | **BAI** | **STAXI-2** | **HRQOL** | **PHQ-9** | **GAD-7** | **WHODAS** | **Demographic variables** |
| Number of items | Number of Individuals | 28 | 29 | 22 | 20 | 21 | 39 | 12 | 10 | 7 | 12 | 8 |
| 116 | 300 | x | x | x | | | | x | x | x | | x |
| 104 | 300 | x | x | x | | | | | x | x | | x |
| 116 | 300 | x | x | x | | | | | x | x | x | x |
| 106 | 300 | x | x | | x | x | | | | | | x |
| 117 | 300 | x | | x | x | | x | | | | | x |
| 119 | 300 | | x | x | | x | x | | | | | x |
| | **1800** | **1500** | **1500** | **1500** | **600** | **600** | **600** | **300** | **900** | **900** | **300** | **1800** |

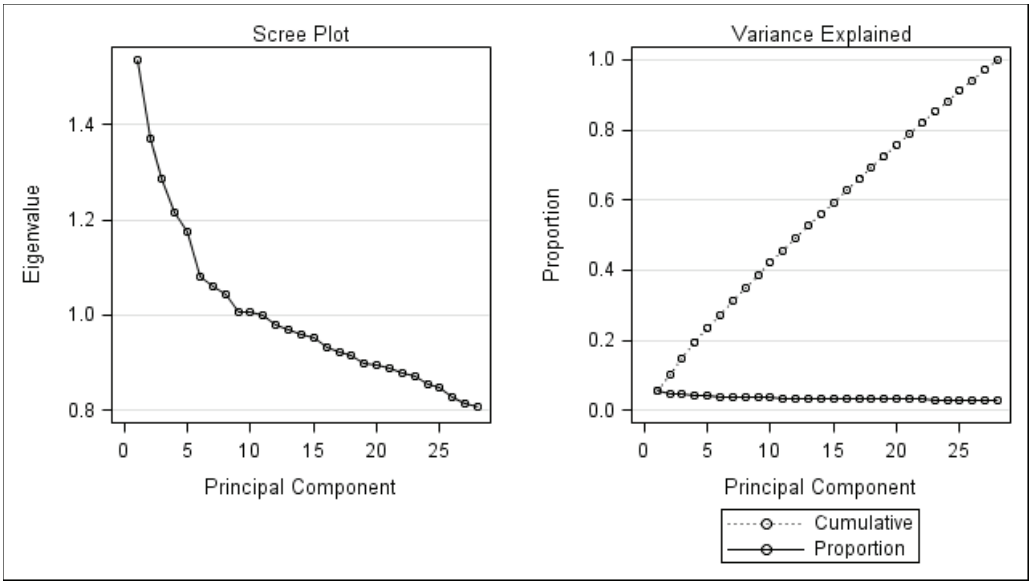
**Supplementary table 2. Characteristics of the sample. The Spanish PROMIS Depression Panel data. (year 2015)**

| | | n | Unweighted % | Weighted % | Population distribution* |
|---|---|---|---|---|---|
| Age, | Mean (SD) | | 48.5 (14.8) | 48.2 (16.0) | 49.2 (25.4) |
| age groups | 18-34 | 322 | 21.4% | 25.0% | 25.0% |
| | 35-49 | 452 | 30.1% | 30.5% | 30.4% |
| | 50-64 | 497 | 33.1% | 22.8% | 22.7% |
| | 65+ | 232 | 15.4% | 21.6% | 22.0% |
| sex | Male | 712 | 47.4% | 48.6% | 48.7% |
| | Female | 791 | 52.6% | 51.4% | 51.3% |
| Marital Status | Married/living with someone | 1063 | 70.7% | 69.0% | |
| | Single | 310 | 20.6% | 21.4% | |
| | Separated/divorced | 78 | 5.2% | 5.4% | |
| | Widow | 52 | 3.5% | 4.2% | |
| Employment situation | Working | 818 | 54.4% | 51.0% | |
| | Not working | 173 | 11.5% | 10.9% | |
| | homework | 109 | 7.3% | 7.6% | |
| | Student | 85 | 5.7% | 6.6% | |
| | Disabled | 32 | 2.1% | 1.6% | |
| | Retired | 258 | 17.2% | 20.4% | |
| | Other | 28 | 1.9% | 1.8% | |
| Education | Incomplete Primary | 33 | 2.2% | 3.0% | |
| | Complete primary | 91 | 6.1% | 6.1% | |
| | Incomplete secondary | 224 | 14.9% | 14.4% | |
| | Complete secondary | 574 | 38.2% | 38.4% | |
| | University degree | 573 | 38.1% | 37.6% | |
| | other | 8 | 0.5% | 0.5% | |

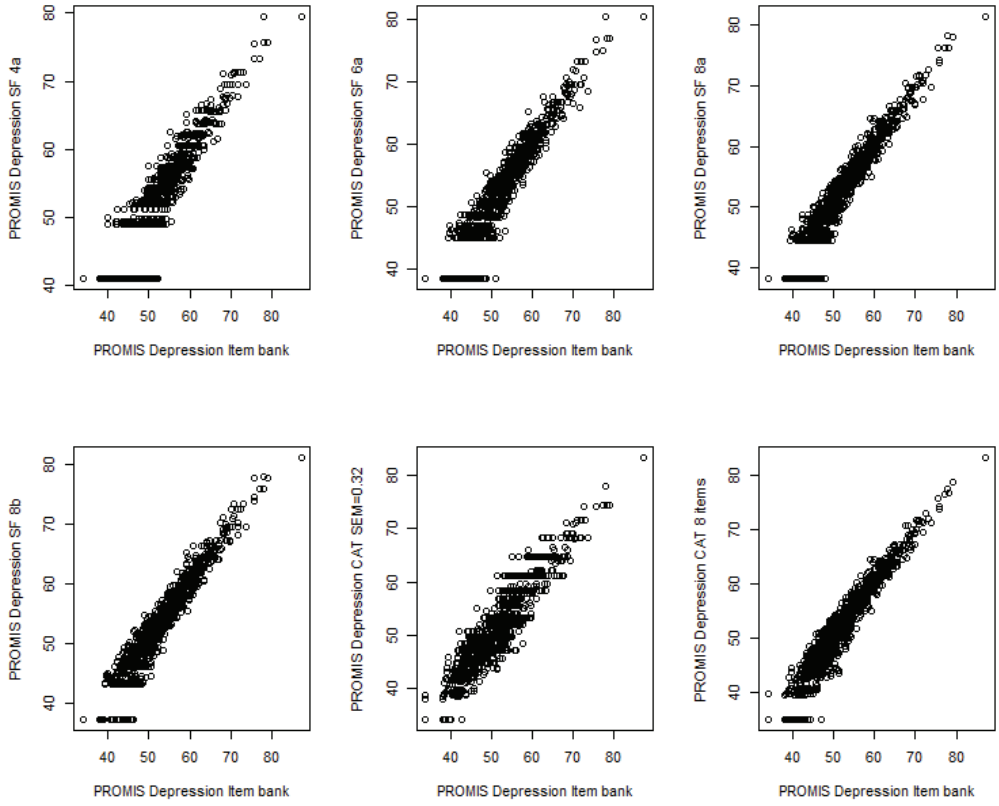
*Padrón de Municipios 2014

**Supplementary table 3. Pseudo $R^2$ and Relative change of coefficient statistics for DIF analysis by language (Spanish versus English)**

| | Pseudo R2 (McFadden) | | | Relative $\beta_1$ Change |
|---|---|---|---|---|
| **Item** | **Uniform DIF** | **Non-uniform DIF** | **Overall DIF** | |
| EDDEP04. I felt worthless | 0.001 | 0.003 | 0.005 | 0.004 |
| EDDEP05. I felt that I had nothing to look forward to | 0.011 | 0.003 | 0.014 | 0.024 |
| EDDEP06. I felt helpless | 0.005 | <0.001 | 0.005 | 0.012 |
| EDDEP07. I withdrew from other people | <0.001 | 0.001 | 0.001 | <0.001 |
| EDDEP09. I felt that nothing could cheer me up | 0.007 | <0.001 | 0.007 | 0.02 |
| EDDEP14. I felt that I was not as good as other people | <0.001 | <0.001 | <0.001 | <0.001 |
| EDDEP17. I felt sad | 0.003 | <0.001 | 0.003 | 0.008 |
| EDDEP19. I felt that I wanted to give up on everything | 0.010 | <0.001 | 0.010 | 0.027 |
| EDDEP21. I felt that I was to blame for things | 0.002 | 0.0002 | 0.002 | 0.003 |
| EDDEP22. I felt like a failure | <0.001 | <0.001 | <0.001 | 0.001 |
| EDDEP23. I had trouble feeling close to people | 0.015 | <0.001 | 0.015 | 0.028 |
| EDDEP26. I felt disappointed in myself | 0.032 | 0.002 | 0.035 | 0.068 |
| EDDEP27. I felt that I was not needed | 0.001 | <0.001 | 0.001 | <0.001 |
| EDDEP28. I felt lonely | 0.002 | 0.001 | 0.003 | 0.003 |
| EDDEP29. I felt depressed | 0.001 | 0.001 | 0.001 | 0.001 |
| EDDEP30. I had trouble making decisions | 0.002 | 0.004 | 0.006 | 0.005 |
| EDDEP31. I felt discouraged about the future | 0.002 | 0.002 | 0.004 | 0.005 |
| EDDEP35. I found that things in my life were overwhelming | <0.001 | <0.001 | <0.001 | <0.001 |
| EDDEP36. I felt unhappy | 0.001 | 0.003 | 0.005 | 0.003 |
| EDDEP39. I felt I had no reason for living | 0.004 | <0.001 | 0.004 | 0.014 |
| EDDEP41. I felt hopeless | 0.004 | 0.002 | 0.006 | 0.011 |
| EDDEP42. I felt ignored by people | 0.006 | <0.001 | 0.006 | 0.010 |
| EDDEP44. I felt upset for no reason | <0.001 | 0.001 | 0.001 | <0.001 |
| EDDEP45. I felt that nothing was interesting | 0.009 | 0.010 | 0.019 | 0.015 |
| EDDEP46. I felt pessimistic | 0.004 | 0.001 | 0.005 | 0.007 |
| EDDEP48. I felt that my life was empty | 0.008 | 0.001 | 0.009 | 0.016 |
| EDDEP50. I felt guilty | 0.003 | 0.001 | 0.004 | 0.006 |
| EDDEP54. I felt emotionally exhausted | 0.004 | <0.001 | 0.004 | 0.007 |




**Supplementary figure 1. Results of principal components analysis of residual correlations from a one-factor Confirmatory factor analysis of PROMIS depression item bank**




**Supplementary figure 2. Comparison of scores for full item bank with short forms and simulated CATs.**

# 9. ANNEX 3: Supplementary material for article 4

Supplementary material for article:

Vilagut G, Forero CG, Adroher ND, Olariu E, Cella D, Alonso J, on behalf of the INSAyD investigators. Testing the PROMIS® Depression measures for monitoring depression in a clinical sample outside the US. J Psychiatr Res. 2015; 68:140–50.