Computerized Adaptive Testing in Dutch Mental Health Care
A new tool to assess Depression and Anxiety

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 5 oktober 2022
klokke 11.15 uur

door

Gerard Slok - Flens

Promotores:

Prof.dr. Edwin de Beurs

Prof.dr. Philip Spinhoven

Co-promotor:

Dr. Niels Smits

Promotiecommissie:

Prof.dr. Sander T. Nieuwenhuis (voorzitter)

Prof.dr. Bernet M. Elzinga

Prof.dr. Rob R. Meijer (University of Groningen)

Prof.dr.ir. Anne M. Stiggelbout

Dr. Kim de Jong

Dr. Muirne C. S. Paap (University of Groningen)

This dissertation was supported by the Foundation for benchmarking mental health care (Stichting Benchmark GGZ; SBG) and the Quality alliance mental health (Alliantie kwaliteit in de geestelijke gezondheidszorg; Akwa GGZ).

Table of Contents

Chapter 1

Introduction

1.1 The need for good measurement instruments in Dutch mental health care

In the Netherlands, two large psychiatric epidemiological population studies were conducted between 1996 and 1999 (NEMESIS; Bijl, van Zessen, Ravelli, de Rijk, & Langendoen, 1998; Bijl, Ravelli, & van Zessen, 1998) and between 2007 and 2018 (NEMESIS-2; de Graaf, ten Have, & van Dorsselaer, 2010, 2012; www.trimbos.nl/kennis/feiten-cijfers-ggz-nederland/nemesis-2). The results showed that more than 4 out of 10 persons have had one or more mental health disorders in their life. Moreover, the prevalence rates of the most common mental health disorders remained highly similar over time. These results lead to an alarming conclusion: mental health disorders occur frequently in the Netherlands with no clear change over time. Consequently, we have to keep looking for new solutions to reduce the large Dutch mental health problem.

Every year, approximately 6% of the Dutch population is treated by clinicians for their mental health problems (de Beurs, Barendregt, & Warmerdam, 2017). The main goal of these clinicians is treating patients effectively and efficiently: patients should recover as much as possible, as sustainably as possible, in as little time as possible. To aid clinicians and patients in achieving this goal, several tools are available to them. One of these tools concerns measurement instruments. Measurement instruments used in mental health care consist of a set of questions that assesses a patient on one or several (mental health) indicators (i.e., constructs), such as depression, anxiety, the ability to participate in social roles and activities, health related quality of life, or treatment goal attainment. By administering these instruments periodically (also known as Routine Outcome Monitoring [ROM]; Carlier et al., 2012a; de Beurs et al., 2011), clinicians are aided in clinical decision-making concerning the patient’s diagnosis, treatment selection and termination, treatment of nonresponders, and relapse prevention (de Beurs et al., 2018; Greenhalgh et al., 2018; Lambert, 2010; Lewis et al., 2015; Martin-Cook et al., 2021). As a result, patients’ motivation to continue treatment may be increased, and patients’ treatment outcomes may be improved (de Jong et al., 2021; Fortney et al., 2018; Guo et al., 2015; Rush & Thase, 2018; Scott & Lewis, 2015). Moreover, if the use of measurement instruments is combined with *shared-decision-making*, in which patients are supported to participate in the decisions concerning treatment, patient’s treatment outcomes may be further improved (Metz et al., 2019; van der Feltz-Cornelis et al., 2014). Finally, aggregation of an instrument’s scores allow for comparisons between groups, and, when combined with data of patient characteristics and treatment process aspects, aggregated data can be used to improve the overall quality and value of care for patients (de Beurs et al., 2018; Porter, 2009).

Measurement instruments used in mental health care should preferably meet two criteria before they are used to aid clinicians and patients in treatment. First, an instrument should have good psychometric properties (Maruyama & Ryan, 2014). This means that the instrument measures reliably, validly, and responsively. An instrument is considered reliable when it produces consistent results over replications (Crocker & Algina, 1986); valid when it adequately measures the concept that is supposed to be measured (Cook & Campbell, 1979); and responsive when it sufficiently detects change over time in the construct to be measured

(Mokkink et al., 2010). In addition, the second criterium is that an instrument should be efficient. This means that the number of questions (i.e., the items) is as small as possible to burden patients as little as possible.

In Dutch mental health care, there is a need for efficient measurement instruments with good psychometric properties. Due to the growing recognition of ROM, more patients are being asked to respond to instruments, and more instruments are being administered to each patient (de Beurs, Barendregt, & Warmerdam, 2017). The currently used instruments, however, have three limitations that can make it difficult to obtain complete and high-quality information from patients. First, it is unclear for many instruments whether their quality is sufficient because relevant psychometric properties have not been studied. For example, it is known for only a few instruments whether patients interpret the items in the same way over time (e.g., Symptom Questionnaire [SQ-48], Carlier et al., 2019; Outcome Questionnaire [OQ-45], Jabrayilov, Emons, de Jong, & Sijtsma, 2017; Short Health Anxiety Inventory [SHAI], te Poel, Hartmann, Baumgartner, & Tanis, 2017). To clarify the relevance of this psychometric property, also known as longitudinal measurement invariance (LMI; Fokkema et al., 2013; Liu et al., 2017; Oort, 2015; Sawatzky et al., 2021; Soland, 2021; Verdam et al., 2021), consider the following example. During treatment, patients may receive psychoeducation. In this form of therapy, clinicians explain to patients what their disorder comprises, and that many of their experienced symptoms are part of that disorder. A consequence of this therapy may be that patients change regarding their understanding or awareness of the behaviors and symptoms that constitute their disorder. When this happens, patients' frame of reference may also change, possibly resulting in altered response-behavior to items that does not reflect change in the measured construct. What it does reflect, however, is that the measured construct itself has changed over time. In other words, score comparisons over time may be biased because the measured construct changed between measurements. This can lead to wrong inferences about patients' treatment progress (Fokkema et al., 2013; Fried et al., 2016).

Second, the instruments currently used in Dutch mental health care use fixed item sets, which makes it challenging for measurement to be both highly reliable and highly efficient. To illustrate this, consider that fixed item sets can make patients respond to items that may not be relevant for their severity level. For example, administering the item *I felt sad* to measure depression may not be relevant for patients whose previous item scores point to a severe mood disorder because feelings of sadness will obviously be present. The instrument, however, needs to measure patients with mild mood disorders reliably too. For these patients, an item about sad feelings can be relevant. This implies that using a fixed item set to reliably measure respondents with a wide range of severity levels may only succeed when the number of items (and the items' coverage of the measured construct) is sufficient. In other words, reliable measurement is achieved at the cost of efficient measurement. Administering a fixed item set with too few items, on the other hand, may lead to insufficiently reliable scores because relevant items are missing. In this case, efficient measurement is achieved at the cost of reliable measurement.

In Dutch clinical practice, the currently used instruments show a large variation in the number of administered items. Some instruments administer less than 10 items while others comprise over 50. As a result, the total number of administered items can become quite substantial with an increasing number of administered instruments. This, in turn, may lead patients to rush through the administered items, affecting the reliability of the item responses, or even ignore the request to respond to instruments at all. Also, clinicians may be forced to measure less constructs to lessen the burden on patients even though they deem the excluded constructs relevant for treatment evaluation. When clinicians do choose to measure all relevant constructs, they could deliberately select shorter instruments to limit the burden on patients, affecting the reliability of measurement. For these reasons, instruments should preferably be both highly efficient and highly reliable to increase the probability of obtaining complete and high-quality information from patients.

A final limitation of the currently used Dutch mental health instruments concerns the large variety used to measure the same construct(s). For example, the construct *general psychopathology* is measured, among others, with the 90-item Symptom Checklist (SCL-90; Arrindell & Ettema, 1981), the 53-item Brief Symptom Inventory (BSI; de Beurs & Zitman, 2005), the 45-item OQ-45 (de Jong et al., 2007), the 48-item SQ-48 (Carlier et al., 2012b), and the 13-item Short List of Complaints (SLC; Lange, Schrieken, van de Ven, & Blankers, 2000). Instruments intended to measure the same construct(s), however, always differ to some degree in their item content, response categories, and psychometric properties. Such differences have raised doubts whether instruments can indeed measure the same construct(s) adequately, even when the scores are expressed on the same scale (de Beurs et al., 2012; de Beurs, Barendregt, & Warmerdam, 2017). Consequently, it may be challenging for clinicians to compare treatment outcomes and learn from each other’s practice when using different instruments. To solve this issue, clinicians should preferably use the same set of instruments in their practice. This generally applies only to clinicians from different mental health providers as clinicians within the same provider often use the same instruments.

In this thesis, we aim to lay the foundation for a new set of Dutch mental health instruments that solve the discussed limitations of the currently used instruments by using modern methodologies. More specifically, we aim to contribute to the psychometric evaluation of new instruments that are investigated on a wide collection of psychometric properties, and that measure efficiently, reliably, validly, and responsively. These instruments should not only lead to an increase in the completeness of information deemed relevant to evaluate patients’ treatment, they should also lead to more high-quality information that reduces the probability of a clinician making biased inferences. For these reasons, a new set of measurement instruments may have the potential to be the new standard in the Netherlands for evaluating patients’ treatment. In this case, clinicians from different mental health providers can compare and learn from each other’s treatment outcomes more easily without having to account for possible bias because they use different instruments. Ultimately, a new set of high-quality measurement instruments may lead to more effective and efficient treatment of patients in general, assuming clinicians and patients make good use of the measurement information. This improved effectiveness and efficiency, in turn, may even lead to mental health care providers having more capacity to help other persons recover from their mental health problems.

As starting point for the new set of Dutch mental health instruments, this thesis will focus on the psychometric evaluation of two instruments that measure depression and anxiety symptoms. These symptoms were chosen because depression and anxiety disorders are the most common mental health problems in the Dutch population (de Graaf, ten Have, van Gool, & van Dorsselaer, 2012), and a worldwide problem in general (Baxter, Scott, Vos, & Whiteford, 2013; Marcus, Yasamy, van Ommeren, Chisholm, & Saxena, 2012). Furthermore, symptoms of depression and anxiety are present in other mental disorders as well, and are usually the prime reason to seek mental health care treatment (Frank, 1974; Clarke & Kissane, 2002).

This introduction is continued with section 1.2 in which the psychometric properties are discussed that are relevant for measurement instruments used for patients' treatment evaluation. Computerized adaptive testing (CAT) is also introduced in this section, which is a promising methodology that will be adopted for the new set of instruments. This methodology is then discussed in further detail in section 1.3. Finally, an outline of the thesis is provided in section 1.4.

1.2 Psychometric properties

1.2.1 Reliability

A common definition of reliability is the consistency (or reproducibility) of a person's score over replications (Crocker & Algina, 1986). To illustrate this definition, consider that a person completes an instrument measuring depression several times in a row under consistent conditions (e.g., the person does not change in the measured construct). The instrument is said to measure depression reliably when it shows similar results among measurements. If the instrument, however, results in depression scores that are substantially different each measurement, the instrument is not of much use. In this case, the instrument is said to measure depression unreliably.

A common reason for unreliable measurement of mental health constructs is that respondents make simple “mistakes” due to fatigue, inattention, or feeling sick (Maruyama & Ryan, 2014). Also, a longer test is generally more reliable than a shorter test, assuming all else being equal (e.g., the item response categories and the overlapping items; Bernstein & Nunnally, 1994). To illustrate this, consider that a fatigued respondent provides a somewhat lower item response on one item and a somewhat higher item response on another. In this case, the administration of multiple items ensures that the effects of unreliable measurement cancel each other out. The more items are administered, the more likely it may be that effects of unreliable measurement cancel each other out, the higher the reliability of measurement. On the other hand, administering *too many* items can also lead to an increase in the degree of unreliability because the respondent is becoming increasingly more tired. Therefore, instruments should preferably have a good balance between reliable measurement and efficient measurement.

To assess the reliability of measurement, it first has to be decided how reliability is treated. This depends on the measurement theory that is adopted in the development stage of the instrument. A measurement theory is a specific paradigm that is chosen for the design, analysis, and scoring of an instrument to estimate and interpret a person's score on one or several constructs. In other words, a measurement theory comprises a set of assumptions that are needed to relate a person's score on one or several measured constructs to his actual experience regarding those constructs.

Most instruments have been developed under the measurement theory *classical test theory* (CTT; Lord & Novick, 1968). In CTT, reliability is treated as a common estimate that is assumed to be equal for all individuals irrespective of their construct level (Jabrayilov, Emons, & Sijtsma, 2016). This means that the reliability of measurement is always considered to be the same between and within persons of a specific population. In addition, instruments can also be developed using a modern measurement theory that is less familiar among clinicians: *item response theory* (IRT; Embretson & Reise, 2000). In IRT, reliability is not considered a common estimate, but a specific estimate that is assumed to depend on a person's construct score (also known as the latent trait level). This means that the reliability of measurement can actually be different between and within persons because latent trait levels can be different (Jabrayilov, Emons, & Sijtsma, 2016).

IRT is slowly gaining popularity in the field of instrument development, partly because this measurement theory can be applied more easily to develop instruments that are both efficient and reliable. When adopting CTT, instruments commonly need to administer a fixed item set that is equal for all measurements. In this way, a construct score can simply be obtained by summing or averaging the item responses, a higher number of items resulting in a higher common reliability estimate, assuming all else being equal. However, it also means that instruments cannot be made shorter without a decrease in reliability for all measurements. By contrast, instruments adopting IRT do not have to administer all of the items to estimate a person's latent trait level. Consequently, the number of administered items can be different between measurements without the reliability of measurement having to vary substantially.

In addition to the possibility of administering *a different number* of items, instruments adopting IRT can use the item responses of previous administered items to select and administer only *the most appropriate* items. With computerized adaptive testing (CAT; Embretson & Reise, 2000), a new item is selected and administered from a set of items, usually in such a way that the reliability of measurement will increase as much as possible. For example, when previous item responses point to a severe anxiety level, a new item that mostly covers lower anxiety levels, such as *I felt worried*, will likely not contribute much to a latent trait estimate with a higher reliability level. To accomplish this, it is preferable to administer an item that covers higher anxiety levels, such as *I felt terrified*. This dynamic selection of items can then continue for as long as needed to estimate a person's latent trait level with sufficient reliability. As a result, measurement can be both reliable and efficient because the item administration is tailored to each patient's latent trait level, and stops as soon as a sufficient reliability level has been reached.

CAT has yet to be implemented at large scale in Dutch clinical practice, but this may be about to change with the introduction of the patient-reported outcomes measurement information system (PROMIS$^{®}$; Cella et al., 2007, 2010). This introduction is now first continued with a discussion of the psychometric properties validity and responsiveness before returning to CAT and PROMIS in section 1.3 to elaborate on these promising developments in further detail.

1.2.2 Validity

A classic definition of (construct) validity is the degree to which a test measures the concept it is supposed to measure (Cook & Campbell, 1979). The validity of an instrument can be studied by collecting sources of evidence commonly claimed as indicative of validity (Maruyama & Ryan, 2014; Newton & Shaw, 2014). For example, it can be studied whether an instrument corresponds to the theorized construct based on face value (also known as face validity) or a systematic review of the content (also known as content validity). Also, empirical sources of evidence can be studied to evaluate whether an instrument is able to relate to or predict some criterion. Three common examples of such sources of evidence are convergent validity, divergent validity, and concurrent validity. Convergent validity is studied to evaluate whether an instrument relates to another instrument that it is expected to relate to; divergent validity to evaluate whether an instrument does not relate to another instrument that it is not expected to relate to; and concurrent validity to evaluate whether an instrument is able to distinguish between groups that it is expected to distinguish.

In addition to the sources of evidence presented in the previous paragraph, the complete psychometric toolbox currently available to us contains even more statistical tools to evaluate additional assumptions on the validity of measurement. These tools include (uni)dimensionality, local independence (LI), monotonicity, and measurement invariance (Reeve et al., 2007; Vanderberg & Lance, 2000; Fokkema et al., 2013). To begin with (uni)dimensionality, this is the assumption that persons' item responses are the result of their level on the measured construct(s) and not of their level on other constructs. In addition, LI and monotonicity are tools that apply specifically to instruments adopting IRT as measurement theory. LI is the assumption that the item responses on an item pair show no association when controlling for the latent trait level. This means that the construct of interest, and no other constructs, explains why the items are related to each other. Monotonicity is the assumption that the probability of selecting an item response which suggests a higher latent trait level increases as the person's latent trait level is higher. This means that the item responses are related to the latent trait level of the measured construct, and not to the latent trait level of other constructs.

Finally, measurement invariance, which is related to differential item functioning (DIF; Reeve et al., 2007), can be investigated to assess whether the measured construct is sufficiently similar between subgroups, such as men and women, or within (sub)groups over time (Fokkema et al., 2013; Oort, 2015; Vanderberg & Lance, 2000). In section 1.1, we already provided an example of the latter by pointing out that patients' constructs may change over the course of therapy. An example of the former can be illustrated with a measure of depression that includes an item about crying. Consider that a woman indicates that she cries more often than a man even though they have the same level of depression. This may mean that women cry more easily than men if they are upset, but it does not have to mean that they are also more depressed. Consequently, items about crying may not be suited as a valid measure of depression because they may also measure an additional construct such as the willingness to express emotions openly, on which woman may score higher.

There are four types of measurement invariance and two types of DIF that can be evaluated (Liu et al., 2017; Reeve et al., 2007). These types assess whether (a) the construct is measured by the same content, (b) the item scores are similarly related to the construct (also known as absence of nonuniform DIF), (c) respondents with a similar score on the construct choose the same item response categories (also known as absence of uniform DIF), and (d) the items measure the construct with the same reliability level. For (Dutch) mental health instruments, measurement invariance is sometimes evaluated between demographic subgroups and only occasionally within clinical groups over time (e.g., Carlier et al., 2019; Jabrayilov, Emons, de Jong, & Sijtsma, 2017; te Poel, Hartmann, Baumgartner, & Tanis, 2017). Consequently, LMI is likely assumed for many instruments without having been studied. This, in turn, may increase the probability of a test user making biased inferences when interpreting an instrument's scores.

1.2.3 Responsiveness

A definition of responsiveness is the ability of an instrument to detect change over time in the construct to be measured (Mokkink et al., 2010). The higher the responsiveness, the more an instrument can discriminate between persons who differ only little in their degree of change on the measured construct. Consequently, it is preferred to evaluate patients' treatment with highly responsive instruments because subtle improvements and deteriorations are detected better.

Studies that investigated the responsiveness of Dutch mental health instruments suggest that instruments often differ in their degree of responsiveness (e.g., Carlier et al., 2017; de Beurs et al., 2012; de Beurs, Barendregt, & Warmerdam, 2017). This may pose a problem for mental health care providers that use different instruments and want to learn from each other's practice and outcomes (de Beurs et al., 2018). The comparison of treatment outcomes with different instruments is only justified when they sufficiently measure the same construct and are similarly responsive (de Beurs, Barendregt, & Warmerdam, 2017). As these criteria are often not met, comparing scores between groups may easily result in invalid inferences. Consequently, it may be best to use a uniform set of highly responsive instruments when comparing treatment outcomes. In this thesis, we want to contribute to the use of a uniform set of instruments by evaluating high-quality instruments that are not only highly responsive, but also highly reliable, efficient, and valid.

1.3 Computerized Adaptive Testing (CAT)

In section 1.2.1, CAT was introduced as a solution to the challenge of developing measurement instruments that are both efficient and reliable. In this section, CAT is discussed in further detail. First, six CAT components are described in section 1.3.1 to gain more understanding in this promising IRT-application. Then, both international CAT developments (section 1.3.2) and national CAT developments (section 1.3.3) are discussed that will be built upon in this thesis.

1.3.1 CAT components

A CAT instrument is a computer-based test that uses one or more item banks and an algorithm to select and administer items based on the answers to previous administered items, and that terminates when a specific stopping rule is met. An item bank is a set of items that measures a latent trait such as depression or anxiety. To measure the latent trait to its full extent (i.e., from light to severe problems), the item bank needs to include items that allow for precise measurement of a wide range of latent trait levels. Therefore, the number of items may preferably be substantial.

The CAT algorithm usually consists of five components: (a) the item parameters, (b) a starting item of the CAT administration, (c) a method to estimate the latent trait level and the accompanying measurement precision after each collected item response, (d) a rule to terminate the administration of new items, and (e) a method to select and administer new items (Smits et al., 2012). First, for every item in the item bank, a set of item parameters is estimated using an appropriate IRT-model and a sample of the population that will be measured by the instrument. The estimated item parameters are used by the CAT instrument to select and administer items that are tailored to the respondent's latent trait estimate, based on the responses to previously administered items (Reeve et al., 2007). For latent traits concerning mental health, two types of parameters are usually estimated: the item slope (also known as the discrimination parameter) and the item difficulty (also known as the threshold parameter). Each item has one discrimination parameter that expresses the extent to which persons with similar latent trait levels can be differentiated by the item. In addition, an item has one or more threshold parameters (the number is equal to the number of response categories minus one) that express the latent trait level(s) on which a person is expected to choose a higher item response category over a lower item response category (Smits et al., 2012).

Second, each CAT administration starts by presenting a first item. The starting item can be selected at random, but usually it is chosen in such a way that it has the highest information value for persons with an average latent trait level. As a result, the initial latent trait estimate is as precise as possible for as many respondents as possible, assuming that no previous information is available on the respondent's latent trait level.

Third, after a response is given on the first item, the respondent's latent trait level is estimated along with the accompanying measurement precision. The latent trait level can be estimated with two types of scoring methods: maximum likelihood (ML) and Bayesian estimation (Embretson & Reise, 2000). Both methods have their strengths and limitations (Penfield & Bergeron, 2005; Wang & Vispoel, 1998). Bayesian estimation is chosen more often because it ensures that a latent trait level can be estimated for all response patterns. A drawback, however, is that the latent trait estimate is pulled somewhat toward the center of the distribution, which may result in bias. ML, by contrast, is not able to estimate a latent trait level for response patterns that only consist of either the lowest or the highest response categories. It may be, however, a more stable estimator considering possible bias.

Fourth, after the respondent's latent trait level and measurement precision are estimated, the CAT algorithm evaluates the results against a stopping rule to determine whether the administration of new items can be terminated. Commonly used stopping rules are a specific measurement precision, an upper limit number of items, or a combination of the two (Smits et al., 2012). For CAT instruments, the latter may be most suitable when the goal is to measure both reliably and efficiently. Assuming an adequate item bank with good psychometric properties, CAT will probably measure a wide range of latent trait levels sufficiently precise with a small number of items. It may be more challenging, however, to reliably measure mental health constructs in persons with a rather high or low latent trait level (Reise & Waller, 2009). The reason for this is that item banks generally do not include a sufficient number of items that discriminate well among these latent trait levels because it is harder to compose such items. Therefore, when reliable measurement is not feasible with a small number of items, the selection and administration of new items may best be terminated after a fixed number of items to keep the administration efficient.

Finally, a new item is selected and administered to the respondent as long as the stopping rule has not been met. There are several methods to select a new item, but usually it is the same method used to determine the starting item. Consequently, a new item is selected based on its information value, only this time where it is highest at the provisional latent trait estimate. This iterative procedure of administering a new item and estimating the respondent's latent trait level and the measurement precision continues until the stopping rule is reached.

1.3.2 International CAT developments

CAT was first implemented in the educational field (e.g., Weiss & Kingsbury, 1984). In (mental) health care, the first major CAT implementation developments started just after the turn of the century with the founding of PROMIS (Cella et al., 2007, 2010; www.healthmeasures.net/explore-measurement-systems/promis). PROMIS started as an initiative of eight major United Stated (US) research centers financed by the National Institute of Health. Their goal was to develop a new and uniform set of instruments to assess self-reported health and well-being. More specifically, they wanted to assist clinicians in assessing patients' responses to interventions, and in modifying treatment plans based on these responses. Furthermore, they wanted to stimulate and standardize clinical research dealing with health-related measurements. To accomplish this, PROMIS started in 2002 to inventory commonly assessed measurement domains for a large part of the health spectrum. Measurement domains can be a symptom area (e.g., pain or depression) or an ability or capacity (e.g., physical or cognitive functioning). In the following years, PROMIS developed, psychometrically evaluated, and normed tens of item banks. The items in the final item banks were based on an extensive selection procedure, and were primarily chosen for their discriminative ability and coverage of the latent traits. In addition, PROMIS founded an assessment center through which the item banks can be administered as CAT instruments (www.assessmentcenter.net).

For measuring aspects of mental functioning, PROMIS initially developed item banks for the symptom areas depression, anxiety, and anger to measure children and adolescents (Irwin et al., 2010) and adults (Pilkonis et al., 2011). Gradually, PROMIS has developed additional item banks to measure aspects of mental health. For children, item banks were developed to measure cognitive function, life satisfaction, meaning and purpose, positive affect, and psychological stress experiences (e.g., Bevans et al., 2018; Forrest et al., 2018). For adults, item banks were developed to measure alcohol use, cognitive function, life satisfaction, meaning and purpose, positive affect, psychosocial illness impact, self-efficacy for managing chronic conditions, smoking, and substance use (e.g., Pilkonis et al., 2016; Valentine, Weiss, Jones, & Andersen, 2019). More information on these item banks can be found in the official PROMIS manuals (www.healthmeasures.net/promis-scoring-manuals).

As expected, it was demonstrated that PROMIS CAT instruments (further referred to as PROMIS CATs) have several advantages over other instruments used in practice (also known as legacy instruments). This seems especially the case for the administration efficiency and the reliability of measurement. For example, US PROMIS CATs for Depression and Anxiety were shown to be more efficient and reliable than several legacy instruments (e.g., Pilkonis et al., 2014; Schalet et al., 2016). For validity and responsiveness, the evaluated instruments were found to be somewhat similar (e.g., Kroenke, Baye, & Lourens, 2019; Pilkonis et al., 2014). It should be noted, however, that these conclusions were based on those psychometric properties that were actually studied. For other psychometric properties such as LMI, it remains unclear whether PROMIS CATs have additional advantages over legacy instruments.

1.3.3 National CAT developments

The international response to the PROMIS initiative has been promising indeed: many of the item banks have been translated in numerous languages with more translations being produced and evaluated every year (www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis/available-translations). In the Netherlands, the Dutch-Flemish (DF) PROMIS group was established in 2009 to investigate whether the PROMIS methodology could also be successfully implemented for the DF population (Flemish is a variant of the Dutch language spoken in Belgium; www.dutchflemishpromis.nl). As starting point, they translated 17 of the adult PROMIS item banks (Terwee et al., 2014) and 9 of the pediatric PROMIS item banks (Haverman et al., 2016) into DF. For measuring aspects of mental health, item banks have been translated for the symptom areas depression, anxiety, and anger. These item banks are now ready to be psychometrically evaluated for CAT implementation.

A strict condition for PROMIS CATs to be successful in Dutch mental health care is a generic solution to administer these dynamic instruments to patients. For a long time, CAT software was only available in small research settings, for example at universities. In 2016, however, the DF PROMIS group collaborated with IRT-experts of the University of Twente and the University of Groningen and founded an assessment center. The DF Assessment Center can provide CAT instruments and fixed-item instruments through a software link to a PROM administration platform or a ROM-questionnaires provider. In 2017, Vital Health was the first ROM-questionnaires provider that successfully connected to the DF Assessment Center (www.philips.nl/healthcare/sites/vitalhealth/products/questmanager-vragenlijstenbeheer-proms-rom). By 2021, many other ROM-questionnaire providers have also made this connection successfully. These include EasyROM (www.kgvp.org/nl), OnlinePROMS (www.onlineproms.nl), Datec (www.datec.nl), Qualizorg (www.qualizorg.nl), BrightFish (www.brightfish.nl), MobileCare (www.mobile-care.nl), KLIK (www.hetklikt.nu), and Fysiomanager (www.fysiomanager.nl).

1.4 The current thesis

1.4.1 Summary

In section 1.1, it was argued that the currently used instruments in Dutch mental health care have several limitations for aiding patients and clinicians in achieving effective and efficient treatment. First, it is unclear for many instruments whether their quality is adequate for treatment evaluation because relevant psychometric properties have been studied insufficiently. Second, the use of fixed item sets has made it challenging to develop instruments that are both highly efficient and highly reliable. Finally, the large variety in used instruments has made it difficult for mental health care providers to learn from each other's treatment outcomes.

To work towards a solution for these issues, section 1.2 presented an overview of the psychometric properties that are relevant for instruments used in treatment evaluation. Furthermore, section 1.3 presented information on CAT methodology along with relevant (inter)national CAT developments of the PROMIS initiative. Following this line of thought, the current thesis builds on the existing work by evaluating the psychometric properties of the DF PROMIS item banks for CAT administration in Dutch mental health care. Well studied measurement instruments may reduce the probability of a clinician making biased inferences because information is available on a wide range of relevant psychometric properties. Moreover, PROMIS CATs in particular may lead to valid and responsive measurement that will decrease the burden on patients due to a small number of administered items while still providing the clinician with a sufficient reliable score. As a result, PROMIS CATs may increase the probability of obtaining complete and high-quality information from patients to aid clinicians in achieving effective and efficient treatment. For this reason, PROMIS CATs may even have the potential to be the new standard for evaluating patients' treatment in the Netherlands. If this will be the case, clinicians from different mental health providers can compare treatment outcomes more easily without having to account for probable bias due to the use of different instruments.

1.4.2 Outline

The main goal of this thesis is to evaluate the DF PROMIS adult v1.0 item banks for Depression and Anxiety for CAT administration in the Dutch clinical (i.e., patients with common mental disorders in ambulatory mental health care) and general population. While collecting the data for this goal, it was first investigated what *the potential* of CAT is based on one of the instruments that is currently used to measure depression and anxiety: the Mood and Anxiety Symptom Questionnaire (MASQ; Watson & Clark, 1991). The MASQ was already evaluated for CAT implementation in Dutch clinical practice, but only for one of its three subscales (i.e., positive affect; Smits et al., 2012). In Chapter 2, their study is expanded by including the subscales for the other two elements of the tripartite model: negative affect and somatic anxiety. Several psychometric properties are evaluated for the three MASQ scales, including unidimensionality, LI, monotonicity, absence of DIF, and fit of the adopted IRT-model. Furthermore, post hoc CAT simulations are used to evaluate the efficiency gains of CAT under a specific measurement precision. A post hoc CAT simulation is not an actual CAT administration, but selects the item responses from a full item bank, and evaluates them as if they had been collected adaptively. It has been shown that the outcomes of post hoc CAT simulations and real CAT administrations tend to be very similar (Kocalevent et al., 2009).

Turning to the PROMIS CATs, the following chapters describe the evaluation of a wide collection of psychometric properties of the DF PROMIS adult v1.0 item banks for Depression and Anxiety. As starting point for Chapter 3 and 4, data of a clinical and general population sample are used because the aim of this thesis is to develop instruments that measure depression and anxiety to their full extent. Furthermore, similar to the US PROMIS initiative, the DF PROMIS CAT studies start with an *extended* version of the original PROMIS item banks. The reason for this is that PROMIS' selection of items in the US may be strongly influenced by the English language and American culture. It should therefore first be assessed whether completion by Dutch persons results in a similar selection of items.

In Chapter 3, the extended version of the DF PROMIS adult v1.0 item bank for Depression is evaluated on several sources of evidence for validity to establish a valid item set as input for CAT. Similar to Chapter 2, these sources of evidence include unidimensionality, LI, monotonicity, absence of DIF, and fit of the adopted IRT-model. In addition, it is assessed whether the extra items in the extended item bank have added value. For this purpose, post hoc CAT simulations are used to compare the extended and the original PROMIS item bank on administration efficiency, measurement precision, and concurrent validity.

In Chapter 4, the DF PROMIS v1.0 adult item bank for Anxiety is evaluated on several sources of evidence for validity to establish a valid item set as input for CAT. In this study, however, the items are limited to those of the original PROMIS item bank. The reason for this is that it is shown in Chapter 3 that the extended item bank for Depression does not differ substantially from the original item bank regarding administration efficiency, measurement precision, and concurrent validity. Similarly, these results were also found for the Anxiety item bank in the pre-analysis stage of Chapter 4. Therefore, it was decided to use the original PROMIS item banks for the remainder of the thesis to enhance international comparability. This choice also makes it possible to focus on other analyses in this chapter. Consequently, post hoc CAT simulations are used to compare the clinical and general population on administration efficiency, measurement precision, and concurrent validity.

In Chapter 5, it is investigated whether the depression and anxiety constructs as measured by the DF PROMIS v1.0 adult item banks are sufficiently invariant over time in a clinical sample. As full LMI rarely holds (van de Schoot et al., 2015), the analyses are not limited to the standard methodology used to evaluate this psychometric property (e.g., Fokkema et al., 2013). Additionally, the expected invariance violations are evaluated on their significance for clinical practice. It is investigated when (i.e., which test-occasion) and where (i.e., which item and response category) the invariance violations have substantial impact on the measured constructs (Liu et al., 2017), and to what degree changes in test scores are affected (Liu & West, 2018). In this way, specific information will be available to test users that can help them to better account for potential bias in the scores.

In Chapter 6, the DF PROMIS adult v1.0 item banks for Depression and Anxiety administered as genuine CATs are compared with a popular Dutch legacy instrument: the BSI (de Beurs & Zitman, 2005). Using pretest and retest data of a clinical sample, responsiveness and several sources of evidence for validity are evaluated to convince test users that changing to PROMIS CATs results *at least* in similar (and preferably even better) assessment of patients. Furthermore, the usability of several change indicators is compared between the instruments to facilitate the use of the PROMIS CATs in clinical practice.

Finally, the thesis is concluded with a general discussion. In Chapter 7, it is first evaluated whether the main goal of the thesis can be considered achieved. This includes a discussion of the study results in light of the existing research literature, the study strengths, and the study limitations. Furthermore, several directions for follow-up research are suggested to make even better use of CAT methodology. Finally, the discussion is closed by elaborating on several points of attention for the implementation of CAT instruments in clinical practice.

Chapter 2

Simulating Computer Adaptive Testing with the Mood and Anxiety Symptom Questionnaire

Published as:

Flens, G., Smits, N., Carlier, I., van Hemert, A. M., & de Beurs, E. (2016). Simulating computer adaptive testing with the Mood and Anxiety Symptom Questionnaire. *Psychological Assessment*, *28*(8), 953.

2.1 Abstract

In a post hoc simulation study (*N* = 3,597 psychiatric outpatients), we investigated whether the efficiency of the 90-item Mood and Anxiety Symptom Questionnaire (MASQ) could be improved for assessing clinical subjects with computerized adaptive testing (CAT). A CAT simulation was performed on each of the 3 MASQ subscales (positive affect, negative affect, and somatic anxiety). With the CAT simulation's stopping rule set at a high level of measurement precision, the results showed that patients' test administration can be shortened substantially; the mean decrease in items used for the subscales ranged from 56% up to 74%. Furthermore, the predictive utility of the CAT simulations was sufficient for all MASQ scales. The findings reveal that developing a MASQ CAT for clinical subjects is useful as it leads to more efficient measurement without compromising the reliability of the test outcomes.

Keywords: computer adaptive test, clinical assessment, Mood and Anxiety Symptom Questionnaire, item response theory

2.2 Background

In the Netherlands, routine outcome monitoring (ROM) has been implemented for mental health care patients nationwide (Carlier et al., 2012a; de Beurs et al., 2011). ROM is the repeated administration of questionnaires to monitor patients' progress over time and use the information to adjust treatment, if indicated. In the clinical setting, care providers and patients have limited time and to keep costs at a minimum, assessments should preferably be short and test outcomes reliable for all patients. A successful methodology that addresses these needs is computerized adaptive testing (CAT). CAT uses information from questions that have been answered so far by an individual in order to select the most appropriate next question. By administering questions tailored to each patient, CAT can reduce respondent burden while maintaining or even improving the reliability of the test outcomes for all patients (Fliege et al., 2005). Ideally, these CAT benefits would decrease respondent burden, increase response rates and reduce possible bias due to selective loss of respondents (Dillman, Sinclair, & Clark, 1993).

Building a full functioning CAT takes a considerable effort (Cook, O'Malley, & Roddey, 2005). One of the reasons is that in most countries, large item banks are generally unavailable for mental health constructs and have to be developed (Gibbons et al., 2014). A solution to this problem could be the use of existing mental health questionnaires as item banks. Although CAT versions of existing clinical scales have already shown to be useful in undergraduate students (Forbey & Ben-Porath, 2007; Gardner et al., 2004; Smits, Cuijpers, & van Straten, 2011), Smits and colleagues specifically assessed in a post hoc simulation study whether a CAT would be useful for measuring clinical subjects (Smits, Zitman, Cuijpers, den Hollander-Gijsman, & Carlier, 2012). As a first proof of principle for using an existing questionnaire to develop a CAT for clinical subjects, they simulated a CAT on one of the Mood and Anxiety Symptom Questionnaire (MASQ; Watson & Clark, 1991) subscales (i.e., the 22-item Anhedonic Depression subscale) by treating patients' responses as if they had been collected adaptively. With the outcomes of the CAT simulation set to a high level of measurement precision, their analysis showed that patients' burden was reduced substantially; the administration of the MASQ Anhedonic Depression scale was shortened for most of the patients with a mean decline of 59% (from 22 to 9 items). Moreover, the outcomes of the CAT remained diagnostically accurate.

The full 90-item MASQ is an extensive questionnaire which has a unique way of assessing symptoms of the two most prevalent psychiatric syndromes, depression and anxiety disorders (according to the tripartite model), and takes into account the high comorbidity between both syndromes and high level of symptom overlap (Watson & Clark, 1991). It is used as research- and clinical assessment instrument, and has been validated in multiple countries, for multiple age groups, and for multiple disorders (e.g., de Beurs, den Hollander-Gijsman, Helmich, & Zitman, 2007; Deng, Jiang, & Li, 2012; Lee, Kim, & Cho, 2015). Ideally, for efficient measurement of clinical subjects, all subscales of the MASQ are transformed into a CAT. Previous studies have generally confirmed three subscales of the 90-item MASQ: a positive affect scale (PA), a negative affect scale (NA), and a somatic anxiety (SA) scale (Bedford, 1997; Clark & Watson, 1991; de Beurs et al., 2007, Keogh & Reidy, 2000; Watson et al., 1995). Other studies that developed shorter versions of the MASQ also applied this three factor structure in their item design (Osman et al., 2011; Wardenaar et al., 2010). In these studies, the number of items for each MASQ scale was fixed, but by doing so, the measurement precision for test outcomes could vary among respondents with different trait levels. By contrast, CAT is more dynamic: it fixes the test outcomes' measurement precision for all trait levels and allows for the number of administered items to vary among respondents (Embretson & Reise, 2000). In other words, CAT is essentially more efficient than fixed questionnaires because CAT administers only the most informative items to each individual respondent.

In this paper, we assessed in a post hoc CAT simulation study whether the administration of three MASQ subscales could be made more efficient for measuring patients receiving mental health care. We present a comprehensive account of the psychometric evaluation of the MASQ scales, which is a prerequisite for applying CAT. As point of departure for the CAT simulations, we have used data from a large Dutch clinical sample (Smits, Zitman, Cuijpers, den Hollander-Gijsman, & Carlier, 2012) applying a three-factor structure to the MASQ from clinically-based MASQ subscales (de Beurs et al., 2007). We assessed to what extent the administration of each MASQ scale can be shortened for clinical subjects and whether the CAT estimates are diagnostically accurate compared with the full-scale estimates.

2.3 Method

2.3.1 Participants

The sample for this study consisted of 3,597 patients (63% female) from three Dutch outpatient Mental Healthcare Centres of the Regional Mental Health Care Provider Rivierduinen. The mean age of the patients was 38.8 years for the entire sample (*SD* = 13.2), 38.2 years for females (*SD* = 13.3), and 39.9 years for males (*SD* = 13.1). Patients were referred to Rivierduinen by their general practitioner for treatment of mood, anxiety and/or somatoform disorders. The patient's diagnosis was assessed with the Dutch translation of the Mini International Neuropsychiatric Interview (MINI-plus; Sheehan et al., 1998) administered by a psychiatric nurse who was extensively trained. The MINI-plus is a standardized interview for clinical diagnosis of mental disorders following the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV*; American Psychiatric Association, 1994). According to the MINI-plus, the sample for this study was classified as follows: 23% of the patients had a singular mood disorder, 20% had a singular anxiety disorder, 8% had a singular somatoform disorder, and 23% did not meet the criteria of these disorders. Furthermore, 18% of the patients had a comorbid mood and anxiety disorder, 4% had a comorbid mood and somatoform disorder, 3% had a comorbid anxiety and somatoform disorder, and 2% suffered from all three disorders.

Rivierduinen collaborated with the Department of Psychiatry of the Leiden University Medical Centre (LUMC) in developing ROM (de Beurs et al., 2011). At intake, patients were informed that ROM is a part of the general policy of Rivierduinen and LUMC, designed to monitor treatment outcome, that their data could be used for research purposes in anonymous form, and that their personal outcome data would be made available only to their therapist. If patients did not consent with the procedure, their data were removed from the database.

Anonymity of the patients and proper handling of the data were assured by a comprehensive policy protocol (Psychiatric Academic Registration Leiden). This policy protocol was made available for patients upon request. The procedure was approved by The Medical Ethical Committee of the LUMC (for more details, see de Beurs et al., 2011).

2.3.2 The MASQ

The MASQ is a 90-item self-report questionnaire that contains feelings, sensations, problems and experiences that people can have associated with mood and anxiety disorders (Watson & Clark, 1991). The full 90-item MASQ was designed to measure symptoms of mood and anxiety disorders according to the tripartite model (Clark & Watson, 1991). The tripartite model aims to account for the high concordance among symptom measures for affective disorders, by assigning symptoms to one of three groups: a group unique to mood disorders (anhedonia or lack of positive affect [PA]), a group unique to anxiety disorders (somatic anxiety [SA]), and a group common to both mood and anxiety disorders (negative affect [NA]). Of the 90 MASQ items, 27 are stated positively (e.g., Item 1 *Felt cheerful*) and 63 are stated negatively (e.g., Item 2 *Felt afraid*). For this study, the Dutch adaptation of the MASQ was used (de Beurs et al., 2007). Patients were asked by computer to indicate on a Likert scale (1 = *not at all*, 2 = *a bit*, 3 = *moderately*, 4 = *much*, and 5 = *very much*) how frequently they experienced the stated feelings, sensations, problems and experiences in the past 7 days, including today. For scoring, the positively stated items were reversed (1 = 5, 2 = 4, 3 = 3, 4 = 2, 5 = 1). Thus, all MASQ scale scores had the same meaning: the higher the score, the more severe the mood or anxiety problems.

As input for the CAT simulations, multiple MASQ factor solutions were available (e.g., Bedford, 1997; Clark & Watson, 1991; Keogh & Reidy, 2000; Watson et al., 1995). In this study, the MASQ items from the Dutch factor solution were used (de Beurs et al., 2007). Firstly, because this factor solution was based on a large Dutch clinical sample. Secondly, because the Dutch subscales showed satisfactory psychometric properties and results that were similar to factor solutions from United States and British datasets (Keogh & Reidy, 2000). The Dutch factor solution grouped 22 of the 90 MASQ items in the lack of PA, 20 items in the NA, and 18 items in the SA. Table 2.1 displays the items from the three Dutch MASQ subscales.

Table 2.1 Items from the three Dutch MASQ scales (PA, NA and SA).

Scale	Item
PA	1, 11, 14, 18, 23, 27, 30, 35, 36, 38, 40, 41, 43, 46, 49, 54, 58, 62, 68, 72, 78, and 86
NA	4, 6, 8, 13, 16, 17, 20, 22, 24, 26, 28, 29, 42, 47, 53, 64, 74, 77, 84, and 89
SA	9, 25, 45, 48, 52, 55, 57, 61, 63, 65, 67, 69, 73, 75, 79, 81, 87, and 88

2.3.3 Psychometric evaluation of the MASQ scales

We undertook a psychometric evaluation of the three MASQ scales (Reeve et al., 2007), which is a prerequisite for applying CAT. It was evaluated whether each of the scales met the three main item response theory (IRT) assumptions of unidimensionality, local independence (LI) and monotonicity. Violation of these assumptions may cause bias in the scaling of persons and items on a common latent trait, which could result in over- or underestimated trait scores. In addition, we evaluated differential item functioning (DIF; Embretson & Reise, 2000) among key demographic groups. Items containing DIF cause bias in latent trait scores because persons from different groups with the same latent trait score have different probabilities of selecting item response categories.

The IRT assumption of *unidimensionality* states that a person's item response results from the person's trait level that the item measures and not from other factors. Because mental health constructs are generally complex, item response results are rarely strictly unidimensional (Reise, Morizot, & Hays, 2007). For IRT applications, it is therefore assessed whether the degree of unidimensionality in item response assessments is sufficient. The degree of unidimensionality in each MASQ scale was explored with both confirmatory factor analyses (CFA) and exploratory factor analyses (EFA) conducted on the polychoric correlation matrix of the items (Bollen, 1989). CFA was evaluated by the fit indices comparative fit index (CFI; > 0.95 for good fit), Tucker-Lewis index (TLI; > 0.95 for good fit), root-mean-square error of approximation (RMSEA; < 0.06 for good fit), and the average absolute residual correlations (< 0.10 for good fit; Reeve et al., 2007), using the R package lavaan (Version 0.5-17; Rosseel, 2012). EFA (varimax rotated) was evaluated with the proportion of variance explained by the resulting factors using the R package psych (Version 1.3.2; Revelle, 2013). Proportion of variance explained in the first factor should be above the Reckase criterium of 20% (Reckase, 1979, cited in Hambleton, 1988), and the ratio of variance explained in the first and second factor should be higher than the minimal requirement of 4 (Reeve et al., 2007).

The assumption of *LI* states that no association should exist among item responses when controlling for the trait level. LI was evaluated among the polytomous response items by inspecting the residual correlation matrix resulting from CFA using the R package lavaan (Version 0.5-17; Rosseel, 2012). Items with residual correlations above 0.20 are considered to be possibly locally dependent (Reeve et al., 2007). Further investigation of LI was done with Yen's Q3 statistic (Yen, 1993). This statistic calculates the residual item scores under the graded response model (GRM; Samejima, 1969) and correlates these among items. For this purpose, we fitted the GRM to each of the MASQ scales using the R package ltm (Version 1.0; Rizopoulos, 2006). As suggested by Smits et al. (2012), the lack of model fit was assessed by Cohen's rules of thumb to interpret effect size; Q3 values between 0.24 and 0.36 imply a moderate deviation, Q3 values above 0.37 imply a large deviation (Cohen, 1988). Item pairs with large deviations were evaluated according to their effect on the item parameter estimates (Reeve et al., 2007). First, we estimated the item parameters of the corresponding MASQ scale. Second, we removed one of the items with a large deviation from the scale and estimated the item parameters for the remaining items. Last, we compared the item parameters from the full scale with the restricted scale (minus one item) to assess whether substantial differences occurred between the remaining parameters. This process was repeated for each item with a large deviation.

The IRT assumption of *monotonicity* states that the probability of selecting an item response that suggests a better health status on a scale should increase as the underlying level of health status on that scale is higher. We evaluated monotonicity by examining graphs of item mean scores conditional on rest scores (total raw score minus the item score). Furthermore, we performed the nonparametric IRT approach Mokken (1971) scale analysis using Mokken scaling with the R package mokken (van der Ark, 2007). In this analysis, persons are ranked on a unidimensional scale according to their trait level and items with regard to their location. According to the rule of thumb of Mokken (1971), a scale has low quality when the scalability coefficient is between 0.3 and 0.4, moderate quality when the scalability coefficient is between 0.4 and 0.5, and high quality when the scalability coefficient is above 0.5.

Finally, *DIF* (Embretson & Reise, 2000) was evaluated for the demographic variables age and gender, using the R package lordif (Version 0.2-2; Choi, Gibbons, & Crane, 2011). An item contains DIF if the probability of responding in different response categories differs across groups, while the trait level influencing a person's response to an item is controlled for. As a consequence, each group should have their own item parameter estimations for items containing DIF. For example, when men with a high level of PA have a higher probability of being more cheerful than women with an identical level of PA, then the MASQ Item 1 *Felt cheerful* contains probably DIF and should have separate item parameter estimations for men and women. DIF comes in two kinds: uniform and nonuniform (Embretson & Reise, 2000; Reeve et al., 2007). Uniform DIF has the same magnitude of DIF across the entire range of the trait. Nonuniform DIF has a different magnitude or direction of DIF across the trait. We explored both kinds of DIF using ordinal logistic regression (OLR; Crane, Gibbons, Jolley, & van Belle, 2006). OLR has the advantage of being a flexible and robust framework for DIF detection, especially with trait level scores from IRT. Effect size was evaluated by means of change in McFadden's R^2 between groups, following the suggestion of a critical value of 0.02 (Choi et al., 2011) for rejecting the hypothesis of no (uniform or nonuniform) DIF. For each scale, differences were evaluated for gender (men and women) and age (divided by means of the median).

2.3.4 CAT Simulation

We simulated a separate CAT on each of the three MASQ scales (PA, NA and SA) from the item responses that were obtained from the patients. The item responses were selected for each patient from all the item responses in the corresponding scale and were evaluated as if they were collected adaptively. Basically, the CAT simulation started with the same item for every individual and then estimated the latent scale score and measurement precision using both item response and item properties. From here, either a new item was selected according to the item properties and the estimated latent trait level, or the simulation stopped when the prespecified value of measurement precision was obtained. The selection of new items, and the estimation of latent trait score and measurement precision using all collected item scores so far, continued until this prespecified measurement precision was reached, or when all items were used; items were used only once. To apply this procedure, we made several decisions regarding (a) the IRT model that estimates the item parameters, (b) the methods for selecting new items and (c) estimating patients' latent scale scores (θ), and (d) the starting level and (e) stopping rule for the CAT. A program (Smits et al., 2011; Smits et al., 2012) was written in the statistical environment R (R Core Team, 2014) to implement these decisions into three separate CAT simulations. Below, we will present the details concerning the decisions rules.

First, as an appropriate IRT model for estimating item parameters, we used Samejima's (1969) GRM for polytomous items. The GRM is often the preferred IRT model, because it is easier to illustrate to test users than other models, and the item parameters are easy to interpret with regard to responder behavior (Ostini, Finkelman, & Nering, 2015; Smits et al., 2011). These advantages are especially desirable when CAT is implemented on a large scale, as is mostly the case in clinical measures, because clinicians should generally understand how CAT works. The GRM model uses two types of parameters. The discrimination parameter *a* specifies to what extent persons with similar scores on the latent trait can be differentiated by the item. Furthermore, the GRM uses the location parameters *b* (the number of location parameters for an item is equal to the number of response categories minus one) which specifies the θ location on which a patient is expected to choose from a lower to a higher item response. We fitted the GRM to the data separately for each scale using the R package ltm (Version 1.0; Rizopoulos, 2006). The GRM was evaluated for each scale by examining model fit and evaluating item properties. Model fit was evaluated by correlating the estimated latent trait scores under the GRM with the traditional MASQ scale scores. Item properties were evaluated by examining the *a* and *b* parameters estimated from the GRM models.

Next, we chose a method for selecting new items and estimating patients' latent scale scores (θ). New items were selected using item information, which is the most used method in other CATs (Embretson & Reise, 2000; Wainer, Dorans, Flaugher, Green, & Mislevy, 2000). Item information specifies how precisely an item can measure the latent trait given the location of the person's estimate. The CAT selected each time a new item which had the highest information at the provisional estimate of θ. In addition, θ was estimated with the maximum a posteriori method (MAP; Embretson & Reise, 2000). MAP is a Bayesian method, which estimates θ as the value with the highest likelihood of bringing forth the observed item responses using a prior standard normal distribution of θ. This Bayesian method was chosen over the maximum likelihood method (ML; Thissen, 1991) for being able to provide a θ estimate for item response patterns consisting exclusively of either extreme low or extreme high response categories.

Finally, we chose a starting level and stopping rule for the CAT. The starting level was set to the average value of the latent trait ($\theta = 0$). As a first item for all respondents, we therefore chose the MASQ item which had the highest information at this starting level: Item 86 for the PA scale (*Felt really good about myself*), Item 22 for the NA scale (*Felt hopeless*), and Item 79 for the SA scale (*Was trembling or shaking*). In addition, there are generally two types of stopping rules for a CAT: (a) a fixed number of administered items, or (b) a prespecified value of measurement precision (*SE*). Because this study was set out to find both reliable and shorter measures, we specified that the CAT simulation stopped applying new items when the latent trait estimate of a patient reached a $SE(\theta) < 0.3$, comparable to a marginal reliability of .90 (Green, Bock, Humphreys, Linn, & Reckase, 1984). This value of measurement precision is generally required for minimal reliability for individual assessments (Bernstein & Nunnally, 1994, p. 265). When a $SE(\theta) < 0.3$ was not obtained after administering all items, the CAT simulation stopped.

We split the data randomly into two equally sized datasets for the simulations: one for estimating the item parameters and one for simulating the CAT. After all, when one uses the same sample to estimate the item parameters and to simulate the CAT, the procedure might lead to overfitting (Hastie, Tibshirani, & Friedman, 2001), resulting in outcomes which are too optimistic. Several statistics were recorded separately for each scale: (a) the mean and standard deviation of the number of administered items, (b) the percentage of patients for whom all items had to be administered, and (c) the mean *SE* of the final θ estimate for all patients.

2.3.5 Comparing full-scale data with CAT data

A CAT may be considered efficient when it shows a substantial decline in administered items compared with the full item bank administration, and outcomes with sufficient reliability. Furthermore, the good psychometric properties of the scale have to be retained, such as sufficient criterion validity for diagnostic status of the patient. This was investigated by comparing CAT outcomes to the full-scale outcomes of the questionnaire.

We performed two analyses to assess whether the CAT scores show sufficient similarity with the full MASQ scale scores. In the first analysis, we assessed whether the CAT *outcomes* are similar to the full MASQ scales. The CAT θ estimates were compared for each MASQ scale with the full-scale θ estimates (PA, 22 item scores; NA, 20 item scores; SA, 17 item scores), using Pearson correlations and scatterplots. Furthermore, we assessed the size of difference between the outcomes expressed as Cohen's *d* (using pooled *SD*'s for the CAT and the full MASQ scale). Cohen's *d* was evaluated using the guideline proposed by Cohen (1988): 0.2 = small effect, 0.5 = medium effect, 0.8 = large effect.

In the second analysis, we assessed whether the *predictive utility* (i.e., criterion validity; McDonald, 1999) of the CATs was similar to that of the full MASQ scales. We formed three patient classifications based on the MINI-plus diagnosis (Sheehan et al., 1998): (a) a mood disorder or no disorder, (b) an anxiety disorder or no disorder, and (c) a comorbid mood and anxiety disorder or no disorder. We then assessed whether the CAT simulation scores and the full MASQ scale scores could predict the patients classifications to a similar degree using the area under the curve (AUC) of the receiver operating curve, an effect size for diagnostic accuracy (Rice & Harris, 2005). In this study, AUC can be interpreted as the probability that a randomly selected person with a disorder has a higher outcome on the corresponding MASQ scale (i.e., more severe problems) than a randomly selected person without an disorder (Zweig & Campbell, 1993). We evaluated the AUC values using the guideline proposed by Rice and Harris (2005): .56 = small effect, .64 = medium effect, .71 = large effect; a higher effect meaning a higher predictive utility for the scale.

2.4 Results

2.4.1 Psychometric qualities of the MASQ scales

Table 2.2 displays the CFA fit statistics for the MASQ scales. All statistics showed a good fit, with the exception of the RMSEA for the NA and SA scales, which resulted in a moderate fit (both 0.08). In addition, EFA results showed that the proportion of variance explained in the first factor of each MASQ scale were all above the Reckase criterium of 20% (PA = 60%, NA = 60%, SA = 52%; Reckase, 1979, as cited in Hambleton, 1988). Furthermore, the ratio of variance explained in the first and second factor were all higher than the minimal requirement of 4 (PA = 15, NA = 10, SA = 9; Reeve et al., 2007). According to these results, we concluded that all three scales sufficiently met the assumption of unidimensionality.

Table 2.2 Confirmatory factor analysis fit statistics for all MASQ scales (PA, NA and SA).

Statistic	PA	NA	SA
CFI	.996	.992	.982
TLI	.996	.992	.980
RMSEA	.057	.077	.082
Average absolute residual correlations	.031	.043	.051

Note. CFI = scaled comparative fit index; TLI =Tucker-Lewis index; RMSEA = scaled root-mean-square error of approximation.

One item pair (Items 9 – 63) in the SA scale was considered to be possibly locally dependent as its residual correlation was above 0.20; both items are associated with assessing the feeling “belly ache”. In addition, deviations of local independence (LI) according to Yen’s Q3 statistic were found in the NA and the SA scales: the NA scale showed moderate deviations in four item pairs (Items 16 - 47, 16 - 64, 47 - 64, and 53 - 64), the SA scale showed moderate deviations in two item pairs (Items 55 – 79 and 69 - 81) and large deviations in two item pairs (Items 9 – 63 and 57 - 79). These item pairs showed that all items in the NA scale are associated with “feeling inferior to others”, while the items in the SA scale are mostly associated with “belly or muscle aches” and “feeling shaky”. Removing Item 9 or 63 from the SA scale resulted in a negligible difference in parameter estimates (max 0.07 for *a* and 0.05 for *b*). However, removing Item 57 or 79 resulted in more substantial differences (max 0.39 for *a* and 0.15 for *b*); both items are associated with “feeling shaky”. We finally decided to remove Item 57 from the SA scale for discriminating between persons in the least degree (i.e., it had the lowest *a* parameter). After removing Item 57, Yen’s Q3 statistic still marked item pair 9 - 63 from the SA scale with a high deviation. However, the difference in *a* and *b* parameters remained negligible when removing the items from the GRM; both items were preserved in the scale. According to these results, all scales (SA without Item 57) sufficiently met the LI assumption.

The graphs of item mean scores conditional on rest scores showed monotonicity for all items as the underlying level of the scale was higher. This result was confirmed by the Mokken scale analysis (van der Ark, 2007). The scalability coefficient for the PA and NA scales was high (0.53 and 0.55), and for the SA scale it was moderate (0.42). Furthermore, the scalability coefficients for all items were above the lower bound of 0.30. According to these results, we concluded that all three scales sufficiently met the monotonicity assumption.

For each MASQ item, change in McFadden's R^2 between men and women, and between age groups divided by means of the median was below 0.02 (Choi et al., 2011). According to these results, we concluded for each scale that no items contained uniform or nonuniform DIF for the variables gender and age.

In sum, the psychometric evaluation of the MASQ scales (PA, NA and SA) showed favorable results. All scales suggested sufficient unidimensionality, complied with the monotonicity assumption, and the items contained no DIF according to gender and age. However, based on the analyses evaluating LI, Item 57 of the SA scale was removed from the scale. After removing this item, the results of all analyses showed slightly improved psychometric characteristics. We concluded that all three MASQ scales could be used as inputs for a CAT simulation.

Table 2.3 Location and discrimination parameters values for the items of the MASQ Scales (PA, NA and SA).

PA						NA						SA					
Item	Item parameter estimates					Item	Item parameter estimates					Item	Item parameter estimates				
	a	b_1	b_2	b_3	b_4		a	b_1	b_2	b_3	b_4		a	b_1	b_2	b_3	b_4
1	2.21	-2.67	-1.19	-0.15	0.77	4	1.62	-0.46	0.57	1.26	2.47	9	1.35	0.45	1.45	2.12	3.44
11	2.01	-3.57	-2.24	-1.18	-0.45	6	2.12	-1.19	-0.08	0.53	1.62	25	1.87	0.48	1.20	1.76	2.72
14	2.34	-2.72	-1.57	-0.66	0.12	8	2.25	-0.80	0.18	0.82	1.90	45	1.75	0.75	1.53	2.11	3.24
18	2.36	-2.85	-1.56	-0.67	0.35	13	2.49	-0.46	0.30	0.88	1.75	48	1.86	0.16	0.85	1.40	2.36
23	2.62	-2.92	-1.62	-0.77	0.03	16	2.61	-0.98	-0.11	0.51	1.44	52	1.76	-0.18	0.71	1.27	2.24
27	1.46	-3.07	-1.38	-0.31	0.62	17	1.43	-1.23	-0.01	0.81	2.20	55	1.70	0.34	1.19	1.82	2.82
30	2.34	-2.25	-1.08	-0.33	0.59	20	1.91	-1.01	0.08	0.81	2.07	61	1.78	1.11	1.70	2.30	3.13
35	1.92	-3.43	-2.04	-1.10	-0.26	22	3.14	-0.43	0.26	0.77	1.60	63	1.39	0.17	1.21	1.77	2.98
36	2.04	-3.35	-1.91	-1.04	-0.23	24	1.70	-0.46	0.48	1.07	2.13	65	1.59	0.22	1.03	1.65	2.71
38	1.12	-3.16	-1.47	-0.36	0.87	26	1.74	-0.40	0.59	1.16	2.25	67	1.32	0.02	0.83	1.46	2.58
40	2.16	-2.76	-1.52	-0.69	0.22	28	1.44	-0.25	0.73	1.36	2.44	69	2.05	0.13	0.82	1.32	2.15
41	0.98	-3.58	-1.55	-0.23	1.09	29	2.42	-0.62	0.27	0.86	1.81	73	1.07	1.29	2.12	2.82	3.64
43	1.85	-3.20	-1.82	-0.88	-0.01	42	1.62	-1.01	-0.02	0.66	1.83	75	1.94	0.18	0.90	1.49	2.39
46	1.46	-3.38	-1.82	-0.63	0.37	47	2.62	-0.21	0.56	1.03	1.85	79	2.32	0.07	0.88	1.32	2.09
49	2.14	-2.89	-1.74	-0.83	0.24	53	1.32	-0.57	0.43	1.16	2.04	81	1.52	-0.37	0.43	0.96	1.93
54	1.67	-3.22	-1.70	-0.61	0.19	64	1.65	-0.50	0.46	1.06	2.19	87	1.67	1.14	1.95	2.60	3.48
58	2.93	-2.82	-1.73	-0.88	-0.27	74	2.18	-0.73	0.23	0.77	1.74	88	1.58	0.01	0.73	1.28	2.33
62	1.94	-2.50	-1.11	-0.16	0.86	77	1.73	-1.82	-0.52	0.17	1.41						
68	2.15	-3.05	-1.62	-0.65	0.25	84	1.81	-1.87	-0.58	0.01	1.18						
72	2.08	-3.05	-1.85	-0.90	-0.20	89	1.46	0.74	1.51	1.99	2.93						
78	1.91	-2.72	-1.50	-0.60	0.40												
86	2.72	-2.95	-1.64	-0.75	0.16												

Note. all PA items are positively stated items. These items were score-reversed.

2.4.2 Calibration of the MASQ scales

The correlations between the GRM's estimated theta's and the traditional MASQ scale scores were high for all scales (PA: r = .98, NA: r = .98, SA: r = .96), indicating the GRM as a good model to represent the MASQ scale scores. In Table 2.3, the estimated parameter values of the GRM model are displayed. The *a* parameters showed a considerable variation and similar patterns for all scales, ranging from *a* = 0.98 (Item 41 *Thoughts and ideas came to me very easily*) to *a* = 2.93 (Item 58 *Felt really 'up' or lively*) for the PA scale, from *a* = 1.32 (Item 53 *Felt unattractive*) to *a* = 3.14 (Item 22 *Felt hopeless*) for the NA scale, and from *a* = 1.07 (Item 73 *Was afraid I was going to die*) to *a* = 2.32 (Item 79 *Was trembling or shaking*) for the SA scale. The *b* parameters showed considerable variation in location for all scales, ranging from *b* = -3.57 (Item 11 *Felt successful*) to *b* = 1.09 (Item 41 *Thoughts and ideas came to me very easily*) for the PA scale, from *b* = -1.87 (Item 84 *Worried a lot about things*) to *b* = 2.93 (Item 89 *Thought about death or suicide*) for the NA scale, and from *b* = -0.37 (Item 81 *Muscles were tense or sore*) to *b* = 3.64 (Item 73 *Was afraid I was going to die*) for the SA scale. On the basis of these results, we concluded that the GRM model fitted the data sufficiently, and decided not to remove any further items from the item banks.

2.4.3 Characteristics of the CATs

Table 2.4 displays the CAT simulation statistics for the three MASQ scales (PA, NA and SA) under stopping rule $SE(\theta) < 0.3$: mean number of administered items (*SD*), the percentage of respondents who completed all items, and the mean $SE(\theta)$. Under this stopping rule, the mean number of administered items declines substantially for all scales (Table 2.4, column 3; PA = 56%, NA = 64%, SA = 74%). Furthermore, the standard deviation of the number of administered items is relatively high for all scales, indicating individual differences among patients (Table 2.4, column 4). This was illustrated by the fact that for some patients all items in the scales needed to be administrated (Table 2.4, column 5). For these small groups of patients, the CAT simulations showed a $SE(\theta)$ above the stopping rule's limiting value of 0.3, $0.3 < SE(\theta) < 0.6$, which caused only a slightly higher mean $SE(\theta)$ for the PA scale, $SE(\theta) = 0.31$ (Table 2.4, column 6). This result was due to having a large number of patients that had to complete all items compared with the NA and SA scales.

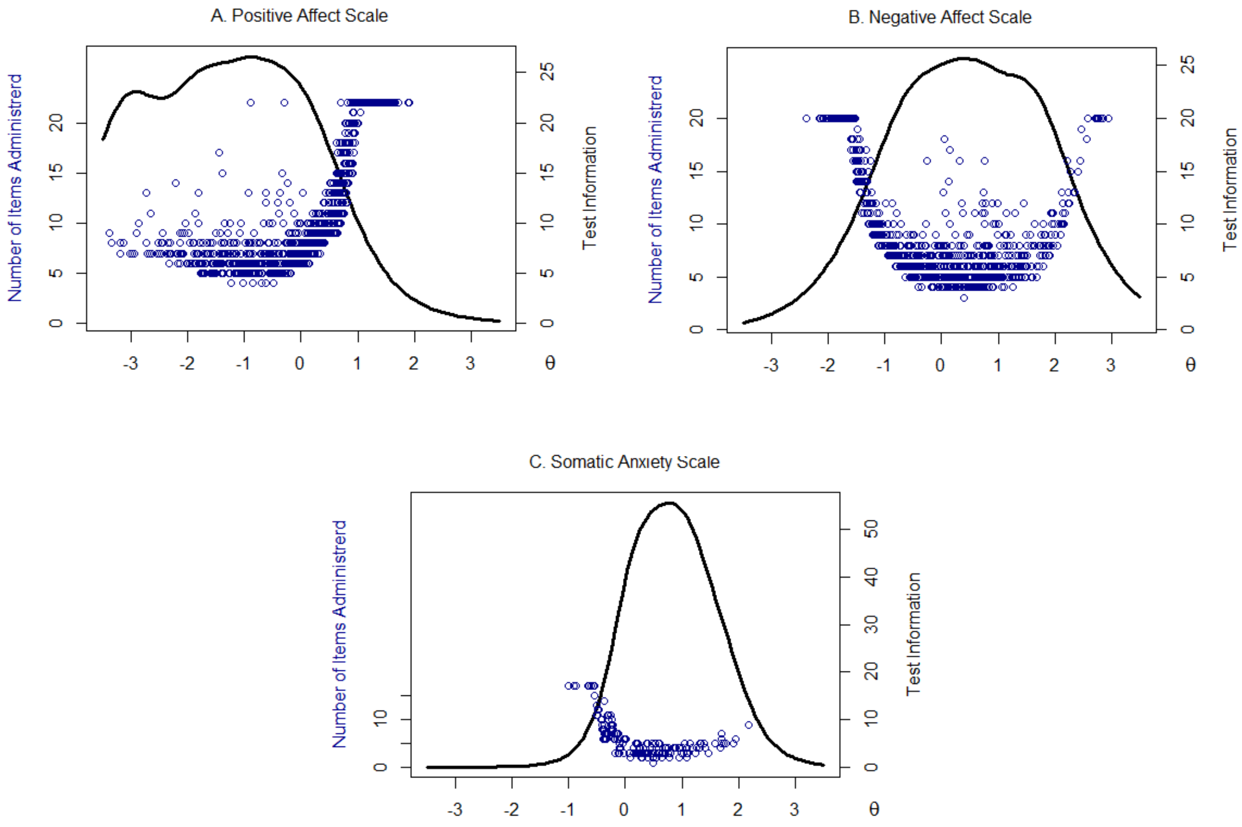
Table 2.4 CAT simulation statistics for the three MASQ scales (PA, NA and SA) under stopping rule $SE(\theta) < 0.3$.

	Number of items				
Scale	Total	*M*	*SD*	% All	*M SE*(θ)
PA	22	9.73	6.05	15	.31
NA	20	7.18	4.49	7	.29
SA	17	4.42	4.35	8	.29

Figure 2.1 shows the number of administered items for each scale as a function of the final θ estimate under stopping rule $SE(\theta) < 0.3$, a higher θ meaning more severe problems. Furthermore, Figure 2.1 shows for each scale the test information function, which specifies how precisely a test can measure the latent trait given the location of the person's estimate. Test information is calculated as the sum of all item information at any relevant θ level. For all

scales, the data confirm our finding of individual differences among patients. For example, we found that patients who completed the PA scale CAT (Figure 2.1A) with the minimum of 4 items (7%) have θ values near the middle of the scale (-1.97 < θ < 0.20), while patients who completed the PA scale CAT with the maximum of 22 items (15%) have θ values at the right end of the scale (0.70 < θ < 1.90; i.e., patients with a severe lack of PA). The NA scale (Figure 2.1B) and the SA scale (Figure 2.1C) also show that patients who completed the CAT with the maximum number of items in the scale mostly have θ estimates in the extremes (i.e., patients with a mild/severe NA, or a mild SA). This is a result of the relatively low test information in these θ estimate regions. In contrast, the SA scale has relatively high test information over almost the entire range. As a result, the mean number of administered items declined most in this scale (74%). Moreover, for 10% of the patients a single item was sufficient to complete the CAT simulation with a *SE*(θ) < 0.3 (i.e., Item 79 *Was trembling or shaking*).

Figure 2.1 Number of administered items shown as a function of the final θ estimate under stopping rule *SE*(θ) < 0.3 for the three MASQ scales (PA, NA and SA).



2.4.4 Validity of the CATs

Table 2.5 displays Pearson's correlations and sizes of differences (Cohen's *d*) for each subscale (PA, NA, SA) between the CAT θ estimates and the full-scale θ estimates. The correlations were high for all scales (PA: *r* = .98, NA: *r* = .98, SA: *r* = .89), indicating a high similarity between the CAT and the full-scale θ estimates. Next, we investigated the scatterplots for each scale and did not identify notable outliers. Finally, the Cohen's *d* values were small for all scales, indicating no structural differences between the CAT θ estimates and the full-scale θ estimates.

Table 2.5 Pearson's correlations and sizes of differences (Cohen's *d*) for each MASQ scale (PA, NA and SA) between the CAT θ estimates and full-scale θ estimates.

	Full scale		CAT			
Scale	*M*	*SD*	*M*	*SD*	*r*	*d*
PA	.05	1.01	.04	.98	.98	-.01
NA	.16	.98	.15	.95	.98	-.01
SA	.18	.52	.18	.49	.89	.00

Table 2.6 displays the AUC values of all MASQ scales (with 95% confidence intervals) for the mood disorder classification, the anxiety disorder classification, and the comorbid mood and anxiety disorder classification. The AUC values were medium to high when no stopping rule was applied and either remained equal or diminished only somewhat under the stopping rule *SE*(θ) < 0.3. These results indicate a similar predictive utility for the CAT administrations and the full MASQ scales administrations.

Table 2.6 AUC statistics for all MASQ scales (PA, NA and SA) under several stopping rules, and 95% confidence intervals.

Scale	Stopping rule	Any mood disorder	Any anxiety disorder	Any mood and anxiety disorder
PA	None: Sum score	.81 (.79, .84)	.69 (.66, .72)	.83 (.80, .86)
	None: θ	.82 (.79, .84)	.69 (.66, .72)	.83 (.80, .86)
	SE(θ) < .3	.81 (.79, .84)	.69 (.66, .72)	.83 (.80, .86)
NA	None: Sum score	.80 (.78, .83)	.71 (.68, .74)	.82 (.79, .85)
	None: θ	.80 (.77, .83)	.71 (.68, .74)	.82 (.79, .85)
	SE(θ) < .3	.80 (.77, .83)	.70 (.66, .73)	.81 (.78, .84)
SA	None: Sum score	.73 (.70, .76)	.71 (.68, .74)	.78 (.75, .81)
	None: θ	.73 (.70, .76)	.71 (.68, .74)	.78 (.75, .81)
	SE(θ) < .3	.71 (.68, .74)	.68 (.65, .71)	.76 (.72, .79)

2.5 Discussion

Until recently, most of the studies that build a CAT version for an existing clinical scale were executed with undergraduate students (Forbey & Ben-Porath, 2007; Gardner et al., 2004; Smits et al., 2011). In this study, we used data from clinical subjects to assess whether the efficiency of the MASQ could be improved with a CAT version. For this purpose, we performed a psychometric evaluation and a CAT simulation on each of the three MASQ scales. Performing a simulation enabled us to compare the full-scale assessments and the CAT simulations within the same patient group. Thus, we could directly assess to what extent the CAT simulations reduced the number of administered items and whether the θ estimates of the CAT simulations had similar outcomes and diagnostic accuracy compared with the full-scale θ estimates.

The present findings suggest that all MASQ scales are good candidates for developing an actual CAT for clinical subjects. First of all, all MASQ scales (with Item 57 removed from the SA scale) showed sufficient psychometric quality to develop a CAT. Second, the administration of all MASQ scales was shortened substantially by the CAT simulations. Third, the θ estimates of the CAT simulations were highly similar to the full-scale θ estimates and also showed highly similar predictive utility. These results are strengthened by the fact that we used data from a large sample of real-life patients in clinical care. Furthermore, the findings are in line with other studies, showing that CAT is a useful method to increase the efficiency of a questionnaire (Fliege et al., 2005; Forkmann et al., 2009; Gardner et al., 2004; Gibbons et al., 2012, 2014; Walter et al., 2007). Previously, Smits et al. (2012) demonstrated that the PA scale could be shortened by a CAT while maintaining reliable outcomes for clinical subjects. This finding can now be extended to all MASQ scales, and patients' administration burden can decrease substantially with a CAT version of the MASQ.

Although another study has shown that the outcomes of CAT simulations and real CAT administrations can be very similar (Kocalevent et al., 2009), actual CATs of the MASQ still have to be built and validated with new patient data to replicate the present results. With such a replication study, it could be investigated whether the correlations between the CAT administration and the full assessment will remain high using a separate CAT measure and full MASQ measure within the same patients. In the present study, the correlations might be inflated because the same data was used to assess a CAT outcome and a full-scale outcome. Moreover, if our results are replicated with an actual CAT, then using CAT simulations on clinical data from existing mental health questionnaires administered by computer could be considered a useful method for selecting candidate questionnaires for CAT transformation. The CAT simulation provides information about the potential efficiency increase and comparability of CAT- with the full-scale scores, which could be used to decide whether a CAT transformation is worth the investment. Compared with the development of a new item bank, this approach would save a lot of time, money, and effort. Be aware that this assumption would hold up for computer administered tests, which was the manner of administration in the present study. If the questionnaires are administered by paper and pencil, the results might be different compared with a computer-based administration due to format influences (Booth-Kewley, Larson, & Miyoshi, 2007; Hayslett & Wildemuth, 2004; Kays, Gathercoal, & Buhrow, 2012).

After replication of the present study's results with an actual CAT, the MASQ CAT could be used in clinical practice for single measure purposes. When the final goal is to use the MASQ CAT in ROM, two additional requirements have to be met. First, the MASQ CAT has to measure the same three scales (PA, NA and SA) at different points in time (factorial invariance over time). When patients' values or their internal standards for measurement are changed, comparing observed scale scores could be biased (response-shift; Fokkema, Smits, Kelderman, & Cuijpers, 2013). Second, the responsiveness to change of the MASQ CATs should be equal to the full-scale scores. When instruments' sensitivity to detect change is different, treatment outcomes could be biased (de Beurs et al., 2012). In future research, these requirements have to be investigated to assess the utility of the MASQ CAT in ROM.

When deciding to use a CAT version of the MASQ, either in single measure purposes or in ROM, one has to take into account that all MASQ scales were noninformative for patients on either one or both sides of the latent trait. These patients had a severe lack of PA, a mild SA, or a mild or severe NA. Patients that were located at these sides of the latent trait had no efficiency gain with a CAT administration. Moreover, these patients could have less reliable change scores between different CAT administrations over time. As a future line of research, we propose to investigate whether adding items with either milder or stronger content will result in more uniform test information because of the increased information in the extremes. Adding items in the extremes of the scales with more item information might enhance the reliability of the outcomes and reduce the number of administered items. These benefits would especially apply to ROM, because prior knowledge about the patient can be used more easily to maximize the efficiency gain of the CATs. For example, clinical interviews with the patient could result in expectations about the patient's treatment outcomes. These expectations could be used to personalize the patient's starting value for the CAT administration. Therefore, when the CAT administration starts with an item that links reasonably well to a patient's location on a scale, it could be expected that the number of administered items would drop even further.

Another factor that should be taken into account when using a CAT version of the MASQ is the psychometric quality of the SA scale. For all IRT assumptions (unidimensionality, local independence, monotonicity), this was somewhat lower than for the PA and NA scale. For example, the SA scale had a moderate scalability coefficient while the NA and PA scale both had a high scalability coefficient. As a consequence, the θ estimates of the SA scale contain more error than one would expect on the basis of the specified standard error, $SE(\theta) < 0.3$. This might explain the lower correlation between the CAT θ estimates of the SA scale and the full-scale θ estimates ($r = .89$) compared with the other scales ($r = .98$). A solution to deal with the lower psychometric quality of the SA scale is by setting a minimum number of items the CAT should administer or by specifying a more strict *SE*; for example, $SE(\theta) < 0.25$. For future CATs, researchers should decide on a minimally required correlation between the CAT and the full-scale θ estimates, which might be met by the proposed solutions.

Two last lines of future research, which can be pursued with the MASQ, are the investigation of clinical cut points and their sensitivity and specificity for mood and/or anxiety diagnosis, and the factor structure for patients. In clinical practice, the patient's diagnosis is usually determined by a standardized clinical interview without using the MASQ. When cut points are available for determining mood/anxiety diagnosis with the MASQ, these could be used to assess whether a CAT classification would differ from the classification according to the full MASQ score. Moreover, the MASQ could be used in clinical practice for diagnostic prediction. In addition, some hold the view that the distinction between depressive and anxiety symptoms could best be described by a hierarchical model instead of a three-factor model (Lin et al., 2014; Simms, Gros, Watson, & O'Hara, 2008; Simms, Prisciandaro, Kruger, & Goldberg, 2012). A hierarchical model assumes that anxiety and depression are measured by a general factor and several underlying factors. When this hierarchical model also applies to MASQ data from (Dutch) clinical subjects, then the MASQ could be used to develop a CAT which takes into account the dimensional structure of the combined scales (Reckase, 1985). This type of CAT could enhance the reliability of the scores and the administration efficiency even further.

In this study we investigated CAT for clinical subjects using an existing clinical questionnaire: a potential solution for the time consuming development of new item banks and the administration burden to patients who are completing clinical self-report questionnaires. As a first step to study CAT for the assessment of mental health patients using existing clinical questionnaires, Smits et al. (2012) simulated a CAT on one of the three MASQ scales and suggested that CAT may result in an equally reliable, but more efficient method to collect self-report data. In this study, we repeated the procedure on all three MASQ scales and found that these findings generalize to all scales. Our findings support the feasibility of future development of genuine CATs for using the MASQ to measure clinical subjects.

Chapter 3

Development of a Computer Adaptive Test for Depression Based on the Dutch-Flemish Version of the PROMIS Item Bank

Published as:

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., & de Beurs, E. (2017). Development of a computer adaptive test for depression based on the Dutch-Flemish version of the PROMIS item bank. *Evaluation & the health professions*, *40*(1), 79-105.

3.1 Abstract

We developed a Dutch-Flemish version of the patient-reported outcomes measurement information system (PROMIS) adult v1.0 item bank for Depression as input for computerized adaptive testing (CAT). As item bank, we used the Dutch-Flemish translation of the original PROMIS item bank (28 items) and additionally translated 28 United States (US) depression items that failed to make the final US item bank. Through psychometric analysis of a combined clinical and general population sample ($N = 2{,}010$), 8 added items were removed. With the final item bank, we performed several CAT simulations to assess the efficiency of the extended (48 items) and the original item bank (28 items), using various stopping rules. Both item banks resulted in highly efficient and precise measurement of depression and showed high similarity between the CAT simulation scores and the full item bank scores. We discuss the implications of using each item bank and stopping rule for further CAT development.

Keywords: clinical assessment, computer adaptive test, depression, item response theory, PROMIS

3.2 Background

Routine outcome monitoring (ROM) is the repeated administration of questionnaires over time to monitor patients' progress towards recovery and to adapt the treatment, if indicated (Carlier et al., 2012a; de Beurs et al., 2011). In 2011, ROM has been implemented nationwide in Dutch mental health care. As ROM is used for various aims (treatment monitoring, benchmarking of institutes, and scientific research), the set of questionnaires administered to patients may become extensive which may result in diminished compliance and data loss. Consequently, more efficient measurement in (Dutch) mental health care is essential.

In 2002, the National Institutes of Health started the patient-reported outcomes measurement information system (PROMIS) initiative (Cella et al., 2007, 2010). Their main goal was to develop a new state of the art assessment system for measuring patient-reported health with highly accurate, precise, and short measures. In 2016, this ongoing initiative already brought forward a wide range of item banks (a set of questions with item parameters to measure a construct), which could be used for computerized adaptive testing (CAT). With CAT, the selection of questions is based on the answer(s) to previous questions and the assessment continues until a precise score of the measured latent construct is obtained (i.e., a score is sufficiently free of random error). For example, a patient answers the first item of a depression questionnaire with 5-point Likert-type scale items with response option 1 or 2. Consequently, the next question will be Item 5; otherwise (when response option 3 - 5 would have been chosen) the next item is Item 7. The various response categories for the follow-up question will then, in turn, lead to other items. This selection procedure based on previously given responses continues until the depression score meets the prespecified precision. By asking questions tailored to each patient, CAT can reduce administration burden with a shorter test while maintaining or even improving the precision of the test outcomes for all respondents (Fliege et al., 2005). Furthermore, CAT can select different sets of questions for patients with varying latent trait levels (θ) while the final test outcomes maintain comparability. By administering varied assessments to monitor patients over time, lack of interest in patients may also be avoided. Ultimately, these CAT benefits should decrease respondent burden, increase response rates and reduce possible bias due to selective loss of respondents (Dillman, Sinclair, & Clark, 1993).

The PROMIS initiative showed that the application of CAT results in highly efficient measurement; the PROMIS item banks show highly desirable psychometric properties (Fries, Krishnan, Rose, Lingala, & Bruce, 2011; Fries, Rose, & Krishnan, 2011; Khanna et al., 2011; Magasi et al., 2012; Pilkonis et al., 2011) and, used with CAT, result in highly accurate, precise, and short measures (Pilkonis et al., 2014). In response to these developments, the Dutch-Flemish PROMIS initiative (www.dutchflemishpromis.nl) was started in 2009 to investigate whether the PROMIS methodology could also be successfully implemented in the Netherlands. As a starting point, they translated 17 adult PROMIS item banks and 9 pediatric PROMIS item banks into Dutch-Flemish (Flemish is a variant of the Dutch language spoken in Belgium; Terwee et al., 2014). Among these item banks were the adult v1.0 item banks for mental health constructs Depression, Anxiety, and Anger (Pilkonis et al., 2011). Depression is the leading cause of disability worldwide in terms of total years lost due to disability (Marcus, Yasamy, van Ommeren, Chisholm, & Saxena, 2012), and is the most common mental health disorder in Dutch adults (de Graaf, ten Have, van Gool, & van Dorsselaer, 2012). Therefore, the Depression item bank is an obvious choice to assess whether the PROMIS methodology could be implemented successfully in (Dutch) mental health care.

The aim of the present study was to develop a Dutch-Flemish version of the United States (US) PROMIS adult v1.0 item bank for Depression that could be used for measuring the full latent depression continuum in the Netherlands (i.e., all persons with no symptoms of depression to patients with severe depression). The US item bank comprises 28 items and is based on a selection of items from a larger item bank of 56 items (Pilkonis et al., 2011). In this 56-item bank, items were selected according to favorable psychometric qualities such as unidimensionality, local independence (LI) and monotonicity (Reeve et al., 2007). However, the selection of items for the final 28-item bank was based on the responses of a US sample. As a consequence, the selection of items may be strongly influenced by the American culture/language. Therefore, we chose to translate the original 56-item bank to investigate whether completion by Dutch respondents would result in a similar selection of items for the final item bank. For this purpose, we evaluated the psychometric properties of all 56 items. In addition, we compared the efficiency of the final item bank with the original 28-item bank by performing several post hoc CAT simulations. One of the PROMIS initiative's goals is to implement identical item banks in every country to increase uniformity and enhance comparability. By comparing the extended item bank with the original item bank, we can appraise the implications of using each item bank for further CAT development.

3.3 Method

3.3.1 Participants

For this study, data were collected in two samples: a clinical sample and a general population sample. We chose to include both samples in the item bank construction because our goal is to develop an instrument that covers the full range of possible latent depression levels in the Netherlands. Within this range, the clinical sample mostly covers moderate to high depression levels while the general population mostly covers low to moderate depression levels. We aimed to include a minimum number of 1,000 respondents per sample. A sample size of at least 1,000 is deemed sufficient for adequate item parameter estimates in the item bank calibration (Reise & Yu, 1990).

For the clinical sample, 3,296 patients were invited by the Dutch Mental Health Care provider Parnassia Psychiatric Institute to complete the item set. Patients were referred to this institute by their general practitioner for treatment of common mental disorders in ambulatory mental health care. The patient's diagnosis was assessed with the Dutch translation of the Mini International Neuropsychiatric Interview (MINI-plus; Sheehan et al., 1998) administered by phone by a psychiatric nurse who was extensively trained in the interview. The MINI-plus is a standardized interview for clinical diagnosis of mental disorders following the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV*; American Psychiatric Association, 1994). After the need for treatment was ascertained, the diagnosis was confirmed in a clinical face-to-face assessment.

According to Dutch law, use of data that are collected in the process of routine clinical practice does not require informed consent from patients. However, in accordance with the mental health-care center’s policy, written informed consent was obtained.

From the general population, we needed a random sample to ensure representativeness. Respondents were invited to partake by the data collection panel Desan Research Solutions until at least 1,000 persons participated. Response rates are generally high for this panel, approximately between 60% and 80% (the total number of invitations to panel members was not registered). Respondents participated on a voluntary basis for a small financial compensation. The sample was composed to be in accordance with the Dutch general population distribution regarding five variables in 2013 (www.cbs.nl): gender (male, 49%; female, 51%), age (18-39, 34%; 40-64, 44%; 65+, 22%), education (low, 32%; middle 40%, high 28%), ethnicity (natives, 80%; western immigrants, 10%; nonwestern immigrants, 10%), and region (north, 10%; east, 21%; south, 22%; west, 47%). Deviations in each subgroup were allowed up to 2.5%.

3.3.2 Measures

The Depression item bank consisted of 28 items from the Dutch-Flemish PROMIS adult v1.0 item bank for Depression (Terwee et al., 2014), and 28 US items that did not make it to the final US PROMIS item bank (Pilkonis et al., 2011). The translation of the additional 28 US PROMIS items was performed by four researchers with ample experience in translation of self-report measures; two researchers performed a forward translation of the items, two researchers performed an independent review of these translations. Adjustments were made, until consensus was reached and the translation was approved by all four researchers. Respondents were asked for all 56 items to indicate on a Likert-type scale (1 = *never*, 2 = *rarely*, 3 = *sometimes*, 4 = *often*, and 5 = *always*) how frequently they experienced a wide range of depression symptoms in the past 7 days. All items reflected symptoms, problems, or negative affective states (e.g., Item 1 *I felt fearful*), a higher score meaning more severe depression.

3.3.3 Psychometric Evaluation

We performed a psychometric evaluation of the 56-item bank on the combined patient and general population sample, following the guidelines proposed by Reeve et al. (2007). First, we evaluated several descriptive statistics to assess the performance of the individual depression items and the full Depression item bank. Individual items were evaluated with response frequencies and range, mean, standard deviation (*SD*), skewness, and kurtosis. Furthermore, we explored the interitem correlation matrix, the item-scale correlations, and the drop in coefficient α for each item when removed from the item bank. In addition, the full item bank was evaluated with the sum score range, mean, SD, skewness, kurtosis, and the reliability coefficient for internal consistency.

Second, we evaluated the main item response theory (IRT) assumptions of unidimensionality, LI, and monotonicity to assess whether the Depression item bank is fit to scale respondents and items on a common latent trait. Item banks are considered *unidimensional* when a person's item response results from the person's trait level that the item measures and not from other factors. However, mental health constructs are generally complex and rarely strictly unidimensional. For IRT applications, it is therefore assessed whether the degree of unidimensionality in item banks is sufficient (Reise, Morizot, & Hays, 2007). Unidimensionality was evaluated with exploratory factor analyses (EFA) using the R package psych (Version 1.5.4; Revelle, 2013), and with confirmatory factor analyses (CFA) using the R package lavaan (Version 0.5-18; Rosseel, 2012), both conducted on the polychoric correlation matrix of the items (Bollen, 1989). With EFA, unidimensionality is deemed sufficient when the first factor accounts for at least 20% of the variance (Reckase, 1979, as cited in Hambleton, 1988), and the ratio of explained variance in the first and second factor is higher than 4 (Reeve et al., 2007). With CFA, unidimensionality of the Depression item bank is deemed sufficient when the comparative fit index (CFI) > 0.95, the Tucker-Lewis index (TLI) > 0.95, the root-mean-square error of approximation (RMSEA) < 0.06, and the average absolute residual correlations < 0.10 (Reeve et al., 2007).

The second IRT assumption we evaluated is *LI*. Item pairs are locally independent when, controlling for the trait level, item responses show no association. LI in the depression items was evaluated by inspecting the residual correlation matrix that resulted from the single-factor CFA. Residual correlations higher than .20 were considered as possibly locally dependent (Reeve et al., 2007). Further investigation of LI was done with Yen's Q3 statistic (Yen, 1993), in which the residual item scores under Samejima's graded response model (GRM; Samejima, 1969), fitted with R package mirt (Version 1.10; Chalmers, 2012), are correlated among items. As suggested by Smits, Cuijpers, and van Straten (2011), model fit was evaluated with Cohen's (1988) rules of thumb to interpret effect size; Q3 values between 0.24 and 0.36 imply moderate deviations, Q3 values above 0.37 imply large deviations.

The third IRT assumption we evaluated is *monotonicity*. Items show monotonicity when the probability of selecting an item response that suggests a better health status on a scale increases as the underlying level of health status on that scale is higher. Monotonicity in the depression items was evaluated by examining graphs of item mean scores conditional on rest scores (total raw score minus the item score), using the R package mokken (Version 2.7.7; van den Ark, 2007). This analysis additionally results in scalability coefficients for the full scale and the individual items. A scale or item has low quality when the scalability coefficient is between 0.30 and 0.40, moderate quality when the scalability coefficient is between 0.40 and 0.50, and high quality when the scalability coefficient is above 0.50 (Mokken, 1971).

Subsequently, we evaluated *differential item functioning* (DIF; Embretson & Reise, 2000) to assess whether persons from different groups have equal probabilities of selecting item response categories. An item shows DIF when the probability of responding in different response categories differs across independent groups, controlling for the trait level influencing a person's item response. We explored DIF for gender (men, women), age (18-39, 40-64, 65+), and education level (low, medium, and high). DIF among the depression items was evaluated with ordinal logistic regression (OLR; Crane, Gibbons, Jolley, & van Belle, 2006), using the R package lordif (Version 0.2-2; Choi, Gibbons, & Crane, 2011). Effect size was evaluated by means of change in McFadden’s pseudo R^2, following the suggestion of a critical value of 0.02 (Choi et al., 2011) for rejecting the hypothesis of no DIF.

Finally, we calibrated the extended item bank with Samejima’s GRM (Samejima, 1969), using the R package mirt (Version 1.10; Chalmers, 2012). We fitted the GRM with multiple group estimation (McDonald, 1999; Smits, 2016) for which we used the combined clinical and general population sample and specified population as grouping factor with constraints on equal discrimination and threshold parameters. The latent trait was scaled to a mean of 0 and a *SD* of 1 for the general population. In addition, we performed a calibration on the original 28 items of the PROMIS adult v1.0 item bank for Depression to compare efficiency results with the extended item bank (see “CAT Simulations” subsection). Note that from here on all items from the original item bank are mentioned as “original item” and all additional items from the extended item bank are mentioned as “added item”.

The calibrations of the extended and original item bank under the GRM were evaluated by examining item fit and item properties. First, item fit was evaluated with the $S\text{-}X^2$ statistic (Orlando & Thissen, 2000, 2003), which compares observed and expected response frequencies under the used IRT model and quantifies differences between these frequencies. Items with a $S\text{-}X^2$ p < .001 are considered to have a poor fit in the IRT model (Reeve et al., 2007). Second, item properties were evaluated by examining *a* (discrimination) and *b* (threshold) parameter estimates. The discrimination parameter represents the extent to which persons with similar scores on the latent trait can be differentiated by the item. The four threshold parameters *b* (the number of threshold parameters for an item is equal to the number of response categories minus one) represent the θ locations on which a person is expected to choose from a lower to a higher item response. In addition, we compared the item parameter estimates of the first 28 items between the extended and the original item bank, using differences in means and SD’s (extended minus original), and Pearson’s correlations.

3.3.4 CAT Simulations

To assess the efficiency of the extended and the original item bank, we performed an individual post hoc CAT simulation with each item bank, using the R package mirtCAT (Version 0.5; Chalmers, 2015). A CAT simulation is not an actual CAT administration, but selects the item responses and evaluates them as if they had been collected adaptively. We split the clinical and general population samples randomly into half; the first half of both samples was used for estimating the item parameters, the second half for simulating CAT. This method will prevent overfitting (Hastie, Tibshirani, & Friedman, 2001), which would have resulted in outcomes that are too optimistic. Note that we estimated the item parameters again to perform this analysis. Thus, the item parameters of the full clinical and general population sample are to be used as input for a future CAT, the item parameters of half of the clinical and general population sample are used in this study as input for simulating CAT.

We chose to perform the primary CAT simulations on the clinical sample because clinical subjects were deemed the most relevant group to measure depression. In addition, we also performed CAT simulations with each item bank using the general population sample and briefly mention some main results. It could be expected that the efficiency gains are higher for the clinical sample compared to the general population sample because the information value of items is generally lower for respondents with low values of the latent trait (low levels of depression; Reise & Waller, 2009).

The CAT simulations started with the item that had the highest item information value at the average value of the latent trait (θ = 0; Embretson & Reise, 2000; Wainer, Dorans, Flaugher, Green, & Mislevy, 2000). Consequently, the CAT simulations started with the original item Emotional Distress – Depression item bank, Item 36 (EDDEP36) *I felt unhappy* for both the extended and the original item bank (note that we use the original US item coding; www.assessmentcenter.net). The depression latent trait scores (θ) were then estimated with the Bayesian method maximum a posteriori (MAP; Embretson & Reise, 2000), and a standard error (*SE*) was calculated. The CAT simulation stopped selecting new items when the patient's θ reached a prespecified value of the *SE*. Otherwise, new items were selected using the highest item information at the provisional θ estimate until the prespecified value of the *SE* was obtained or when all items were selected without obtaining the *SE*. We evaluated several stopping rules: $SE(\theta) < 0.1$, $SE(\theta) < 0.2$, $SE(\theta) < 0.3$, and $SE(\theta) < 0.4$. For each stopping rule, several statistics were recorded individually for both the extended and the original item bank to assess the efficiency of CAT: (a) the mean and *SD* of the number of selected items, (b) the percentage of all patients for whom all items had to be selected, and (c) the mean *SE* of the final θ estimate for all patients. In addition, we investigated the efficiency of CAT under each stopping rule by plotting the number of selected items for each patient with the test information of each item bank. Test information displays how precisely an item bank can measure a latent trait, given the location of the person's estimate. It is calculated as the sum of all item information values at any relevant θ level.

3.3.5 Comparing full-scale data with CAT data

Through CAT simulations, we could assess the similarity between patients' estimated CAT θ scores and patients' estimated full item bank θ scores. For this analysis, we used the patients from the (CAT simulation) clinical sample ($n = 504$). First, similarity between the depression scores was assessed with Pearson's correlation. Second, we assessed the effect size between both depression scores using Cohen's *d* (with pooled *SD*'s), which was evaluated using the guideline proposed by Cohen (1988): 0.2 = small effect, 0.5 = medium effect, and 0.8 = large effect. We performed these analyses for the original and the extended item bank under all stopping rules.

The CAT simulations also enabled us to assess whether depressed persons systematically differed in θ estimates from persons without a diagnosis, as a minimal requirement for predictive validity. For this analysis, we compared scores of persons with a mood disorder ($n = 161$) to the scores from persons without a diagnosis ($n = 449$). Cohen's *d* (with pooled *SD*'s) was assessed for the original and the extended item bank under all stopping rules, including the full-scale estimates (no stopping rule).

3.4 Results

3.4.1 Demographic characteristics

From the 3,296 invited patients, 1,032 completed the questionnaire (response rate = 31.3%). We did not find differences between responders and nonresponders for the variables gender and age. Among the 1,032 respondents, 24 patients were excluded from the analyses because they did not complete all 56 items. Therefore, the final clinical sample consisted of *n* = 1,008 patients (61.7% female). The mean age of the patients was 40.2 years (*SD* = 12.9, range 19–76). Patients' diagnoses (*DSM-IV*) were classified as follows: 44% had a mood disorder, 33% had an anxiety disorder, and 23% had another disorder (e.g., attention deficit disorder, somatoform disorder, personality disorder, etc.).

From the 1,055 respondents of the general population, 53 persons were excluded because they showed suspicious response patterns (e.g., all responses in one category). Therefore, the final general population sample consisted of *n* = 1,002 persons from the Dutch population. The mean age of the general population sample was 50.5 years (*SD* = 16.5, range 19–102). Regarding demographics, the sample was composed as follows: gender (male, 49%; female, 51%), age (18-39, 34%; 40-64, 44%; 65+, 22%), education (low, 31%; middle 40%; high 29%), ethnicity (natives, 80%; western immigrants, 13%; nonwestern immigrants, 7%), and region (north, 12%; east, 20%; south, 21%; west, 47%).

3.4.2 Psychometric evaluation

For the psychometric evaluation of the data, the clinical sample and the general population sample were combined (56 items; *N* = 2,010). The extended item bank data did not show outliers in response frequencies of the depression items, mean, *SD*, range, skewness, and kurtosis. Furthermore, the data showed a high internal consistency reliability (Cronbach's α = .99). However, the added item EDDEP11 *I ate more than usual* showed a very small negative interitem correlation (*r* = -0.02) with the added item EDDEP49 *I lost weight without trying*. This negative correlation is also implied by the content of the items, as the item *I ate more than usual* is implicitly about gaining weight.

All CFA fit indices resulted in a good fit (CFI = 0.99; TLI = 0.99; average absolute residual correlations = 0.04), except for the RMSEA, which resulted in a moderate fit (RMSEA = 0.09). With EFA, the proportion of variance explained by the first factor was 68% which is above the Reckase criterion of 20% (Reckase, 1979, as cited in Hambleton, 1988). In addition, the ratio of variance explained by the first and second factor was 17, which was also higher than the minimal requirement of 4 (Reeve et al., 2007). We concluded that the extended item bank sufficiently met the assumption of unidimensionality.

Of all item pairs, 8 added item pairs were considered possibly locally dependent as their residual correlations were above .20 (Reeve et al., 2007). Further investigation of these items with Yen's Q3 statistic showed 3 item pairs with high deviations (item pairs EDDEP32 *I wished I were dead and away from it all* – EDDEP33 *I thought about suicide*, EDDEP32 *I wished I were dead and away from it all* – EDDEP40 *I felt that others would be better off if I were dead*, and EDDEP33 *I thought about suicide* – EDDEP40 *I felt that others would be better off if I* *were dead*), 3 item pairs with moderate deviations (item pairs EDDEP11 *I ate more than usual* – EDDEP15 *I disliked the way my body looked*, EDDEP49 *I lost weight without trying* – EDDEP53 *I had little desire to eat,* and EDDEP16 *I felt like crying* – EDDEP34 *I had crying spells*), and 2 item pairs with no deviations (item pairs EDDEP11 *I ate more than usual* – EDDEP49 *I lost weight without trying,* and EDDEP11 *I ate more than usual* – EDDEP53 *I had little desire to eat*). Items with residual correlations > .20, high deviations in Yen's Q3 statistic, and other poor psychometric properties were removed from the item bank.

The graphs of item mean scores conditional on rest scores showed monotonicity for all 56 depression items. In addition, the scalability coefficient of the Depression item bank was high (.64), and the scalability coefficient for all depression items was above the lower bound of .30 (Mokken, 1971). We concluded that the extended item bank sufficiently met the assumption of monotonicity.

The 56 depression items showed no DIF for age and education level. For gender, the added items EDDEP16 *I felt like crying* and EDDEP34 *I had crying spells* were flagged for DIF. Change in McFadden's R^2 was .03 for both items, which was above the threshold of .02 (Choi et al., 2011).

Based on the statistical results, we chose to remove the added items EDDEP11 *I ate more than usual,* EDDEP49 *I lost weight without trying,* EDDEP16 *I felt like crying,* and EDDEP34 *I had crying spells*. First, item EDDEP11 *I ate more than usual* and EDDEP49 *I lost weight without trying* for having a small negative correlation with each other. Both are symptoms of depression, but cannot occur at the same time in a single person. Therefore, the item response for one of these items could result in bias because it is not clear which item can be seen as a depression symptom in a person. Second, we removed item EDDEP16 *I felt like crying* and EDDEP34 *I had crying spells* for having DIF on gender. Based on content, we additionally chose to remove the added items EDDEP53 *I had little desire to eat* and EDDEP55 *I felt like I needed help for my depression*. First, EDDEP53 *I had little desire to eat* because just as items EDDEP11 *I ate more than usual* and EDDEP49 *I lost weight without trying* both the confirmation and the rejection of this item can be seen as a depression symptom in different persons. Second, EDDEP55 *I felt like I needed help for my depression* because this item is not appropriate for healthy respondents. After removing these items, we reevaluated all psychometric qualities of the extended 50-item bank and found that they had all improved slightly.

In the calibration of the remaining items, we found five $S\text{-}X^2$ *p*-values below .001 for the extended 50-item bank (original items EDDEP42 and EDDEP46; added items EDDEP32, EDDEP38 and EDDEP40) and seven $S\text{-}X^2$ *p*-values below .001 for the original 28-item bank (original items EDDEP09, EDDEP21, EDDEP27, EDDEP39, EDDEP42, EDDEP44 and EDDEP54). Based on content and other psychometric properties, we chose to remove the added items EDDEP32 *I wished I were dead and away from it all* and EDDEP40 *I felt that others would be better off if I were dead*; both items showed a high degree of local dependency with item EDDEP33 *I thought about suicide*. After the 48-item bank was recalibrated, we did not find other items that needed to be removed. In addition, the correlation between the estimated latent trait scores (θ) under the full item banks (extended item bank, 48 items; original item bank, 28 items) and the sum of raw scores under the full item banks was high for both the original and the extended item bank (r = .99). We concluded that the GRM fitted the extended 48-item bank and the original 28-item bank sufficiently.

Table 3.1 Discrimination and threshold parameter estimates for the extended and original PROMIS item bank for Depression.

		Extended item bank					Original item bank				
Item code	Item	Item parameter estimates					Item parameter estimates				
		a	b_1	b_2	b_3	b_4	a	b_1	b_2	b_3	b_4
EDDEP04	I felt worthless	2.718	-0.109	0.606	1.412	2.528	2.940	-0.030	0.680	1.493	2.618
EDDEP05	I felt that I had nothing to look forward to	2.638	-0.063	0.615	1.423	2.500	2.747	0.007	0.687	1.508	2.611
EDDEP06	I felt helpless	2.781	0.038	0.772	1.530	2.657	2.818	0.104	0.847	1.625	2.793
EDDEP07	I withdrew from other people	2.400	-0.170	0.573	1.408	2.610	2.336	-0.113	0.646	1.506	2.765
EDDEP09	I felt that nothing could cheer me up	3.446	0.091	0.746	1.467	2.321	3.493	0.156	0.818	1.561	2.456
EDDEP14	I felt that I was not as good as other people	2.672	0.006	0.641	1.329	2.257	2.746	0.074	0.714	1.418	2.373
EDDEP17	I felt sad	2.693	-0.504	0.218	1.195	2.309	2.700	-0.446	0.285	1.279	2.437
EDDEP19	I felt that I wanted to give up on everything	2.700	0.143	0.810	1.604	2.599	2.672	0.204	0.885	1.707	2.746
EDDEP21	I felt that I was to blame for things	2.666	0.246	0.934	1.614	2.493	2.689	0.312	1.010	1.711	2.626
EDDEP22	I felt like a failure	3.243	0.360	0.904	1.494	2.254	3.498	0.430	0.978	1.581	2.364
EDDEP23	I had trouble feeling close to people	2.048	0.040	0.865	1.711	2.837	1.963	0.095	0.946	1.826	3.011
EDDEP26	I felt disappointed in myself	3.209	-0.141	0.497	1.249	2.196	3.272	-0.076	0.564	1.335	2.325
EDDEP27	I felt that I was not needed	2.627	0.121	0.805	1.545	2.522	2.644	0.186	0.877	1.641	2.661
EDDEP28	I felt lonely	2.702	-0.120	0.556	1.314	2.214	2.723	-0.055	0.628	1.401	2.334
EDDEP29	I felt depressed	3.450	0.114	0.713	1.304	2.126	3.388	0.174	0.784	1.400	2.265
EDDEP30	I had trouble making decisions	2.429	-0.289	0.533	1.381	2.436	2.241	-0.253	0.597	1.489	2.621
EDDEP31	I felt discouraged about the future	3.232	-0.113	0.512	1.198	2.100	3.316	-0.049	0.577	1.280	2.219
EDDEP35	I found that things in my life were overwhelming	2.564	0.216	0.919	1.704	2.660	2.422	0.272	0.999	1.825	2.843
EDDEP36	I felt unhappy	3.946	-0.103	0.557	1.186	2.068	4.111	-0.034	0.626	1.269	2.192
EDDEP39	I felt I had no reason for living	2.578	0.842	1.349	1.965	2.732	2.581	0.913	1.437	2.079	2.870
EDDEP41	I felt hopeless	3.935	0.482	1.016	1.690	2.520	3.904	0.540	1.090	1.803	2.691
EDDEP42	I felt ignored by people	2.276	0.383	1.156	2.051	3.185	2.143	0.441	1.247	2.199	3.403
EDDEP44	I felt upset for no reason	2.717	0.377	0.980	1.670	2.821	2.488	0.432	1.061	1.795	3.042
EDDEP45	I felt that nothing was interesting	3.283	0.266	0.885	1.554	2.496	3.075	0.322	0.960	1.667	2.682
EDDEP46	I felt pessimistic	3.057	0.047	0.671	1.377	2.350	2.978	0.105	0.736	1.471	2.502
EDDEP48	I felt that my life was empty	3.431	0.204	0.769	1.342	2.196	3.469	0.268	0.842	1.435	2.333
EDDEP50	I felt guilty	2.773	0.195	0.807	1.502	2.429	2.683	0.255	0.880	1.605	2.585
EDDEP54	I felt emotionally exhausted	3.496	-0.017	0.556	1.085	1.952	3.124	0.032	0.617	1.171	2.107
EDDEP12	I had mood swings	2.592	0.133	0.715	1.453	2.418					
EDDEP43	I felt slowed down	2.688	0.167	0.737	1.484	2.502					
EDDEP10	I was critical of myself for my mistakes	2.219	-0.027	0.590	1.344	2.256					
EDDEP56	I had trouble enjoying things that I used to enjoy	3.137	0.041	0.560	1.142	2.080					
EDDEP13	I felt that other people did not understand me	2.845	-0.052	0.601	1.328	2.294					
EDDEP51	I lost interest in my appearance	1.694	0.297	1.147	2.089	3.171					
EDDEP03	I felt that I had no energy	2.458	-0.229	0.391	1.113	2.091					
EDDEP08	I felt that everything I did was an effort	2.919	-0.255	0.413	1.142	2.173					
EDDEP15	I disliked the way my body looked	1.503	-0.031	0.795	1.635	2.586					
EDDEP18	I got tired more easily than usual	2.246	-0.029	0.508	1.184	2.252					
EDDEP24	I felt like being alone	1.863	-0.123	0.503	1.397	2.641					
EDDEP01	I reacted slowly to things that were said or done	2.447	0.327	1.051	1.899	2.767					
EDDEP20	My thinking was slower than usual	2.365	0.179	0.885	1.676	2.679					
EDDEP33	I thought about suicide	1.868	1.342	1.887	2.658	3.547					
EDDEP38	I felt unloved	2.572	0.624	1.220	1.802	2.651					
EDDEP47	I had trouble keeping my mind on what I was doing	2.447	-0.051	0.575	1.356	2.619					
EDDEP52	I had trouble thinking clearly	2.720	0.128	0.755	1.540	2.582					
EDDEP37	I was unable to do many of my usual activities	2.146	0.629	1.259	1.973	3.086					
EDDEP02	I felt lonely even when I was with other people	3.229	0.507	0.937	1.590	2.478					
EDDEP25	I had bad dreams that upset me	1.746	0.735	1.407	2.150	3.207					

In Table 3.1, the final item parameter estimates of the extended 48-item bank and the original 28-item bank are displayed (N = 2,010; clinical sample, n = 1,008 and general population sample, n = 1,002). The item parameter estimates of both the extended 48-item bank and the original 28-item bank showed considerable variation. For the extended 48-item bank, the item parameter estimates ranged from a = 1.503 (added item EDDEP15 *I disliked the way my body looked*) to a = 3.946 (original item EDDEP36 *I felt unhappy*), and from b_1 = -0.504 (original item EDDEP17 *I felt sad*) to b_4 = 3.547 (added item EDDEP33 *I thought about suicide*). For the original 28-item bank, the item parameter estimates ranged from a = 1.963 (EDDEP23 *I had trouble feeling close to people*) to a = 4.111 (EDDEP36 *I felt unhappy*), and from b_1 = -0.446 (EDDEP17 *I felt sad*) to b_4 = 3.403 (EDDEP42 *I felt ignored by people*). In addition, the comparison between the matching 28 items of the extended and original item bank showed high Pearson's correlations (r_a = .97, r_{b1} = 1.00, r_{b2} = .99, r_{b3} = .97, r_{b4} = .96), small differences in means (m_a = 0.02, m_{b1} = 0.17, m_{b2} = 0.19, m_{b3} = 0.25, m_{b4} = 0.36), and small differences in *SD*s (SD_a = -0.04, SD_{b1} = 0.00, SD_{b2} = -0.02, SD_{b3} = -0.07, SD_{b4} = -0.12).

3.4.3 Efficiency of CAT using different stopping rules

In Table 3.2, the CAT simulation outcomes for the clinical sample are displayed for the extended and original item bank under each stopping rule (n = 504). Evidently, both the mean number of selected items and the number of patients for whom the full item banks were selected declined, as the stopping rule was less strict.

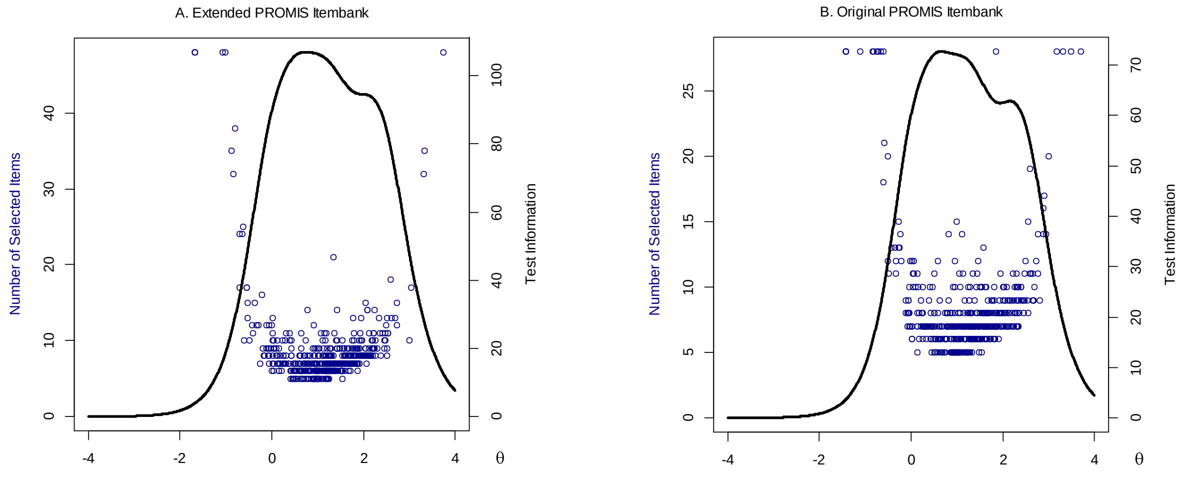
Table 3.2 Patients' CAT simulation statistics for the extended and original PROMIS item bank for Depression under several stopping rules.

	Extended item bank				Original item bank			
Stopping rule	Number of items				Number of items			
	M	*SD*	% All	Mean *SE*(θ)	*M*	*SD*	% All	Mean *SE*(θ)
SE(θ) <0.1	44.29	5.07	56.2	.11	28.00	0.00	100.0	.13
SE(θ) <0.2	8.69	5.68	1.2	.20	8.40	4.45	3.6	.20
SE(θ) <0.3	3.48	4.04	0.6	.28	3.40	3.33	1.4	.28
SE(θ) <0.4	2.09	3.83	0.6	.35	2.03	2.76	1.0	.35

Apart from stopping rule *SE*(θ) < 0.1, the extended and original item bank show highly similar results. Apparently, stopping *SE*(θ) < 0.1 is too strict for both item banks as the simulations selected all items for a high percentage of patients (Table 3.2, column 4 and 8). This is especially the case with the original item bank (100% full item bank selections) due to its relative low number of items (28 in the original item bank to 48 in the extended item bank). From stopping rule *SE*(θ) < 0.2, however, the mean number of selected items dropped substantially for both item banks, following a similar pattern (Table 3.2, column 2 and 6). Under stopping rule *SE*(θ) < 0.2, the mean number of selected items is around 8.54, and then dropped even further to 3.44 under stopping rule *SE*(θ) < 0.3, and 2.06 under stopping rule *SE*(θ) < 0.4. These stopping rules also result in a much smaller percentage of patients for whom all items were selected (below 4%). Overall, the efficiency of the original item bank is slightly higher. This result is an effect of the large difference in the number of items in each item bank. As a consequence, the mean number of selected items from the extended item bank is somewhat inflated by the group of patients for whom (almost) all items were selected.

As example, Figure 3.1 shows the test information along with the number of selected items under stopping rule *SE*(θ) <0.2 for both the extended (1A) and the original (1B) item bank. Evidently, test information is higher for most θ values in the extended item bank due to the larger number of items. However, the shape of the test information curve is similar for both item banks, meaning that test information is high for -0.5 < θ < 3 and low for θ < -0.5 (very low depression score) or θ > 3.0 (very high depression score). Obviously, the number of selected items is linked to the test information, because large number of items were selected for patients with θ estimates at the end of the scales (low-test information). In contrast, only 5 or 6 items were selected for most patients with θ estimates in the middle of the scale (high-test information). This pattern was shown for all stopping rules for both item banks, naturally with a decline in number of selected items as the stopping rule was less strict. Under stopping rule *SE*(θ) < 0.4, for example, only 1 or 2 items were selected for patients with θ estimates that showed high-test information.

Figure 3.1 Number of selected items shown as a function of the final θ estimate under stopping rule *SE*(θ) < 0.2 for the extended and original PROMIS item bank.



Finally, the CAT simulation outcomes for the general population showed, as expected, less efficiency gains compared to the clinical sample. Naturally, most respondents from the general population had θ estimates at the lower end of the depression scale (very low depression scores), which indicates very low-test information. Consequently, the mean number of selected items increased. For example, under stopping rule *SE*(θ) < 0.2, the mean number of selected items was 19 with the extended item bank and 14 with the original item bank.

3.4.4 Comparing full-scale data with CAT data

In Table 3.3, Pearson's correlations and sizes of difference (Cohen's *d*) between patients' CAT simulation θ estimates and patients' full item bank θ estimates are displayed for the extended and original item bank under each stopping rule (n = 504). Note that the results regarding the mean and *SD* of both item banks cannot be compared directly. Because the datasets are different (i.e., the number of items), the metric of the scales is also slightly different. As a result, the extended and the original item bank show a small difference in mean and *SD* of the θ estimates (extended item bank: full-scale θ, *M* = 1.15 and *SD* = 0.79; original item bank: full-scale θ, *M* = 1.21 and *SD* = 0.83).

Table 3.3 Pearson's correlations and sizes of difference (Cohen's d) between patients' CAT simulation θ estimates and patients' full item bank θ estimates for the extended and original PROMIS item bank for Depression under several stopping rules.

	Extended Item Bank				Original Item Bank			
	CAT θ				CAT θ			
Stopping Rule	*M*	*SD*	*r*	*d*	*M*	*SD*	*r*	*d*
SE(θ) < 0.1	1.15	.79	1.00	.00	1.21	.83	1.00	.00
SE(θ) < 0.2	1.14	.78	.96	.01	1.19	.81	.97	.02
SE(θ) < 0.3	1.11	.78	.92	.05	1.14	.81	.94	.08
SE(θ) < 0.4	1.03	.78	.87	.14	1.08	.78	.89	.14

Evidently, Pearson's correlations declined and sizes of difference increased as the stopping rule was less strict. Again, the extended and original item bank showed highly similar results. Pearson's correlations were high under all stopping rules, ranging from 1.00 under stopping rule *SE*(θ) < 0.1 to 0.87 (extended item bank) and 0.89 (original item bank) under stopping rule *SE*(θ) < 0.4 (Table 3.3, column 4 and 8). In addition, Cohen's *d* values indicated a negligible to a very small effect under all stopping rules, ranging for both item banks from 0 under the stopping rule *SE*(θ) < 0.1 to 0.14 under stopping rule *SE*(θ) < 0.4 (Table 3.3, column 5 and 9). Specifically, patients' mean CAT simulation θ estimates under stopping rule SE(θ) < 0.1 were equal to patients' mean full item bank θ estimates, and declined as the stopping rule was less strict. For clinical practice this would imply that less strict stopping rules yield slightly lower depression scores with CAT.

Table 3.4 presents the sizes of difference (Cohen's *d*) between the θ estimates of persons with a mood disorder and the θ estimates of persons without a diagnosis for the extended and original PROMIS item bank under several stopping rules. Cohen's *d* was large under each stopping rule and nearly identical for the extended and original item bank, ranging from 1.41 (extended item bank) and 1.45 (original item bank) under the full item bank (no stopping rule) to 1.22 for both item banks under stopping rule *SE*(θ) < 0.4. The results indicate that depressed patients have a much higher θ estimate on the depression scale than persons without a diagnosis, and this difference declines somewhat when the stopping rule is less strict.

Table 3.4 Sizes of difference (Cohen's d) between the θ estimates of persons with a mood disorder and persons without a diagnoses for the extended and original PROMIS item bank for Depression under several stopping rules.

	Extended item bank					Original item bank				
	Mood disorder		No diagnosis			Mood disorder		No diagnosis		
Stopping rule	*M*	*SD*	*M*	*SD*	*d*	*M*	*SD*	*M*	*SD*	*d*
None: θ	.94	.72	-.24	.88	1.41	.99	.75	-.22	.86	1.45
SE(θ) < 0.1	.94	.72	-.24	.88	1.41	.94	.72	-.22	.86	1.41
SE(θ) < 0.2	.91	.69	-.22	.88	1.35	.91	.69	-.21	.87	1.36
SE(θ) < 0.3	.89	.68	-.17	.85	1.32	.89	.68	-.16	.84	1.32
SE(θ) < 0.4	.82	.71	-.13	.80	1.22	.82	.71	-.12	.80	1.22

3.5 Discussion

In this study, we evaluated the Dutch-Flemish version of the US PROMIS adult v1.0 item bank for Depression with data from a sample of patients with mental health problems and a sample from the Dutch general population. We started with a 56-item bank that was also used in the US validation study (Pilkonis et al., 2011). In the US, the validation of the Depression item bank resulted in 28 items (original item bank). Although all 28 items were retained in our study, we found a total of 48 items with desirable psychometric qualities (extended item bank). These psychometric qualities included sufficient unidimensionality, LI and monotonicity, and absence of DIF. Furthermore, the 48-item bank showed a sufficient fit with the GRM (Samejima, 1969).

We compared the original and extended item bank using a post hoc CAT simulation and found that the efficiency of both item banks for patients was highly similar, with a slight superiority for the original item bank. Therefore, based on efficiency, the original bank could also be used for CAT implementation. Using the smaller 28-item bank has the additional benefit of enhanced international comparability between the Dutch-Flemish and the US item banks for the assessment of depression. To investigate comparability further, future research should address factorial invariance and DIF between countries.

CAT methodology is not only aimed at improving efficiency, but also at varied assessments for patients with differing θ estimates. Within the treatment process, this CAT characteristic is also advantageous with repeated administrations over time. By administering varied assessments to monitor patients' health progress, diminished attentiveness may be avoided. This benefit might be more clearly visible in the extended item bank than the original item bank due to the larger number of items. For future research, we therefore recommend the US PROMIS group to assess whether the original 28-item US PROMIS adult v1.0 item bank for Depression could be extended with newly validated items.

Using CAT to assess respondents, it is common to adopt stopping rule *SE*(θ) < 0.3 (e.g., Becker et al., 2008; Gibbons et al., 2014). This stopping rule is comparable to a marginal reliability of .90 (Green, Bock, Humphreys, Linn, & Reckase, 1984), which is generally required for minimal reliability for individual assessments (Bernstein & Nunnally, 1994, p. 265). Our findings suggest that stopping rule *SE*(θ) < 0.3 would be a sound choice for using CAT with the original or the extended Depression item bank. Under this stopping rule, the mean number of selected items was very low (extended item bank, *m* = 3.48; original item bank, *m* = 3.40), patients' CAT simulation θ estimates showed a sufficient similarity with patients' full item bank θ estimates, and patients with a depression diagnosis differed substantially in θ estimates from persons without a diagnosis. However, our ultimate goal is to use CAT in ROM to monitor patients' progress over time. To assess significant change, high levels of individual test precision are required. Significant change can be expressed with the IRT-based *Z*-test (Brouwer, Meijer, & Zevalkink, 2013) using pretest and posttest data. The pre-post difference needed to deem a patient as significantly changed is dependent of the *SE* of measurement. With a lower *SE* of pretest and posttest, we will be better able to detect true change. It may therefore be more suited to use a stopping rule requiring more precision such as *SE*(θ) < 0.2. Under this stopping rule, the mean number of selected items is still very acceptable (extended item bank, *m* = 8.69; original item bank, *m* = 8.40), the similarity between patients' CAT simulation θ estimates and patients' full item bank θ estimates is substantial for both item banks, and depressed patients differed substantially in θ estimates from persons without a diagnosis.

When choosing a stopping rule, researchers should also take into account the maximum number of items the CAT software should administer to increase the efficiency for each individual. This can be done by setting a fixed number of maximum items, or by incorporating (state of the art) stopping rules which take into account whether additional items will increase the precision or change the estimated latent trait value of the assessment (e.g., predicted SE reduction [PSER], Choi, Grady, & Dodd, 2011; change in θ, Babcock & Weiss, 2013). Using one of these methods is especially useful for persons with very high or very low depression levels. For such persons, test information is low which could result in the administration of all items in the item bank without ever meeting the *SE* stopping rule. Limiting the maximum number of items to be administered should therefore not result in an unacceptable diminishment of precision of the test result. Consequently, the slight inferiority of the extended item bank in this study should diminish because the mean number of selected items is no longer affected by individuals for whom all or most of the 48 items were selected.

After choosing the item bank and stopping rule, the Dutch-Flemish version of the PROMIS adult v1.0 item bank for Depression could be used in clinical practice for single measure purposes to assess the level of depressive symptomatology. For utilizing CAT specifically in diagnostic prediction, future research needs to further address predictive validity using patients´ diagnoses. For utilizing CAT in ROM, future research needs to address measurement invariance over time (Fokkema, Smits, Kelderman, & Cuijpers, 2013) and whether responsiveness to change of CAT θ estimates is equal to full item bank θ estimates or to responsiveness of legacy measures (de Beurs et al., 2012).

A possible limitation of the present study is that the results regarding the efficiency of the item banks were assessed with CAT simulations and not with real CAT administrations. Although another study has shown that the outcomes of CAT simulations and real CAT administrations can be very similar (Kocalevent et al., 2009), replications of these results are necessary. For example, the CAT simulations results were based on item parameters from a smaller sample (*n* = 1,004) and therefore could differ somewhat from real CAT administration results that are based on item parameters from a larger sample (*N* = 2,008). Furthermore, the correlations and the sizes of difference between patients' CAT simulation θ estimates and patients' full item bank θ estimates could respectively be inflated and deflated as the data derive from the same assessment. An independent administration with both the full item bank and the CAT in the same subjects could provide useful information about the utility of CAT simulations to assess CAT efficiency gains.

Another factor that should be taken into account when using CAT is the influence of shrinkage in the θ estimates of the Bayesian estimation method MAP (Embretson & Reise, 2000). Shrinkage basically means that the use of a prior normal distribution pulls θ estimations towards the mean, especially with early θ estimations. As a consequence, θ could be somewhat over- or underestimated for patients with a low number of selected items. This effect might explain the slightly diminishing mean in θ estimates as the stopping rules were less strict (Table 3.3). A solution to deal with the influence of shrinkage in Bayesian θ estimation is by setting a minimum number of items the CAT should administer or by using a different estimation method (e.g., maximum likelihood; Smits, 2016).

In this study, we showed that the PROMIS methodology results in efficient measurement of depression in Dutch patients. The Dutch-Flemish PROMIS item banks (extended and original) show desirable psychometric qualities, and applied as a CAT, could result in short and precise measurement. These favorable results were also found in other countries using different translations of the PROMIS item banks (e.g., German, Jakob et al., 2015; Spanish, Vilagut et al., 2015). We therefore encourage researchers in other countries to investigate whether the PROMIS methodology is efficient and valid for assessment of depression in their clinical and general population.

Chapter 4

Development of a Computerized Adaptive Test for Anxiety Based on the Dutch–Flemish Version of the PROMIS Item Bank

Published as:

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., Spinhoven, P., & de Beurs, E. (2019). Development of a computerized adaptive test for anxiety based on the Dutch–Flemish version of the PROMIS item bank. *Assessment*, *26*(7), 1362-1374.

4.1 Abstract

We used the Dutch–Flemish version of the United Stated PROMIS adult v1.0 item bank for Anxiety as input for developing a computerized adaptive test (CAT) to measure the entire latent anxiety continuum. First, psychometric analysis of a combined clinical and general population sample (*N* = 2,010) showed that the 29-item bank has psychometric properties that are required for a CAT administration. Second, a post hoc CAT simulation showed efficient and highly precise measurement, with an average number of 8.64 items for the clinical sample, and 9.48 items for the general population sample. Furthermore, the accuracy of our CAT version was highly similar to that of the full item bank administration, both in final score estimates and in distinguishing clinical subjects from persons without a mental health disorder. We discuss the future directions and limitations of CAT development with the Dutch–Flemish version of the PROMIS Anxiety item bank.

Keywords: assessment, anxiety, clinical subjects, general population, item response theory, computerized adaptive test, PROMIS

4.2 Background

In 2002, the National Institutes of Health started the patient-reported outcomes measurement information system (PROMIS) initiative in the Unites States (US). PROMIS has the ambition to combine and transform all existing patient-reported outcome measures (PROMs) into one state of the art assessment system for measuring self-reported health (Cella et al., 2007, 2010). With this system, self-reported health of adults and children is measured more accurately, precisely, responsively, and efficiently than existing PROMs allow for (Fries, Krishnan, Rose, Lingala, & Bruce, 2011; Fries, Rose, & Krishnan, 2011; Magasi et al., 2012; Pilkonis et al., 2014; Schalet et al., 2016). This is accomplished by the development of item banks (i.e., sets of items that measure the construct of interest) that meet high psychometric standards (i.e., good quality item parameters). These item banks may be administered through a fixed questionnaire with a low number of items (also known as short forms), but preferably through a computerized adaptive test (CAT; Reeve et al., 2007). With short forms, the measurement precision for test outcomes can vary among respondents. A CAT, however, is more dynamic. It is a computer-administered test that selects questions based on the response pattern on previous questions until a precise outcome is obtained. In other words: it fixes the test outcomes' measurement precision and allows for the number of administered items to vary among respondents (Embretson & Reise, 2000). Consequently, administration burden can be reduced with a shorter test while maintaining the precision of the test result (Fliege et al., 2005).

PROMIS has become increasingly popular in the US, and in other countries as well. By early 2017, many countries had developed translations of PROMIS item banks (www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis/available-translations). Moreover, several countries had evaluated at least one item bank psychometrically (e.g., Depression item bank: German, Jakob et al., 2015; Spanish, Vilagut et al., 2015). In the Netherlands, PROMIS is also gradually being implemented. First, 17 adult item banks and 9 pediatric item banks have been translated into Dutch-Flemish (Flemish is a variant of the Dutch language spoken in Belgium; Haverman et al., 2016; Terwee et al., 2014). Second, the item banks for Physical Function (Voshaar et al., 2014), Pain Interference (Crins et al., 2015), Pain Behavior (Crins et al., 2016) and Depression (Flens et al., 2017) have been psychometrically evaluated and meet the PROMIS standards (Reeve et al., 2007). Third, post hoc CAT simulations with the Depression item bank have shown highly efficient and precise measurement for clinical subjects, with a similar accuracy compared with the full item bank administration (Flens et al., 2017).

Following depression, anxiety is the most common disorder in Dutch mental health care (de Graaf, ten Have, van Gool, & van Dorsselaer, 2012), and a worldwide problem in general (Baxter, Scott, Vos, & Whiteford, 2013). Validating the Anxiety item bank as input for a CAT administration is therefore an obvious next step before the PROMIS methodology can be implemented successfully in (Dutch) mental health care. New measurements that are more accurate, precise, responsive, and efficient are always desirable, but considering the nationwide implementation of routinely collected PROM data in the Netherlands, there is an urgent need for state of the art efficient assessment with high-quality instruments (Carlier et al., 2012a; de Beurs et al., 2011).

The present article has two goals. The first goal is to present a psychometric evaluation of the Dutch-Flemish version of the PROMIS adult v1.0 item bank for Anxiety (Pilkonis et al., 2011). The evaluation is conducted on a large sample with both clinical subjects and persons from the general population, because we aimed to develop an instrument that measures the full latent anxiety continuum (i.e., all persons with no symptoms of anxiety to patients with severe anxiety). Furthermore, the evaluation is based on the PROMIS standards to ensure high quality items (Reeve et al., 2007), which is prerequisite for applying a CAT administration (Smits, Zitman, Cuijpers, den Hollander-Gijsman, & Carlier, 2012). Our second goal is to investigate how efficient and precise a CAT version of the Anxiety item bank may be to clinical and general population subjects, and how accurate this CAT version may be compared with a full item bank administration. For this goal, we performed a post hoc CAT simulation with a stopping rule set to a combination of high measurement precision and a fixed number of administered items. The stopping rule was chosen with a primary focus on the measurement precision of average and higher anxiety levels, as these are deemed the most relevant to measure, but without compromising the measurement precision of lower anxiety levels to a considerable extent. Efficiency and measurement precision were investigated both overall and as a function of the anxiety level; accuracy was investigated by comparing both test outcomes and group membership assignment between the CAT simulation and the full item bank administration.

4.3 Method

4.3.1 Participants

We collected data in a clinical and general population sample to cover the full range of possible latent anxiety levels in the Netherlands. For both samples, we aimed to include at least 1,000 respondents to obtain adequate item parameter estimates (Reise & Yu, 1990).

The eligible clinical sample consisted of 3,296 patients with common mental disorders who started their treatment in ambulatory mental health care. Patients were invited by the Dutch mental health care provider Parnassia Psychiatric Institute to digitally complete the item set. Parnassia Psychiatric Institute is by far the largest mental health institute in the Netherlands, and has a broad coverage across departments over the entire country. In accordance with the mental health care center’s policy, the item set was only administered when written informed consent was obtained. The patient’s diagnosis (*Diagnostic and Statistical Manual of Mental Disorders,* 4th ed. [*DSM-IV*]; American Psychiatric Association, 1994) was assessed prior to the study in two steps. First, a psychiatric nurse administered the Dutch translation of the Mini International Neuropsychiatric Interview (MINI-plus; Sheehan et al., 1998) by phone to ascertain the diagnosis. Second, the diagnosis was verified in a clinical face-to-face assessment.

The eligible general population sample consisted of 1,486 respondents that were approached digitally by a data collection panel to complete the item set (Desan Research Solutions; www.desan.nl). Respondents participated voluntary in the panel and received a small financial compensation for the study. To ensure representativeness of the sample, stratified sampling was applied. We used the following five stratification variables to mirror the Dutch population in 2013 (Statistics Netherlands; www.cbs.nl): gender (male, 49%; female, 51%), age (18-39 years, 34%; 40-64 years, 44%; 65+ years, 22%), education (low, 32%; middle 40%; high 28%), ethnicity (Dutch natives, 80%; western immigrants, 10%; nonwestern immigrants, 10%), and region (north, 10%; east, 21%; south, 22%; west, 47%). In each subgroup, deviations were allowed up to 2.5% because stratified sampling becomes increasingly difficult with an increasing number of variables. In addition, we assessed the diagnostic status of respondents by asking whether they were currently under treatment for mental health issues.

4.3.2 Measures

The item set consisted of 29 items from the Dutch-Flemish PROMIS adult v1.0 item bank for Anxiety (Terwee et al., 2014). The content of the items reflected a wide range of anxiety symptoms, problems, or negative affective states, and were stated positively (see Table 4.1; e.g., *I felt fearful*). Respondents were asked to indicate on a Likert scale how frequently they experienced the symptoms, problems or negative states in the past 7 days (1 = *never*, 2 = *rarely*, 3 = *sometimes*, 4 = *often*, and 5 = *always*), a higher score meaning more severe anxiety.

4.3.3 Psychometric evaluation

The psychometric evaluation of the Anxiety item bank was performed on the combined clinical and general population sample. We followed the PROMIS guidelines proposed by Reeve et al. (2007) to investigate whether we should remove any items from the item bank due to poor psychometric qualities. The evaluation focused on descriptive statistics, the main assumptions of item response theory (IRT), differential item functioning (DIF), and the item bank calibration. Below, we provide the details on these evaluation aspects. For more information, see Reeve et al. (2007). All statistical analyses were performed in the statistical environment R (R Core Team, 2015).

First, we evaluated the *descriptive statistics* of the full item bank sum scores (i.e., range, mean, standard deviation [SD], skewness, kurtosis, and internal consistency reliability [coefficient α]) and the individual item scores (i.e., response frequencies, range, mean, SD, skewness and kurtosis, interitem correlations, item-scale correlations, and drop in coefficient α for each item removed from the item bank). Specifically, undesirable patterns in the data were assessed (e.g., small range of item scores, outliers in item means, or negative correlations between items).

Second, we evaluated the IRT main assumptions of unidimensionality, local independence (LI), and monotonicity. *Unidimensionality* was evaluated with confirmatory factor analyses (CFA) using the R package lavaan (Version 0.5-18; Rosseel, 2012), and exploratory factor analyses (EFA) using the R package psych (Version 1.5.4; Revelle, 2013), both conducted on the polychoric correlation matrix (Bollen, 1989). For CFA, we used the following (scaled) fit statistics to assess good fit of the one-dimensional model: comparative fit index (CFI) > 0.95, Tucker–Lewis index (TLI) > 0.95, root-mean-square error of approximation (RMSEA) < 0.08, standardized root-mean-square residual (SRMR) < 0.08 (Reeve et al., 2007). For EFA to indicate sufficient unidimensionality, the first extracted factor should explain above 20% of the variance (Reckase, 1979, as cited in Hambleton, 1988). Furthermore, the ratio of variance explained by the first to second factor should at least be 4 (Reeve et al., 2007).

The assumption of *LI* was evaluated with the residual correlation matrix from the single-factor CFA, and with Yen’s Q3 statistic (Yen, 1993) using the R package mirt (Version 1.10; Chalmers, 2012). With the residual correlation matrix, we marked an item pair as possibly locally dependent when the corresponding coefficient was higher than 0.20 (Reeve et al., 2007). With Yen’s Q3 statistic, the residual item scores are calculated under Samejima’s graded response model (GRM; Samejima, 1969), and are then correlated among items. We assessed lack of model fit with Cohen's (1988) rules of thumb to interpret correlation effect sizes (Smits, Cuijpers, & van Straten, 2011): Q3 values between 0.24 and 0.36 imply moderate deviations of model fit, Q3 values above 0.37 imply large deviations. Item pairs with large deviations were marked as possibly locally dependent. When an item pair was marked by either its residual correlation coefficient or Yen’s Q3 statistic, further investigation was done by evaluating the impact of each item on the item parameter estimates (Reeve et al., 2007). To study this impact, we compared the item parameter estimates of the original GRM with a restricted GRM (i.e., minus one item).

The assumption of *monotonicity* was evaluated by examining graphs of item mean scores as a function of rest scores (total raw score minus the item score) using the R package mokken (Version 2.7.7; van der Ark, 2007). In addition, we evaluated the accompanying scalability coefficients (Mokken’s *H*) for the full scale and the individual items. Mokken’s *H* was interpreted as follows: low quality when $.30 \leq H < .40$, moderate quality when $.40 \leq H < .50$, and high quality when $H \geq .50$ (Mokken, 1971).

Third, we evaluated uniform and nonuniform *DIF* (Embretson & Reise, 2000) for gender, age (recoded into a binary variable by means of a median split), and education level (low, medium, high). Both types of DIF were assessed with ordinal logistic regression methods (OLR; Crane, Gibbons, Jolley, & van Belle, 2006) using the R package lordif (Version 0.2-2; Choi, Gibbons, & Crane, 2011). As measure of effect size, we used the change in McFadden’s pseudo R^2, following the suggestion of .02 as critical value for rejecting the hypothesis of no DIF (Choi et al., 2011).

Last, we estimated the item parameters of the Anxiety item bank (*calibration*) under the normal GRM (Samejima, 1969), an IRT model for polytomous items (Reeve et al., 2007). The GRM was fitted with multiple group estimation (McDonald, 1999; Smits, 2016) using the R package mirt (Version 1.10; Chalmers, 2012). We specified population (clinical and general) as grouping factor, and fixed the item parameters to be equal across groups. The latent trait (θ) was standardized to a scale with a mean of 0 and a standard deviation of 1 for the general population, a higher θ meaning more severe anxiety. The mean and standard deviation of the clinical sample were estimated under the model. As estimation algorithm, we used expectation–maximization. This algorithm is effective with one to three factors (Chalmers, 2012).

We evaluated the fit of the GRM by examining the item parameters and item fit. The GRM uses two types of parameters: the discrimination parameter *a* expresses the extent to which persons with similar θ estimates can be differentiated by the item; the four threshold parameters b_1 *to* b_4 (the number of threshold parameters for an item is equal to the number of response categories minus one) express the values of θ on which a person is expected to choose a higher over a lower item response. In addition, item fit was examined with the $S\text{-}X^2$ statistic

(Orlando & Thissen, 2000, 2003). This statistic compares the observed and expected response frequencies under the used IRT model, and quantifies differences between these frequencies. Items with a $S\text{-}X^2$ $p < .001$ are considered to have a poor fit in the IRT model (Reeve et al., 2007). To study the impact of poor fit, we evaluated the effect of each item on the item parameter estimates by comparing those of the original GRM with those of a restricted GRM (i.e., minus one item).

Finally, we evaluated how well the item bank could measure Anxiety for the full latent continuum. To accomplish this, we plotted the test information of the item bank for $-4 \leq \theta \leq 4$. It is calculated as the sum of all item information values at any relevant θ level.

4.3.4 CAT simulation

We used a post hoc CAT simulation to assess how efficient and precise a CAT version of the Anxiety item bank may be in clinical and general population subjects, and how accurate this CAT version may be compared with a full item bank administration. Previous studies have shown that post hoc CAT simulations are useful for this purpose as the results tend to be very similar to that of a real CAT administration (Kocalevent et al., 2009). Below, we provide the details on the CAT simulation settings and the assessment of efficiency, precision, and accuracy. The CAT simulation was performed using the R package mirtCAT (Version 0.5; Chalmers, 2015).

A CAT administration/simulation consists of four basic building blocks: a starting item, a method for estimating θ, an item selection procedure, and a stopping rule. The administration/simulation starts by presenting a first item. After a response is given, the software estimates θ and calculates the corresponding measurement precision (standard error [*SE*]). It then evaluates whether the obtained results meet the stopping rule. If not, a new item is selected and the procedure is repeated until the stopping rule is met, or all items have been presented.

As starting item, the CAT simulation used the item with the highest Fisher's information (Embretson & Reise, 2000; Wainer et al., 2000) at the average value of the latent trait in the general population (θ = 0). This item was *I felt tense*, which was coded as EDANX54 (Emotional Distress – ANXiety item bank, item 54) in the original US PROMIS item bank (www.assessmentcenter.net).

To estimate θ, we could choose from two methods: maximum likelihood (ML) and Bayesian estimation (Embretson & Reise, 2000). Bayesian estimation is often chosen because it uses an a priori population distribution of the latent variable. This property ensures that θ can be estimated for all response patterns. A drawback of Bayesian estimation, however, is that the estimation of θ is also influenced by the a priori distribution; it pulls θ estimates toward the center of the population distribution, which may result in bias (Flens et al., 2017; Smits, 2016). ML, by contrast, does not use an a priori distribution, and is therefore not able to estimate θ for response patterns that exclusively comprise extreme responses. It is, however, a more stable estimator considering possible bias. ML can also result in bias, but generally to a lesser extent compared with Bayesian estimation, especially using CAT (Wang & Vispoel, 1998). Bias in ML emerges when the respondent's latent trait level is different from the average threshold of the administered items. This means that, under the assumption that the item bank has an adequate number of items to cover the entire latent continuum, bias under CAT should be minimal, as it is specifically designed to select items according to the threshold level at the provisional θ estimate. Consequently, we have chosen to use ML as method for estimating θ. To deal with the issue of estimating θ for response patterns that exclusively comprise extreme responses, we could either set scale boundaries (Kim, Moses, & Yoo, 2015) or temporarily use a different estimation method (Chalmers, 2015). Due to a certain randomness in setting scale boundaries, we chose to temporarily use the commonly adopted Bayesian estimation method maximum a posteriori (Embretson & Reise, 2000). Thus, maximum a posteriori was used to estimate θ for response patterns that only include item scores 1 or 5, ML was used to estimate θ for all other response patterns.

To select additional items, we again used Fisher's information (Embretson & Reise, 2000; Wainer et al., 2000). Consequently, the item which had the highest information at the provisional θ estimate was selected.

As stopping rule, several methods have been proposed: a fixed number of administered items, a prespecified level of *SE*(θ), a prespecified change in θ estimate, or a prespecified change in *SE*(θ) (Babcock & Weiss, 2013; Choi, Grady, & Dodd, 2011; Smits et al., 2012). Each of these methods can be used individually or combined with each other. For this study, we chose to combine a prespecified level of *SE*(θ) with a fixed number of administered items. This combination rule is useful for measurements that are developed for both clinical and general population subjects. While clinical subjects mostly result in highly precise measurement with a low number of administered items (Flens, Smits, Carlier, van Hemert, & de Beurs, 2016), general population subjects often do not, not even when the full item bank is administered (Flens et al., 2017). Including a fixed number of administered items in the stopping rule should therefore result in efficient measurement for general population subjects as well, but without compromising the *SE* substantially.

The combination rule that we used to terminate the CAT simulation is a *SE*(θ) < 0.22 with a fixed number of 12 administered items. We chose a *SE*(θ) < 0.22 because it is comparable to a marginal reliability of .95 (Green, Bock, Humphreys, Linn, & Reckase, 1984), which results in a high standard for precise individual assessments (Bernstein & Nunnally, 1994). Regarding the fixed number of administered items, we aimed for a number that did not have a substantial impact on the precision of clinical subjects' CAT scores. We chose clinical subjects as this group is deemed the most relevant to measure anxiety (i.e., this group predominantly includes average to higher latent trait levels). To accomplish our aim, we used the criterion that at least 90% of the clinical subjects resulted in a *SE*(θ) < 0.22. We found this number to be 12 (92%). Using this fixed number of items, we investigated whether the *SE*(θ) of the general population subjects was not compromised to a considerable extent. This was assessed by comparing the *SE*(θ) of general population subjects that did not end up with a *SE*(θ) < 0.22 after 12 administered items, with their *SE*(θ) when no fixed number of items was applied in the stopping rule. By contrast, we also made this comparison for the number of selected items to assess the increase in administration efficiency by the fixed number of items.

As item parameters for the CAT simulation, it would be obvious to use the estimations of the complete sample. However, this would mean that we use the same data to calibrate the items and simulate the CAT, which would result in overfitting (i.e., results that are too optimistic; Hastie, Tibshirani, & Friedman, 2001). To deal with this issue, we split the clinical and general population sample randomly into half. The first half of the samples were combined to recalibrate the item bank (see “Psychometric evaluation” subsection); the second half of the samples were used as input for the CAT simulation. Thus, the item parameters of the complete sample (N = 2,010) could be used in a future CAT administration; the item parameters of half of the samples (n = 1,005) are used in the CAT simulation of this study. To study the similarity of the item parameters, we compared them using Pearson’s correlation coefficients, and differences in means and *SD*s (complete sample parameters minus CAT simulation parameters).

4.3.5 Precision and efficiency

A first demand for a CAT administration is that its outcome is both efficient (i.e., a low numbers of administered items) and precise (i.e., sufficiently free of random error). Efficiency was assessed by the mean number of selected items by the CAT simulation (and *SD*); precision was assessed by the mean $SE(\theta)$ and the percentage of respondents with a $SE(\theta) < 0.22$. In addition to these analyses, we plotted the number of selected items for each respondent as a function of the final θ estimate, along with the conditional *SE* of the Anxiety item bank. The conditional SE displays how precisely the item bank can measure anxiety at each level of the latent trait. It is calculated as the reciprocal square root of the sum of all item information values at each θ. All results are shown separately for the clinical and the general population sample.

4.3.6 Accuracy

A second demand for a CAT administration is that its outcome represents the construct which it purports to measure (i.e., free of systematic error). The θ estimates of the CAT simulation should therefore at the least be similar to those of the full item bank. We evaluated this demand by comparing the θ estimates of both tests with Pearson’s correlation coefficient and Cohen’s *d* effect size (difference between the average θ estimate divided by the pooled *SD*s). Cohen’s *d* was calculated using the R package effsize (version 0.6.2.; Torchiano, 2016), and was evaluated using the guideline proposed by Cohen (1988): 0.2 = small effect, 0.5 = medium effect, 0.8 = large effect, a higher value meaning more systematic error between the θ estimates of the CAT simulation and those of the full item bank administration. The results are shown separately for the clinical and the general population sample.

A third demand for a CAT administration is that its outcome discriminates group membership accurately (clinical vs. healthy). The group membership assignment of the CAT simulation should therefore at the least be similar to that of the full item bank. We evaluated this demand by comparing the diagnostic accuracy of both tests (McDonald, 1999). Specifically, it was assessed how well the CAT simulation and the full item bank administration could predict the diagnostic status of a person (i.e., anxiety disorder or no disorder). For this analysis, we needed clinical subjects with an anxiety disorder and healthy persons without a disorder. Persons with an anxiety disorder were selected from the clinical sample; healthy persons (i.e., persons without current treatment for mental health issues) were selected from the general population sample. Diagnostic accuracy was assessed with the area under the curve (AUC) of the receiver operating curve, an often-used indicator for diagnostic accuracy (Rice & Harris, 2005). AUC can be interpreted as the probability that a randomly selected person with an anxiety disorder has a higher θ estimate than a randomly selected person without mental health issues (Zweig & Campbell, 1993). We used the guideline proposed by Rice and Harris (2005) to evaluate the AUC values: .56 = small effect, .64 = medium effect, .71 = large effect, a higher value meaning a higher discriminative ability of the scale.

4.4 Results

4.4.1 Demographic characteristics

In the clinical sample, the response rate was 31% (*n* = 1,032). Of the 1,032 respondents, 24 were excluded for failing to complete all 29 items. The final clinical sample therefore consisted of *n* = 1,008 patients (62% female; average age = 40.2 years, *SD* = 12.9, range 19–76). Because the response rate of the eligible sample was only moderate, we performed a chi-square test of independence to examine whether the responders group differed from the nonresponders group. This analysis was performed for the variables gender, age, and diagnosis group (i.e., anxiety, depression, or another disorder, e.g., attention deficit disorder, somatoform disorder, personality disorder). We found no significant differences ($p < .05$) between responders and nonresponders for the variables gender and age. For the variable diagnosis group, we did find a significant difference (χ^2 [2, N = 3,296] = 11.39, $p < .05$), with somewhat less patients with a mood disorder in the responders group (44%) than in the nonresponders group (50%), somewhat more patients with an anxiety disorder in the responders group (33%) than in the nonresponders group (28%), and about an equal number of other disorders (responders group, 23%; nonresponders group, 22%). As measure of effect size, we investigated Pearson's residuals, following the suggestion of 2.00 as critical value for indicating a lack of model fit (Agresti & Kateri, 2011). It was found that only the responders group contained somewhat more patients with anxiety disorders than expected (*r* = 2.03).

In the general population sample, the response rate was 71% (*n* = 1,055). Of the 1,055 respondents, 53 respondents were excluded for showing suspicious response patterns (e.g., all responses in one category in combination with a very low response time). The final general population sample therefore consisted of *n* = 1,002 respondents (average age = 50.5 years, *SD* = 16.5, range 19–102). The demographics of the sample were as follows: gender (male, 49%; female, 51%), age (18-39 years, 34%; 40-64 years, 44%; 65+ years, 22%), education (low, 31%; middle 40%; high 29%), ethnicity (natives, 80%; western immigrants, 13%; nonwestern immigrants, 7%), and Dutch region (north, 12%; east, 20%; south, 21%; west, 47%). Each subgroup remained within the allowed deviation of 2.5% from the Dutch population statistics in 2013.

4.4.2 Psychometric properties of the Anxiety item bank

To begin with, the Anxiety item bank (*N* = 2,010) showed good descriptive statistics. Overall, the item bank showed a high internal consistency reliability (α = .98) that hardly changed when items were deleted from the item bank. Specifically, all items' scores showed a range between 1 and 5, and lacked outliers in response frequencies, mean and *SD* (see Table 4.1, column 3 for the item means and *SD*s). Only the item *I felt terrified* (EDANX33) had a minor deviation in skewness (1.16) and kurtosis (0.34). In addition, we did not find any negative or small correlation coefficients among the items. The lowest coefficient ($r = 0.41$) was found for item pair *I worried about other people's reactions to me* (EDANX37) and *I had twitching or trembling muscles* (EDANX44).

Next, the results from CFA and EFA indicated that the Anxiety item bank was sufficiently unidimensional. CFA showed a good fit of the unidimensional model for three out of four (scaled) fit indices: CFI = 0.97, TLI = 0.97, and SRMR = 0.04; the RMSEA indicated a moderate fit (RMSEA = 0.10). In addition, EFA showed that the first extracted factor explained 71% of the variance, which is far above the Reckase criterium of 20% (Reckase, 1979, as cited in Hambleton, 1988). Furthermore, the second extracted factor explained only 6% of the variance. The ratio of variance explained by the first to second factor was therefore almost 12, which is 3 times higher than the required minimum of 4 (Reeve et al., 2007).

Examining the results from the residual correlation matrix and Yen's Q3 statistics, the Anxiety item bank showed sufficient LI. The residual correlation coefficients were all below the lower bound of .20 (Reeve et al., 2007), which resulted in none of the items to be marked as possibly locally dependent. With Yen's Q3 statistic, we did find two item pairs that were marked. These item pairs were *I felt fearful* (EDANX01) and *I felt frightened* (EDANX02; Q3 = .48), *I felt fearful* (EDANX01) and *I felt anxious* (EDANX05; Q3 = .42). Fortunately, removing each of these items individually from the GRM only showed a minor impact on the item parameter estimates (max 0.11 for *a*, EDANX02 and EDANX05; max 0.04 for *b*, EDANX02).

Turning to the results from the Mokken analyses, the Anxiety item bank showed monotonicity to a high degree. First, the graphs of item mean scores as a function of rest scores showed monotonicity for all items as the underlying level of the scale was higher. Second, Mokken's *H* was .67 for the full Anxiety item bank, which indicates a strong scale. Third, all individual items had Mokken's *H* values above .50 (see Table 4.1, column 9), which is much higher than the lower bound of .30 (Mokken, 1971).

Subsequently, the results of the OLR analyses indicated that uniform and nonuniform DIF was not present among the items of the Anxiety item bank. We confirmed this for the variables gender, age, and education level.

Finally, Table 4.1 (column 4 to 8) displays the GRM item parameter estimates of the Anxiety item bank. The item parameters were parametrized in the scale of the latent trait distribution of the general population sample ($M = 0$, $SD = 1$). The mean and *SD* of the clinical sample was estimated to be 1.42 and 0.70, respectively.

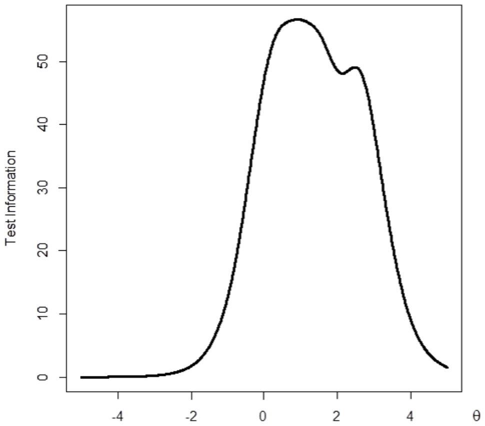
Table 4.1 IRT item characteristics for the Dutch-Flemish PROMIS Anxiety item bank based on a clinical sample and general population sample.

Item code	Item	*M* (*SD*)	a	b_1	b_2	b_3	b_4	H	$S\text{-}X^2$	p
EDANX01	I felt fearful	2.45 (1.19)	2.75	0.03	0.78	1.74	3.13	0.70	340.09	.00***
EDANX02	I felt frightened	2.08 (1.10)	2.63	0.45	1.33	2.19	3.35	0.68	254.57	.14
EDANX03	It scared me when I felt nervous	2.25 (1.21)	2.51	0.38	1.05	1.92	3.09	0.67	343.54	.00***
EDANX05	I felt anxious	2.52 (1.25)	2.99	0.08	0.77	1.56	2.84	0.71	296.81	.00
EDANX07	I felt like I needed help for my anxiety	2.45 (1.39)	3.03	0.39	0.93	1.51	2.42	0.70	339.77	.00
EDANX08	I was concerned about my mental health	2.53 (1.38)	2.35	0.22	0.84	1.49	2.52	0.66	402.13	.00***
EDANX12	I felt upset	2.53 (1.22)	2.75	0.00	0.72	1.65	2.85	0.70	254.08	.19
EDANX13	I had a racing or pounding heart	2.31 (1.21)	1.80	0.19	0.99	2.00	3.37	0.60	396.79	.00
EDANX16	I was anxious if my normal routine was disturbed	2.21 (1.23)	2.20	0.43	1.17	1.95	3.08	0.64	362.27	.01
EDANX18	I had sudden feelings of panic	2.14 (1.24)	3.11	0.61	1.22	1.89	2.84	0.70	272.85	.05
EDANX20	I was easily startled	2.16 (1.12)	1.67	0.18	1.31	2.34	3.77	0.59	401.48	.00***
EDANX21	I had trouble paying attention	2.63 (1.19)	1.89	-0.38	0.63	1.65	3.15	0.63	327.94	.08
EDANX24	I avoided public places or activities	2.42 (1.30)	1.76	0.16	0.95	1.76	3.03	0.59	378.10	.12
EDANX26	I felt fidgety	2.80 (1.29)	2.99	-0.21	0.44	1.29	2.47	0.72	277.05	.06
EDANX27	I felt something awful would happen	2.04 (1.19)	2.28	0.65	1.37	2.14	3.06	0.64	404.70	.00***
EDANX30	I felt worried	3.08 (1.18)	2.35	-0.91	0.02	1.13	2.42	0.70	277.60	.04
EDANX33	I felt terrified	1.80 (1.06)	2.68	0.96	1.66	2.40	3.30	0.68	360.91	.00***
EDANX37	I worried about other people's reactions to me	2.51 (1.31)	1.97	0.02	0.85	1.65	2.70	0.62	400.32	.01
EDANX40	I found it hard to focus on anything other than my anxiety	2.32 (1.28)	3.59	0.40	1.04	1.71	2.64	0.72	244.75	.08
EDANX41	My worries overwhelmed me	2.47 (1.31)	2.87	0.22	0.87	1.61	2.59	0.70	326.91	.00
EDANX44	I had twitching or trembling muscles	1.96 (1.10)	1.36	0.66	1.58	2.75	4.53	0.52	370.35	.03
EDANX46	I felt nervous	2.59 (1.22)	2.87	-0.12	0.66	1.56	2.84	0.71	232.26	.36
EDANX47	I felt indecisive	2.36 (1.26)	2.26	0.25	0.95	1.80	2.97	0.65	378.12	.00
EDANX48	Many situations made me worry	2.52 (1.25)	2.85	0.08	0.74	1.58	2.84	0.70	278.73	.05
EDANX49	I had difficulty sleeping	2.73 (1.36)	1.34	-0.40	0.54	1.49	2.73	0.54	458.13	.00
EDANX51	I had trouble relaxing	2.90 (1.30)	2.51	-0.44	0.38	1.23	2.38	0.71	294.21	.12
EDANX53	I felt uneasy	2.52 (1.26)	3.06	0.07	0.80	1.57	2.72	0.72	282.94	.02
EDANX54	I felt tense	2.87 (1.32)	3.28	-0.27	0.42	1.21	2.28	0.74	267.90	.02
EDANX55	I had difficulty calming down	2.31 (1.24)	3.23	0.33	1.04	1.80	2.79	0.71	268.89	.02

The item parameter estimates showed considerable variation. The discrimination parameters ranged from a = 1.34 (*I had difficulty sleeping*; EDANX49) to a = 3.59 (*I found it hard to focus on anything other than my anxiety*; EDANX40); the threshold parameters ranged from b_1 = -0.91 (*I felt worried*; EDANX30) to b_4 = 4.53 (*I had twitching or trembling muscles*; EDANX44). In addition, the p-values of the $S\text{-}X^2$ statistics ranged from 0.00 to 0.36 (see Table 4.1, column 10). From the 29 items, 6 items had a $p < .001$ (see Table 4.1, column 11). These items were *I felt fearful* (EDANX01), *It scared me when I felt nervous* (EDANX03), *I was concerned about my mental health* (EDANX08), *I was easily startled* (EDANX20), *I felt* *something awful would happen* (EDANX27), and *I felt terrified* (EDANX33). Removing each of these items individually from the GRM only showed a minor impact on the item parameter estimates (max 0.10 for *a*, EDANX01; max 0.05 for *b*, EDANX27). We therefore concluded that the GRM fitted the Anxiety item bank sufficiently. Moreover, based on all results of the psychometric evaluation, we have chosen not to remove any of the items from the Anxiety item bank.

In Figure 4.1, we displayed the test information of the Anxiety item bank. The item bank is highly informative for the average and higher anxiety levels (approximately θ > -0.5), and less informative for the lower anxiety levels (approximately θ < -0.5). These results indicate that although we constructed a scale to measure the full latent Anxiety continuum, the item bank measures Anxiety more precise for the average and higher anxiety levels than for the lower anxiety levels. This was to be expected as low values of the latent trait are generally related to less precise measurement in mental health constructs (Reise & Waller, 2009).

Figure 4.1 Test information of the Anxiety item bank.



4.4.3 Properties of the CAT simulation

4.4.3.1 Item parameter estimates

The comparison between the item parameter estimates of the complete sample (*N* = 2,010) and the CAT simulation sample (*n* = 1,005) resulted in high correlation coefficients (r_a = 1.00, r_{b1} = 1.00, r_{b2} = 1.00, r_{b3} = 1.00, r_{b4} = .99), small differences in means (M_a = -0.02, M_{b1} = 0.08, M_{b2} = 0.07, M_{b3} = 0.09, M_{b4} = 0.06), and small differences in SDs (SD_a = -0.02, SD_{b1} = 0.01, SD_{b2} = 0.01, SD_{b3} = 0.02, SD_{b4} = 0.03). We therefore concluded that the item parameter estimates of the CAT simulation sample are highly similar to those of the complete sample.

4.4.3.2 Efficiency and precision

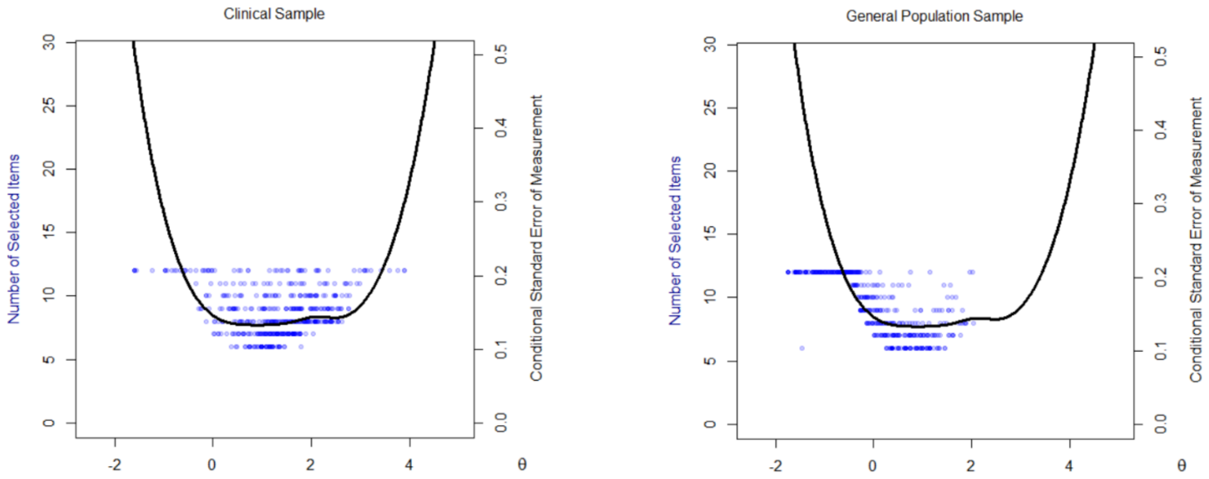
Efficient and highly precise measurement was obtained in both samples, with more gains for the clinical sample (*n* = 504; number of selected items, *M* = 8.64, *SD* = 1.83; mean *SE*(θ) = 0.22) than for the general population sample (*n* = 501; number of selected items, *M* = 9.48, *SD*

= 2.38; *SE*(θ) = 0.28). This was also shown by the percentage of respondents with a *SE*(θ) < 0.22, which was much higher in the clinical sample (92%) than in the general population sample (63%). Considering that the percentage of persons with low anxiety values is higher in the general population, these results were to be expected (Reise & Waller, 2009).

In Figure 4.2, the number of selected items are displayed as a function of the final θ estimate along with the conditional *SE* of the Anxiety item bank. The θ estimates of the general population sample are clearly located more to the left of the scale than those of the clinical sample. At this end of the scale, the conditional *SE* is high. Consequently, the general population sample contained less respondents with a *SE*(θ) < 0.22 than the clinical sample, and received more often all 12 items. By contrast, the conditional *SE* was at its lowest approximately between 0.00 < θ < 2.00. At these scale points, measurement was most efficient for the majority of respondents from both samples, with six items as the lowest number of administered items.

Figure 4.2

Number of selected items by the CAT simulation shown as a function of the final θ estimate along with the conditional SE of measurement of the Anxiety item bank for the clinical sample and the general population sample.



For the general population, we found that subjects whom did not end up with a *SE*(θ) < 0.22 after 12 administered items, had an average *SE*(θ) = 0.39. When the CAT simulation was performed again, but without applying a fixed number of items in the stopping rule, the average *SE*(θ) decreased somewhat to *SE*(θ) = 0.35. By contrast, the mean number of selected items increased from 12.00 to 26.72. These results indicate that applying a fixed number of 12 administered items in our stopping rule did not compromise respondents' *SE*(θ) substantially, but did increase the administration efficiency considerably.

4.4.3.3 Accuracy

Table 4.2 displays Pearson's correlation coefficients and sizes of difference (Cohen's *d*) between the θ estimates of the CAT simulation and those of the full item bank administration.

We found that the coefficients were high for both clinical and general population sample (*r* = 0.98). Furthermore, Cohen’s *d* showed a negligible effect size for both samples (*d* = 0.01). These results indicate that the θ estimates of a CAT administration may be highly similar to those of a full item bank administration.

The AUC analyses consisted of *n* = 204 patients with an anxiety disorder and *n* = 449 healthy persons. We found that the AUC value showed a large effect when the full item bank was administered (AUC = 0.92, 95% *CI* [0.89, 0.94]), which remained highly similar under the CAT simulation (AUC = 0.92, 95% *CI* [0.90, 0.95]). These results indicate that the diagnostic accuracy of a CAT administration may be highly similar to that of a full item bank administration.

Table 4.2 Pearson’s correlation coefficient and effect size of the difference (Cohen’s *d*) between the full Anxiety item bank θ estimates and the CAT simulation θ estimates for the clinical sample and the general population sample.

	Full θ		CAT θ			
Sample	*M*	*SD*	*M*	*SD*	*r*	*d*
Clinical	1.32	0.87	1.33	0.88	0.98	0.01
General Population	-0.11	0.96	-0.09	0.95	0.98	0.01

4.5 Discussion

The first goal of this study was to present a psychometric evaluation of the Dutch-Flemish version of the US PROMIS adult v1.0 item bank for Anxiety (Pilkonis et al., 2011). We used a large sample (*N* = 2,010) with clinical and general population subjects to demonstrate that the Anxiety item bank has desirable psychometric properties according to the PROMIS standards (Reeve et al., 2007). These properties include sufficient unidimensionality, LI, monotonicity, absence of DIF, and GRM fit. We therefore conclude that the Anxiety item bank could be used as input for a CAT administration to measure the full latent anxiety continuum. As expected, the item bank measures Anxiety more precisely for persons with average and higher anxiety levels than for persons with low-anxiety levels (Reise & Waller, 2009).

The second goal of this study was to investigate how efficient and precise a CAT version of the Anxiety item bank may be to clinical and general population subjects, and how accurate this CAT version may be compared with a full item bank administration. For this goal, we performed a post hoc CAT simulation with a stopping rule that combined a high measurement precision with a fixed number of administered items, and that was chosen with a primary focus on the measurement precision of average and higher anxiety levels. First, the simulation showed that our CAT version resulted in efficient and highly precise measurement, with more gains for the clinical sample as compared with the general population sample. For clinical practice, this may imply that measurement precision and efficiency declines somewhat as the severity of anxiety declines. This is to be expected and acceptable as the Anxiety item bank is primarily developed to measure clinical subjects. Second, the simulation showed that our CAT version was similarly accurate compared with the full item bank administration. This was shown by both θ estimates and the assignment of group membership. We therefore conclude that a CAT administration with the Anxiety item bank may not only be efficient and highly precise, but also just as accurate as a full item bank administration.

In this study, we showed that the item parameter estimates of the CAT simulation sample were highly similar to those of the complete sample. This means we can assume that similar results will be obtained when our CAT version is administered with the item parameters of the complete sample. To verify this assumption, replication of the present results is necessary with a genuine CAT administration. Moreover, we need to address other accuracy aspects to validate our CAT version. These aspects include concurrent validity to ensure that our CAT version is similar to other validated anxiety instruments (McDonald, 1999), as well as longitudinal validity aspects to ensure that our CAT version could be used to assess change in respondents. Longitudinal validity aspects include measurement invariance over time (Fokkema, Smits, Kelderman, & Cuijpers, 2013; Fried et al., 2016) and responsiveness to change (de Beurs et al., 2012; Schalet et al., 2016). A measurement invariant scale means that the item bank measures the same construct at different time points; responsiveness to change means that change in θ estimates over time represent real changes in the construct (Mokkink et al., 2010).

Specific consideration should be given to the comparison of the Dutch-Flemish version of the Anxiety item bank and the original US version. PROMIS aims to implement identical item banks and item parameters in every country to increase uniformity and enhance international comparability. This might prove difficult as the meaning of items may vary in different languages, and cultural differences may emerge across the globe regarding the valence of constructs, such as anxiety (van Widenfelt, Treffers, de Beurs, Siebelink, & Koudijs, 2005). Future research should therefore address measurement invariance between countries to assess to what extent comparisons are valid, and whether similar norms can be applied to instruments (e.g., Paz, Spritzer, Morales, & Hays, 2013; Wahl et al., 2015). Furthermore, countries should come to an international agreement about the CAT software, the CAT specifics, and the continued development of the item banks and the CAT methodology.

While awaiting these developments, our CAT version of the Dutch-Flemish PROMIS adult v1.0 item bank for Anxiety can be used in single measures. To increase the efficiency gains in these measures, the required measurement precision may be decreased, for example to a $SE(\theta) < 0.32$, which is generally required as minimal precision for individual assessments (Bernstein & Nunnally, 1994). Using this alternative stopping rule, further simulations (the results of which are not shown herein) showed that the mean number of selected items may be decreased even further from 8.64 to 4.25 items for the clinical sample, and from 9.48 to 6.06 items for the general population sample. When the goal, however, is to assess change over time, we recommend using higher levels of measurement precision. High $SE(\theta)$ values are needed to detect true change in respondents (Brouwer, Meijer, & Zevalkink, 2013). With more precise indicators for true change, treatment providers have more useful information to assess whether to continue, change, or conclude treatment of patients. For this reason, we also recommend future researchers who are interested in change assessment to consider alternative stopping rules for the CAT administration, such as the predicted standard error reduction (Choi et al., 2011). With this stopping rule, new items will be administered for as long as the measurement precision increases to a prespecified degree. For our CAT version, this means that the measurement precision could increase for a substantial number of respondents (see Figure 4.2). This stopping rule is not yet available in the Dutch CAT software, but when it does, Anxiety could be measured even more precisely.

A limitation of this study is the representativeness of the samples used. For the clinical sample, we collected data from the largest mental health institute in the Netherlands which has a broad coverage across departments over the entire country. The response rate, however, was only moderate (i.e., 31%). Furthermore, clinical subjects with an anxiety disorder were slightly overrepresented in the responders group. The difference between the responders and nonresponders group, however, was only small (i.e., approximately 5 percentage points). The effects of this selection bias will therefore likely be small. To deal with this issue in future item bank development based on clinical subjects, we recommend incorporating clinical criteria in a stratified sampling process.

In addition, the representativeness of the samples used to assess diagnostic accuracy could be somewhat improved. First, the sample size for clinical subjects with an anxiety disorder was moderately small (*n* = 204). Second, we did not have any information concerning comorbidity rates in the clinical sample. For future studies, we therefore recommend increasing the sample size for clinical subjects, and using both primary and secondary diagnostic criteria to assign group membership. In addition, the sample for healthy persons contained respondents with moderate- to high-anxiety trait levels, and may have included persons in need of treatment for their anxiety, but who either choose not to reveal being in treatment, or did not seek treatment. Ideally, to ensure a pure healthy sample, the diagnosis-free status of these respondents would be assessed with a diagnostic screener or interview, but the burden may be too high for the possible gains in classification accuracy. We therefore recommend maintaining our adopted approach in which respondents are asked whether they are currently under treatment for mental health issues. Finally, potential inclusion of anxiety disorder patients in the healthy sample likely does not bias the present results in a positive direction, but rather yields a too conservative estimate of the diagnostic accuracy of CAT.

In this study, the Dutch-Flemish version of the PROMIS adult v1.0 item bank for Anxiety was investigated. We found favorable psychometric properties, evidence of efficient and highly precise measurement applying a CAT simulation, and a similar accuracy between this CAT simulation and the full item bank administration. Similar results have been reported for the original US version of the Anxiety item bank (Pilkonis et al., 2011; Schalet et al., 2016), the Dutch PROMIS adult v1.0 item bank for Depression (Flens et al., 2017), and other translations of the Depression item bank (e.g., Spanish, Vilagut et al., 2015; German, Jakob et al., 2015). Considering these results, the PROMIS methodology seems to fulfill its promise to measure - with an internationally applicable assessment battery - patient-reported health of adults and children more efficiently, precisely, and accurately than existing PROMs do. We therefore recommend colleagues from other countries to translate and evaluate the PROMIS item banks as input for a CAT administration.

Chapter 5

Practical Significance of Longitudinal Measurement Invariance Violations in the Dutch-Flemish PROMIS Item Banks for Depression and Anxiety: An Illustration with Ordered-Categorical Data

Published as:

Flens, G., Smits, N., Terwee, C. B., Pijck, L., Spinhoven, P., & de Beurs, E. (2021). Practical Significance of Longitudinal Measurement Invariance Violations in the Dutch–Flemish PROMIS Item Banks for Depression and Anxiety: An Illustration With Ordered-Categorical Data. *Assessment*, *28*(1), 277-294.

5.1 Abstract

We investigated longitudinal measurement invariance in the Dutch–Flemish PROMIS adult v1.0 item banks for Depression and Anxiety using two clinical samples with mood and anxiety disorders (*n* = 640 and *n* = 528, respectively). Factor analysis was used to evaluate whether the item banks were sufficiently unidimensional at two test-occasions and whether the measured constructs remained the same over time. The results indicated that the item banks were sufficiently unidimensional, but the thresholds and residual variances of the constructs changed over time. However, using tentative rules of thumb, these invariance violations did not substantially affect the endorsement of a specific response category of a specific item at a specific test-occasion. Furthermore, the impact on the mean latent change scores of the item banks remained below the proposed cutoff value for substantial bias. These findings suggest that the invariance violations lacked practical significance for test users, meaning that the item banks provide sufficiently invariant latent factor scores for use in clinical practice.

Keywords: depression, anxiety, clinical assessment, longitudinal measurement invariance, PROMIS

5.2 Background

In the Netherlands, Dutch-Flemish versions of the Patient-Reported Outcomes Measurement Information System (PROMIS) adult v1.0 item banks for Depression and Anxiety have been developed. In previous studies, the original United States (US) PROMIS adult v1.0 item banks for Depression and Anxiety were translated from English into Dutch-Flemish (Terwee et al., 2014), and psychometrically evaluated for cross-sectional use in both the Dutch general population and ambulatory clinical populations at the start of treatment (Flens et al., 2017, 2019). These studies showed that both item banks have psychometric properties that complied with the PROMIS standards (Reeve et al., 2007). Consequently, adequate item parameters are available that may be used as input for computerized adaptive testing (CAT). CAT is a computer-based method in which items are selected from an item bank based on a respondent's previous item responses. The administration of items stops when a prespecified criterion is met (e.g., a high measurement precision). Consequently, CAT can reduce administration burden with a shorter test while maintaining a high-measurement precision. For more details on CAT, see for example, Embretson and Reise (2000).

Using the Dutch-Flemish PROMIS item banks in CAT simulations, efficient and highly precise measurement of depression and anxiety was obtained (Flens et al., 2017, 2019). Furthermore, the accuracy of the CAT simulations was highly similar compared with that of the full item bank administrations, both in final score estimations and in distinguishing clinical subjects from persons without a mental health disorder. Based on these results, it was concluded that the Dutch-Flemish PROMIS item banks administered by CAT may measure depression and anxiety accurately, precisely, and efficiently in both the general population and clinical ambulatory populations at the start of treatment. When the final goal, however, is to use these CATs in *repeated assessments* of clinical subjects, research also needs to address their longitudinal measurement properties. One of these aspects includes longitudinal measurement invariance (LMI; Widaman, Ferrer, & Conger, 2010).

An item bank is said to be longitudinally measurement invariant when it measures one or more single constructs in the same way over time. This means that changes in test scores of respondents over time can entirely be attributed to changes in the construct(s) measured by the item bank (Fried et al., 2016; Liu et al., 2017). If this is not the case, for example due to the psychoeducation of clinical subjects (Fokkema, Smits, Kelderman, & Cuijpers, 2013; for more explanations, see Fried et al., 2016), then observed changes in test scores are likely to be biased, possibly resulting in wrong inferences about the (change in) construct level. To our knowledge, this kind of bias is investigated in numerous mental health instruments (e.g., Fokkema et al., 2013; Fried et al., 2016; Jabrayilov, Emons, de Jong, & Sijtsma, 2017; te Poel, Hartmann, Baumgartner, & Tanis; 2017), but not yet in any of the PROMIS item banks. The evaluation of LMI in these item banks is highly relevant because in most of the performed LMI studies, it was concluded that the assumption of invariance did not or only partially hold.

In the present study, LMI was investigated for the Dutch-Flemish PROMIS adult v1.0 item banks for Depression and Anxiety using two clinical samples with mood and anxiety disorders respectively. We evaluated whether (a) the item banks were sufficiently unidimensional at two test-occasions, and (b) the measured constructs remained the same over time. Specifically, LMI was investigated within the framework of factor analysis, using both confirmatory factor analysis (CFA) and exploratory factor analysis (EFA). We modeled the items of the PROMIS item banks explicitly as ordered-categorical. In previous measurement invariance studies, ordered-categorical items were often modeled as continuous because the evaluation of invariance through factor analysis comes with several challenges for ordered-categorial data (Liu et al., 2017; Wu & Estabrook, 2016). Recently, new methodology has become available for CFA which overcomes most of these challenges (Liu et al., 2017). As a result, LMI can be investigated more accurately than would have been the case when the data were modeled as continuous (Rhemtulla, Brosseau-Liard, & Savalei, 2012). In addition, as full LMI rarely holds (van de Schoot et al., 2015), we did not focus solely on statistical significance in the analyses. Additionally, effect sizes based on new methodologies for CFA were evaluated to study the practical significance of the expected invariance violations. Specifically, we investigated two effect sizes that are relevant for test users. This means that we evaluated when (i.e., which test-occasion) and where (i.e., which item and response category) a LMI violation has a substantial impact (Liu et al., 2017), and to what degree changes in test scores are affected (Liu & West, 2018).

5.3 Methods

5.3.1 Participants

Data for this study were collected in two clinical populations that consisted of patients who started ambulant treatment for either a mood disorder or an anxiety disorder. Patients were invited to participate by the Dutch mental health care provider Parnassia Psychiatric Institute, which is the largest mental health institute in the Netherlands and has departments across the country (Flens et al., 2019). Prior to the study, mental health clinicians of Parnassia Psychiatric Institute determined the patient's diagnosis (*DSM-IV*; *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed.; American Psychiatric Association, 1994) with the Dutch translation of the Mini International Neuropsychiatric Interview (i.e., MINI-plus; a structured diagnostic interview used to systematically assess *DSM-IV* diagnoses) in a clinical face-to-face assessment during the intake of treatment. The MINI(-plus) showed sufficient sensitivity, specificity, negative and positive predictive values, and sufficient interrater agreement with other diagnostic instruments; only the interrater agreement on a generalized anxiety disorder and a simple phobia was insufficient (Lecrubier et al., 1997; Muramatsu et al., 2007; Sheehan et al., 1998; van Vliet & de Beurs, 2007). In addition, in accordance with Parnassia Psychiatric Institute's policy, informed consent was obtained before the measurements were administered.

We aimed to include at least 500 patients per sample to be able to adequately examine factor structures (Comrey & Lee, 1992; Liu et al., 2017; MacCallum, Widaman, Zhang, & Hong, 1999). A patient was included when (a) a pretest and posttest were completed without missing item responses, (b) the posttest was administered at least one month after the pretest, and (c) the posttest was administered after the first treatment session. We only included patients that completed a pretest and posttest without missing item responses because our software package (see section Software) could not yet handle missing data using CFA with ordered-categorical data. For more details on handling missing data in assessing LMI with ordered-categorical data, see Liu et al. (2017). Additionally, the manual of the used software package could be evaluated for any new features (e.g., https://cran.r-project.org/web/packages/lavaan/lavaan.pdf).

5.3.2 Measurements

The measurements consisted of the full Dutch-Flemish PROMIS adult v1.0 item banks for Depression (Flens et al., 2017) and Anxiety (Flens et al., 2019). The Depression item bank was administered to patients who were treated for a mood disorder; the Anxiety item bank was administered to patients whom were treated for an anxiety disorder.

Patients were asked to indicate on a Likert-type scale (1 = *never*, 2 = *rarely*, 3 = *sometimes*, 4 = *often*, and 5 = *always*) how frequently they experienced a wide range of either depression or anxiety symptoms in the past 7 days. The items reflected symptoms, problems, or negative affective states (e.g., *I felt worthless* for the Depression item bank, or *I felt fearful* for the Anxiety item bank), a higher score meaning more severe depression or anxiety. In Table 5.1 (Depression item bank; 28 items) and Table 5.2 (Anxiety item bank; 29 items), the items with the original PROMIS coding are presented. We should note that the item banks cannot be used without permission of PROMIS (see also www.healthmeasures.net).

For each patient, an item bank was administered digitally through an automated process twice. According to Parnassia Psychiatric Institute’s policy, the invitation for the pretest was sent before the intake session. To ensure that at least some treatment was administered and some change in clinical severity would be achieved, the invitation of the posttest was sent at least 1 month after the pretest.

In addition to the administration of the PROMIS item banks, the pretest was preceded by several questionnaires depending on the patient’s age and disorder. These questionnaires were not relevant for the purpose of this study and therefore not further described. The posttest administration was not preceded or followed by additional questionnaires.

5.3.3 Statistical analyses

5.3.3.1 Descriptive statistics

The degree of change within patients was evaluated by comparing the mean raw item scores between the pretest and posttest. Uniformity in the pretest to posttest interval was evaluated by calculating quantiles of the days between the pretest and posttest.

5.3.3.2 Unidimensionality

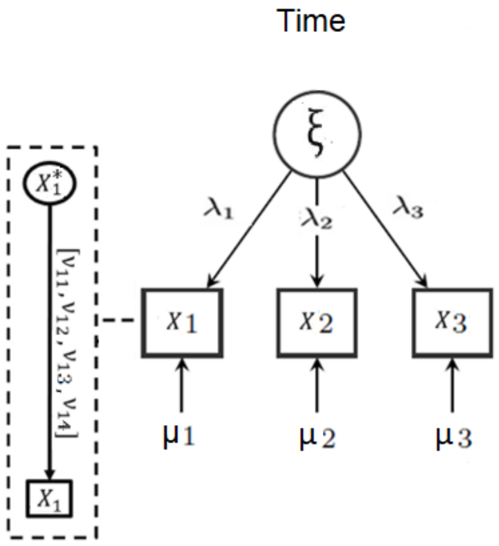
To evaluate LMI in instruments that are theorized to (strongly) reflect one underlying construct, the evaluation of the unidimensionality assumption is a strict condition (Fried et al., 2016). If this assumption is violated, item parameter estimates of CFA will almost inherently be biased, possibly resulting in biased test scores.

Unidimensionality was assessed with EFA as well as CFA (Reeve et al., 2007). With EFA, two factors were extracted from the pretest and posttest data separately. A measurement was considered to be sufficiently unidimensional when the first factor explained more than 20%

of the variance (Reckase, 1979, as cited in Hambleton, 1988), and the ratio of variance explained by the first to second factor was at least 4 (Reeve et al., 2007).

With CFA, a one-factor model was fitted to the pretest and posttest data separately. To illustrate the one-factor CFA model as a first step towards the longitudinal CFA models, it is presented in Figure 5.1, for three example items with five response categories. The model estimates four types of parameters for the ordered-categorial data: (a) *the common factor mean* (ξ) represents the mean of all respondent's latent factor scores; (b) *the factor loadings* (λ) represent for each item the strength and direction of association between the observed item responses and the latent factor scores; (c) *the thresholds* (υ) are cutoff values (the number of thresholds for each item equals the number of response categories minus one) that divide the underlying continuous latent responses into sections, each of which corresponds to endorsing an observed ordinal response category; and (d) *the residual variances* (μ) represent the degree of error with which each item measures the construct of interest. With the resulting model, the degree of unidimensionality was evaluated using the following (scaled [i.e., corrected for nonnormality]) fit-statistics (Fokkema et al., 2013): a scaled CFI ≥ .90 indicates an adequate fit, a scaled CFI ≥ .95 a good fit (Bentler, 1990); a standardized root-mean-square residual (SRMR) ≤ .08 indicates an adequate fit, a SRMR ≤ .05 a good fit (Hu & Bentler, 1999); a scaled root-mean-square error of approximation (RMSEA) ≤ .08 indicates an adequate fit, a scaled RMSEA ≤ .05 a good fit (Browne & Cudeck, 1993).

Figure 5.1 One-factor CFA model for ordered-categorical data with three items and five response categories.

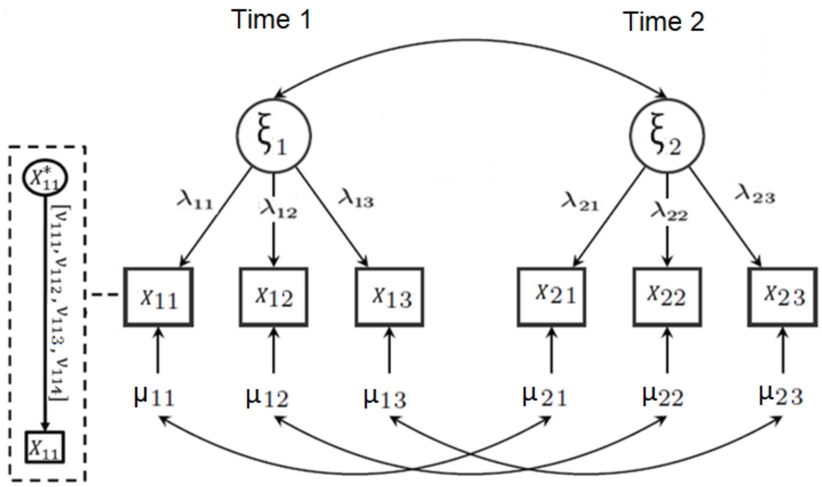


Note. ξ = common factor mean; λ = factor loadings; X^* = continuous latent item responses; X = observed item responses; υ = thresholds; μ = residual variances. For each parameter, the first subscript represents the item, the second subscript the threshold number.

5.3.3.3 Tenability of equality constraints

To investigate whether the measured constructs remain the same over time, a series of nested longitudinal CFA models was evaluated and compared (Liu et al., 2017). To illustrate the modeling sequence for evaluating LMI with ordered-categorical data, the general longitudinal model is presented in Figure 5.2, again using three example items and five response categories.

Figure 5.2 Longitudinal CFA model for ordered-categorical data with three items and five response categories.



Note. ξ = common factor mean; λ = factor loadings; X^* = continuous latent item responses; X = observed item responses; υ = thresholds; μ = residual variances. For each parameter, the first subscript represents the test-occasion, the second subscript the item number, and the third subscript the threshold number. The longitudinal structure of the model is captured by including a factor correlation between test-occasions as well as a residual correlation between test-occasions for each item.

First, the *baseline invariance model* was fitted. This is a two-factor model in which the pretest and posttest were treated as separate factors. To account for the longitudinal design, a factor correlation was included between test-occasions as well as a residual correlation between test-occasions for each item (Oort, 2005; Vandenberg & Lance, 2000). With the resulting model, it was assessed whether the construct of interest is measured by the same items (i.e., the same content) over time. Second, the baseline invariance model was extended with equality constraints on the factor loadings between test-occasions for each item to create *the loading invariance model*. With this model, it was assessed whether the observed item scores have a similar correlation with the latent factor scores over time. Third, the loading invariance model was extended with equality constraints on the thresholds between test-occasions for each item to create the *threshold invariance model*. With this model, it was assessed whether respondents with similar latent factor scores over time would choose the same response categories. Finally, the threshold invariance model was extended with equality constraints on the residual variances between test-occasions for each item to create the *unique factor invariance model*. With this model, it was assessed whether the items measure the construct of interest with a similar amount of error over time. Only if this is the case, then an item bank is said to be sufficiently invariant. In other words: equality constraints on factor loadings, thresholds, and residual variances need to be tenable in the longitudinal model to attribute changes in the observed item responses over time entirely to changes in the latent factor over time. A mathematical explanation that supports this can be found in Liu et al. (2017).

To investigate the tenability of the equality constraints, we first evaluated the fit of the longitudinal CFA models using the same fit statistics and cutoff values as for the one-factor CFA models. Second, we compared the fit between two subsequent models with the chi-square (i.e., χ^2) scaled difference test (Satorra, 2000), using an alpha level of .05 to indicate deterioration of fit. Third, because a χ^2 difference test is known to exhibit inflated Type 1 error rates (Sass, Schmitt, & Marsch, 2014), we also evaluated the modification indices of the imposed equality constraints (Liu et al., 2017). When a model showed a modification index above 5, this was considered a deterioration of fit (Jöreskog & Sörbom, 1996). Finally, it has been suggested to also compare the fit between two subsequent models by calculating differences in CFI's or RMSEA's (Chen, Curran, Bollen, Kirby, & Paxton, 2008; Cheung & Rensvold, 2002; Hu & Bentler, 1998). These difference tests, however, have not been properly studied for models with ordered-categorical data (Liu et al., 2017). We therefore chose not to use these fit statistics in our study.

5.3.3.4 CFA model-identification

To be able to estimate the parameters of a CFA model (i.e., to identify the model), some parameters need to be constrained (i.e., the model-identification parameters). For the one-factor CFA models, the following constraints needed to be imposed at each test-occasion (Wu & Estabrook, 2016): (a) the common factor mean was fixed to 0; (b) the factor loading of one single item was fixed to 1; (c) all intercepts were fixed to 0 (intercepts represent the expected item response when the latent factor score is equal to zero, and are only allowed to be estimated when the data is continuous); and (d) all residual variances were fixed to 1. In addition, we needed to impose the following constraints to identify the longitudinal CFA models (Liu et al., 2017): (a) the common factor mean of the pretest was fixed to 0; (b) the factor loading of a single item (i.e., the marker item) was fixed to 1 for both measurements; (c) one threshold of each item and a second threshold for the marker item was constrained to be equal between pretest and posttest; (d) all intercepts were fixed to 0; and (e) all residual variances of the pretest items were fixed to 1.

The constraints on the common factor mean, the intercepts and the residual variances could be imposed directly because *all* parameters were affected within a test-occasion. In the cases of factor loadings and thresholds, however, we needed to impose constraints on *specific* parameters. For these parameters, it is strictly necessary that they are at least longitudinally invariant. Otherwise, baseline invariance will be violated, which will make further model-comparisons biased. In the case of noninvariant threshold model-identification parameters, for example, a true violation of threshold invariance may mistakenly result in the conclusion that loading invariance is violated (Liu et al., 2017). To deal with the possible issue(s) of noninvariant model-identification parameters, we followed a two-step approach in which the model-identification parameters were selected and evaluated on LMI.

First, the model-identification parameters were selected by comparing the one-factor CFA models of the pretest and posttest based on their factor loading estimates and threshold estimates. Obviously, these models also needed to be identified first before the parameters could be compared. Consequently, we fixed the factor loading of the first item to 1, which is an arbitrary choice. From the remaining items, a marker item was selected based on a trade-off between a high-factor loading estimate for the pretest and posttest (Liu et al., 2017) and a high probability of having an invariant factor loading and two invariant thresholds (i.e., a small difference between the pretest and posttest estimates). Subsequently, we also selected the threshold parameters for the rest of the items based on a high probability of having an invariant threshold (i.e., the thresholds with the smallest difference between the pretest and posttest estimates). The differences between the factor loading estimates as well as the threshold estimates were calculated by subtracting the pretest estimate from the posttest estimate.

Second, we evaluated whether the selected parameters were sufficiently invariant over time. To evaluate LMI, the *baseline invariance model* was compared with the *loading invariance model*. For details about the criteria used to evaluate sufficient invariance, see the section “Tenability of equality constraints over time” above. If these criteria were not met, other parameters were selected for model-identification, and the evaluation of LMI was repeated (Yoon & Millsap, 2007).

5.3.3.5 Practical significance of an invariance violation

When the assumption of LMI is violated, it should be investigated how relevant this violation may be for clinical practice. Specifically, it should be investigated when (i.e., which test-occasion) and where (i.e., which item and response category) the violation has a substantial impact, and to what degree changes in test scores are affected. The findings reveal the usefulness of the measurement to assess change in psychopathology over time. Furthermore, they may help researchers to generate hypotheses as to why the lack of LMI occurs.

5.3.3.5.1 Test-occasions, items, and response categories

Liu and West (2018) proposed to evaluate the practical significance of an invariance violation in ordered-categorical data using two methods. The first methodology is used to investigate to what degree each response category of each item at each measurement occasion is impacted by an invariance violation. To accomplish this, the methodology uses model-predicted probabilities (Liu et al., 2017). These probabilities are *estimations* of the percentage of respondents that endorse each response category on each item at each test-occasion, assuming a specific invariance model. For example, it can be estimated how many respondents would endorse the first item at the pretest with response category *never*, assuming the threshold invariance model. This means that the number of predicted probabilities was 280 for each model of the Depression item bank (2 test-occasions * 28 items * 5 response categories) and 290 for each model of the Anxiety item bank (2 test-occasions * 29 items * 5 response categories).

Model-predicted probabilities were estimated for a model assuming the strictest *achieved* type of LMI (i.e., the strictest model that showed sufficient fit) and a model assuming a stricter, *violated* type of LMI (i.e., the first model that did not show sufficient fit). We then calculated the differences between the predicted probabilities of the models (i.e., the predicted probability of the model assuming a violated type of LMI minus the predicted probability of the model assuming the strictest achieved type of LMI), which can be considered a measure of the magnitude of an invariance violation. For example, when the model-predicted probability of responding to the first item at the pretest with response category *never* is 20% in the loading invariance model and 30% in the threshold invariance model, 10% of the respondents are predicted to choose a different response category under the assumption of different invariance models. Liu et al. (2017) did not suggest a specific cutoff value to interpret this difference, but they used a difference of 5% as illustration for a small impact based on 749 respondents. We chose to follow their example, meaning that when this cutoff was exceeded, more than 5% of the patients are expected to choose a different response category for a specific item at a specific test-occasion, assuming different types of LMI.

5.3.3.5.2 Mean latent change score

The second methodology to evaluate the practical significance of an invariance violation in ordered-categorical data uses the *estimated* mean latent change score (i.e., the difference between the estimated common factor means of the pretest and posttest). This methodology was introduced by Liu and West (2018) for a specific type of longitudinal model (i.e., the latent growth model) and can be extended to the longitudinal model with two factors. This application consists of three steps.

The first step was to create a subset of items (i.e., the anchor set) that showed a specific type of LMI sufficiently (e.g., threshold invariance), which would be used in the second step to evaluate the remaining items on that type of LMI. Some authors use all items (or all items except one) to evaluate them individually on a specific type of LMI, but it has been shown to be more accurate if these evaluations are solely based on a group of invariant items (i.e., between 10% and 20% of the full item set; Woods, 2009). We therefore created an anchor set using the following steps. First, all items were evaluated individually with the χ^2 scaled difference test by comparing the model assuming a stricter, violated type of LMI to the same model minus the model-specific equality constraint(s) for 1 item. Next, 20% of the items (i.e., six items for both item banks) were selected as anchor set based on the lowest $\Delta\chi^2$ values. Finally, it was evaluated with the χ^2 scaled difference test whether the anchor set was sufficiently invariant. This was done by comparing the model assuming the strictest achieved type of LMI (e.g., the loading invariance model) to the same model including equality constraints on the anchor set (in our example that would be the inclusion of equality constraints on the thresholds of the anchor set items). If the anchor set did not show sufficient invariance, we removed the additional equality constraint(s) from the item that earlier showed the highest $\Delta\chi^2$ statistic of the anchor set items, and repeated the evaluation of LMI.

The second step was to establish which additional items showed a specific type of LMI sufficiently. To accomplish this, all items were evaluated individually with the χ^2 scaled difference test by comparing the model assuming the strictest achieved type of LMI including equality constraints on the anchor set with the same model including the equality constraint(s) on one additional item. For example, to evaluate which additional items showed sufficient threshold invariance, the loading invariance model including threshold constraints on the anchor set was compared with the same model including threshold constraints on one additional item, and this was repeated for all items.

The third and final step was to assess the impact of an invariance violation on the mean latent change score. To accomplish this, the *relative mean change* was calculated between the model with equality constraints on all invariant items (i.e., the *partial invariance model*) and the model assuming a stricter, violated type of LMI (i.e., the *full invariance model*). This relative mean change was calculated as the difference between the mean latent change score of these two models, divided by the mean latent change score of the full invariance model. As mean latent change score, we used the estimated common factor mean of the posttest as this equals the mean latent change score in a longitudinal model in which the common factor mean of the pretest was set to 0 for model-identification purposes. Following the suggestion of Kaplan (1989, as cited in Flora & Curran, 2004), a relative mean change value larger than 10% was considered as indicative of substantial bias. When this was the case, the modeling sequence was continued with the partial invariance model. Otherwise, the modeling sequence was continued with the full invariance model.

5.3.3.6 Software

We performed all analyses separately for the Depression and the Anxiety item banks in the statistical environment R (R Core Team, 2017). EFA was conducted with the R package psych (Version 1.5.4; Revelle, 2013); CFA was conducted with the R package lavaan (Version 0.5-18; Rosseel, 2012) using theta parametrization and the diagonally weighted least squares estimator with robust standard errors and a mean and variance adjustment (i.e., WLSMV in lavaan; Liu et al., 2017). Furthermore, both factor analyses were conducted on the polychoric correlation matrix (Bollen, 1989). For some guidelines on selecting a software package, an estimation method, and a framework for analysis (i.e., factor analysis or item response theory [IRT]) for investigating LMI with ordered-categorical data, see Li, 2016; Liu et al., 2017).

5.4 Results

5.4.1 Demographic characteristics

The eligible sample consisted of 13,802 patients (Depression, *n* = 8,372; Anxiety, *n* = 5,430). Of these patients, 13,067 (Depression, *n* = 7,715; Anxiety, *n* = 5,352) were invited to respond to the pretest and 5,383 (Depression, *n* = 3,031; Anxiety, *n* = 2,352) also completed it (pretest response rate Depression item bank = 39.3%; pretest response rate Anxiety item bank = 43.9%). Of the patients with completed pretests, 2,962 patients (Depression, *n* = 1,561; Anxiety, *n* = 1,401) were invited to respond to the posttest and 1,253 patients (Depression, *n* = 664; Anxiety, *n* = 589) also completed it (posttest response rate Depression item bank = 42.5%; posttest response rate Anxiety item bank = 42.0%). None of the patients with a completed pretest and posttest had specific missing item responses. Consequently, we did not have to exclude any more patients for not meeting our first inclusion criterium. We did exclude 85 more patients for not meeting the remaining inclusion criteria (i.e., the posttest was administered less than one month after the pretest, and/or before the first treatment session). Our final study sample therefore consisted of *n* = 640 for the Depression item bank (total sample response rate = 7.6%; 62% female; M_{age} = 41.3 years, *SD* = 13.4, range 18–77) and *n* = 528 for the Anxiety item bank (total sample response rate = 9.7%; 65% female; M_{age} = 37.1 years, *SD* = 12.9, range 18–73). These final samples did not contain sparse data (i.e., missing specific item response *categories* within items). Consequently, LMI could be investigated in a straightforward fashion (Liu et al., 2017).

Table 5.1 Item *M* (*SD*) for the pretest and posttest of the Depression item bank.

Item code	Item	Pretest	Posttest
EDDEP04	I felt worthless	3.44 (1.04)	2.88 (1.06)
EDDEP05	I felt that I had nothing to look forward to	3.48 (1.09)	2.95 (1.16)
EDDEP06	I felt helpless	3.42 (1.01)	2.93 (1.09)
EDDEP07	I withdrew from other people	3.57 (0.93)	3.10 (1.05)
EDDEP09	I felt that nothing could cheer me up	3.49 (0.96)	2.97 (1.11)
EDDEP14	I felt that I was not as good as other people	3.53 (1.10)	3.03 (1.14)
EDDEP17	I felt sad	3.75 (0.89)	3.25 (1.04)
EDDEP19	I felt that I wanted to give up on everything	3.32 (1.05)	2.80 (1.12)
EDDEP21	I felt that I was to blame for things	3.23 (1.11)	2.78 (1.12)
EDDEP22	I felt like a failure	3.28 (1.19)	2.82 (1.18)
EDDEP23	I had trouble feeling close to people	3.20 (1.12)	2.89 (1.14)
EDDEP26	I felt disappointed in myself	3.68 (1.04)	3.20 (1.12)
EDDEP27	I felt that I was not needed	3.35 (1.14)	2.95 (1.18)
EDDEP28	I felt lonely	3.64 (1.08)	3.18 (1.18)
EDDEP29	I felt depressed	3.85 (1.02)	3.15 (1.20)
EDDEP30	I had trouble making decisions	3.57 (0.98)	3.08 (1.12)
EDDEP31	I felt discouraged about the future	3.76 (1.05)	3.20 (1.23)
EDDEP35	I found that things in my life were overwhelming	3.28 (1.11)	2.88 (1.14)
EDDEP36	I felt unhappy	3.78 (1.00)	3.18 (1.13)
EDDEP39	I felt I had no reason for living	2.72 (1.31)	2.28 (1.23)
EDDEP41	I felt hopeless	3.19 (1.09)	2.75 (1.16)
EDDEP42	I felt ignored by people	2.80 (1.05)	2.52 (1.06)
EDDEP44	I felt upset for no reason	3.12 (1.08)	2.70 (1.09)
EDDEP45	I felt that nothing was interesting	3.37 (1.04)	2.87 (1.13)
EDDEP46	I felt pessimistic	3.44 (1.02)	2.98 (1.12)
EDDEP48	I felt that my life was empty	3.43 (1.13)	2.91 (1.23)
EDDEP50	I felt guilty	3.36 (1.13)	2.90 (1.16)
EDDEP54	I felt emotionally exhausted	3.85 (1.06)	3.28 (1.21)

As the response rates were small, additional tests were performed for each item bank to examine whether the composition of the included patients was similar to that of the nonincluded patients. For the variable gender, we investigated the effect size Pearson's residual, following the suggestion of 2.00 as cutoff value for indicating a systematic difference between the observed and expected number of respondents (Agresti & Kateri, 2011). For the variables age and pretest score (i.e., the sum of the item scores), we investigated the effect size Cohen's *d* (i.e., the difference between the mean ages/pretest scores divided by the pooled *SD*), following the guideline proposed by Cohen to interpret the size of the effect (1988): 0.20 = small effect, 0.50 = medium effect, 0.80 = large effect. The results showed for both item banks that Pearson's residuals were all below 2.00 and Cohen's *d*s were below 0.20. We therefore concluded that the included patients for each item bank did not differ substantially from the nonincluded patients regarding the variables gender, age and pretest score.

Table 5.2 Item *M* (*SD*) for the pretest and posttest of the Anxiety item bank.

Item code	Item	Pretest	Posttest
EDANX01	I felt fearful	3.57 (0.86)	3.10 (0.96)
EDANX02	I felt frightened	2.92 (1.11)	2.47 (1.06)
EDANX03	It scared me when I felt nervous	3.16 (1.15)	2.89 (1.04)
EDANX05	I felt anxious	3.54 (0.94)	3.11 (0.96)
EDANX07	I felt like I needed help for my anxiety	3.51 (1.12)	2.81 (1.12)
EDANX08	I was concerned about my mental health	3.29 (1.15)	2.78 (1.12)
EDANX12	I felt upset	3.21 (1.05)	2.83 (1.05)
EDANX13	I had a racing or pounding heart	2.90 (1.16)	2.61 (1.11)
EDANX16	I was anxious if my normal routine was disturbed	2.92 (1.22)	2.65 (1.17)
EDANX18	I had sudden feelings of panic	3.06 (1.19)	2.59 (1.12)
EDANX20	I was easily startled	2.71 (1.23)	2.39 (1.13)
EDANX21	I had trouble paying attention	3.07 (1.10)	2.88 (1.12)
EDANX24	I avoided public places or activities	2.63 (1.33)	2.34 (1.24)
EDANX26	I felt fidgety	3.74 (0.97)	3.32 (1.05)
EDANX27	I felt something awful would happen	2.64 (1.24)	2.29 (1.15)
EDANX30	I felt worried	3.72 (0.95)	3.26 (1.02)
EDANX33	I felt terrified	2.36 (1.23)	1.98 (1.07)
EDANX37	I worried about other people's reactions to me	3.13 (1.23)	2.82 (1.22)
EDANX40	I found it hard to focus on anything other than my anxiety	3.25 (1.13)	2.83 (1.14)
EDANX41	My worries overwhelmed me	3.01 (1.19)	2.53 (1.21)
EDANX44	I had twitching or trembling muscles	2.32 (1.17)	2.15 (1.08)
EDANX46	I felt nervous	3.47 (0.96)	3.14 (0.96)
EDANX47	I felt indecisive	3.15 (1.12)	2.80 (1.11)
EDANX48	Many situations made me worry	3.22 (1.06)	2.81 (1.11)
EDANX49	I had difficulty sleeping	3.22 (1.31)	2.91 (1.29)
EDANX51	I had trouble relaxing	3.73 (0.98)	3.31 (1.12)
EDANX53	I felt uneasy	3.28 (1.00)	2.96 (1.07)
EDANX54	I felt tense	3.74 (0.91)	3.36 (1.00)
EDANX55	I had difficulty calming down	3.15 (1.06)	2.77 (1.12)

5.4.2 Descriptive statistics

Table 5.1 (Depression item bank) and Table 5.2 (Anxiety item bank) display the mean item scores (*SD'*s) of the pretest and posttest. All items showed a decrease in mean from pretest to posttest, ranging from 0.27 to 0.71 for the Depression item bank and from 0.17 to 0.70 for the Anxiety item bank.

Concerning the pretest to posttest interval, the median was 238.50 days for the Depression item bank (range = 43.00–803.00, interquartile range = 219.00–281.00) and 181.50 days for the Anxiety item bank (range = 39.00 – 825.00, interquartile range = 158.00–278.25). These results indicate that the degree of uniformity in the pretest to posttest interval was quite low for both item banks.

5.4.3 Model-identification parameters

For the *Depression item bank*, item EDDEP05 (i.e., *I felt that I had nothing to look forward to*) was selected as marker item because it showed a large factor loading for both pretest ($\lambda_2 = 0.93$) and posttest ($\lambda_2 = 0.93$) that did not differ between test-occasions. Furthermore, we found relatively moderate differences between the test-occasions in the first and second threshold of this item ($\Delta\upsilon_1 = 0.41$, $\Delta\upsilon_2 = 0.78$). In addition, we selected the first threshold of the remaining items for showing the smallest difference between the test-occasions' estimates.

The evaluation of LMI in the selected parameters showed that the loading invariance model was not rejected by the χ^2 scaled difference test (see Table 5.3, line 1 and line 2 of the Depression item bank). Furthermore, all modification indices of the constrained parameters were below 5. We concluded that the selected parameters of the Depression item bank were sufficiently invariant for model-identification.

Table 5.3 Fit statistics for the longitudinal CFA (invariance) models of the Depression and Anxiety item banks.

Item bank	Invariance model	*df*	χ^2	Δdf	$\Delta\chi^2$	*p*	CFI	SRMR	RMSEA
Depression	Baseline	1455	5449.131	-	-	-	0.955	0.051	0.057
	Loading	1482	5472.148	21.165	30.087	0.094	0.955	0.051	0.056
	Threshold	1565	5612.895	52.499	144.376	0.000	0.954	0.051	0.055
	Unique Factor	1593	6068.212	22.635	94.926	0.000	0.956	0.052	0.053
Anxiety	Baseline	1565	5006.248	-	-	-	0.954	0.055	0.054
	Loading	1593	5035.036	22.580	33.380	0.067	0.954	0.055	0.053
	Threshold	1679	5206.926	50.578	144.475	0.000	0.953	0.055	0.052
	Factor Variance	1708	5656.931	23.699	89.184	0.000	0.955	0.057	0.051

Note. *df* = degrees of freedom; χ^2 = unscaled chi-square; Δdf = scaled difference in degrees of freedom based on the preceding model; $\Delta\chi^2$ = scaled difference in chi-square based on the preceding model; *p* = *p*-value for the chi-square scaled difference test; CFI = scaled comparative fit index; SRMR = standardized root-mean-square residual; RMSEA = scaled root-mean-square error of approximation.

For the *Anxiety item bank*, we selected item EDANX40 (i.e., *I found it hard to focus on anything other than my anxiety*) as marker item because the factor loading was adequate for the pretest (λ_2 = 0.69) and posttest (λ_2 = 0.77) and differed only somewhat between test-occasions ($\Delta\lambda_2$ = 0.08). Furthermore, we found relatively moderate differences between the test-occasions in the first and second threshold (Δυ1 = 0.34, Δυ2 = 0.57). In addition, we selected the first threshold for almost all remaining items because the difference between the test-occasions' estimates was the smallest, except for items EDANX03, EDANX21, and EDANX46, for which the smallest difference was found for the second threshold.

The evaluation of LMI in the selected parameters showed that the loading invariance model was rejected. Furthermore, the modification indices of the constrained parameters were above 5 for both the factor loading and the first threshold of item EDANX05 (i.e., *I felt anxious*). When we changed the equality constraint of this item from the first to the second threshold, the loading invariance model was no longer rejected (see Table 5.3, line 1 and line 2 of the Anxiety item bank). Moreover, the modification indices of the constrained parameters were all below 5. We concluded that the (adjusted) selection of parameters for the Anxiety item bank were sufficiently invariant for model-identification.

5.4.4 Unidimensionality of the item banks

EFA showed that the first and second factor of the pretest explained 58% and 6% of the variance for the Depression item bank, and 54% and 6% for the Anxiety item bank, respectively. For the posttest, the first and second factor explained 68% and 4% of the variance for the Depression item bank, and 63% and 5% for the Anxiety item bank, respectively. The variances explained by the first factor were above 20% and the ratios of variance explained by the first to second factor were larger than 4. Both item banks were therefore considered to be sufficiently unidimensional at both measurements. Moreover, as both indices of unidimensionality improved from pretest to posttest, the constructs Depression and Anxiety can be considered to become more homogeneous over time.

Table 5.4 Fit statistics for the one-factor CFA models of the Depression and Anxiety item banks.

Item bank	Measurement	*df*	CFI	SRMR	RMSEA
Depression	Pretest	350	0.916	0.063	0.111
	Posttest	350	0.964	0.042	0.097
Anxiety	Pretest	377	0.910	0.067	0.106
	Posttest	377	0.959	0.052	0.094

Note. *df* = degrees of freedom; CFI = scaled comparative fit index; SRMR = standardized root-mean-square residual; RMSEA = scaled root-mean-square error of approximation.

In Table 5.4, the fit statistics are presented for all evaluated one-factor CFA models. For the pretest, the CFI and SRMR indicated adequate model fit for both item banks; the RMSEA indicated a moderate fit. For the posttest, the model fit improved for both item banks according to all fit statistics. Moreover, the fit changed from adequate to good for the CFI of both item banks and the SRMR of the Depression item bank. These results are in line with the findings of EFA: the item banks showed sufficient unidimensionality at both test-occasions, and the constructs Depression and Anxiety became more homogeneous over time.

5.4.5 Tenability of equality constraints

In Table 5.3, the fit statistics are presented for all evaluated longitudinal CFA models. The results were highly similar for both item banks. According to the CFI, SRMR, and RMSEA, all models showed good model fit. The χ^2 scaled difference test showed that including constraints on factor loadings did not worsen the model fit, but including constraints on thresholds and residual variances did worsen the model fit. Furthermore, for the Depression item bank, modification indices above 5 were found for threshold constraints of 8 items and residual variance constraints of 10 items. For the Anxiety item bank, modification indices above 5 were found for threshold constraints of 9 items and residual variance constraints of 10 items. These results indicate that equality constraints on factor loadings were tenable in the longitudinal model, but equality constraints on thresholds and residual variances were not tenable. In other words, we found for both item banks that loading invariance was achieved, but threshold invariance and unique factor invariance were violated.

5.4.6 The magnitude and practical significance of the invariance violations

5.4.6.1 Threshold invariance

In Table 5.5 (Depression item bank) and Table 5.6 (Anxiety item bank), the differences are presented between the model-predicted probabilities of the loading invariance model and the threshold invariance model. For the Depression item bank, all of the 280 differences were below the cutoff value of 5%. Both the lowest and highest difference were found for response Category 4 (i.e., *often*) of item EDDEP17 (*I felt sad*). The number of respondents that are predicted to endorse this response category on this item at the pretest was 3.9% lower in the threshold model than in the loading invariance model, while at the posttest it was 3.8% higher. In addition, for the Anxiety item bank, only 2 out of 290 differences were somewhat above the cutoff value of 5%. The number of respondents that are predicted to endorse response Category 2 (i.e., *rarely*) on item EDANX07 (i.e., *I felt like I needed help for my anxiety*) at the pretest was 6.1% higher in the threshold model than in the loading invariance model while it was 5.6% lower at the posttest. Consequently, the *overall* results indicate that the rejection of threshold invariance does not substantially affect the endorsement of a specific response category of a specific item administered at a specific test-occasion.

To evaluate to what extent the mean latent change score was impacted by the threshold invariance violation, an anchor set was first created for each item bank. We selected items EDDEP05, EDDEP21, EDDEP28, EDDEP31, EDDEP35, and EDDEP48 as anchor set for the Depression item bank, and items EDANX12, EDANX20, EDANX40, EDANX41, EDANX46, and EDANX49 for the Anxiety item bank. Both of these item sets showed sufficient threshold invariance according to the χ^2 scaled difference test. When we used these item sets to evaluate the other items on threshold invariance, items EDDEP04, EDDEP06, EDDEP07, EDDEP09, EDDEP17, EDDEP23, EDDEP29, EDDEP30, EDDEP36, EDDEP46, and EDDEP54 did not

Table 5.5 Differences between the model-predicted probabilities of choosing specific response categories on specific items at specific test-occasions based on the loading invariance and the threshold models for the Depression item bank.

	Never		Rarely		Sometimes		Often		Always	
Item code	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
EDDEP04	-0,005	0,006	0,009	-0,010	0,016	-0,016	-0,006	0,006	-0,014	0,014
EDDEP05	-0,005	0,006	-0,011	0,013	0,027	-0,031	-0,008	0,009	-0,002	0,002
EDDEP06	-0,005	0,006	-0,002	0,002	0,030	-0,031	-0,036	0,034	0,013	-0,011
EDDEP07	-0,002	0,003	0,015	-0,018	0,014	-0,014	-0,029	0,030	0,002	-0,002
EDDEP09	-0,005	0,005	0,027	-0,030	-0,018	0,020	-0,017	0,015	0,013	-0,010
EDDEP14	-0,004	0,005	0,003	-0,004	0,016	-0,016	-0,011	0,011	-0,004	0,004
EDDEP17	-0,002	0,002	0,018	-0,019	0,017	-0,016	-0,039	0,038	0,005	-0,005
EDDEP19	-0,006	0,007	0,017	-0,020	0,004	-0,003	-0,019	0,019	0,003	-0,003
EDDEP21	-0,005	0,006	0,000	-0,001	0,008	-0,009	-0,004	0,004	0,001	-0,001
EDDEP22	-0,008	0,010	-0,018	0,020	0,029	-0,033	-0,003	0,003	0,000	0,000
EDDEP23	-0,005	0,006	-0,010	0,010	0,001	-0,003	-0,010	0,005	0,023	-0,018
EDDEP26	-0,003	0,003	-0,008	0,009	0,013	-0,015	0,002	-0,002	-0,005	0,004
EDDEP27	-0,006	0,006	-0,001	0,001	-0,011	0,010	0,007	-0,008	0,011	-0,009
EDDEP28	-0,003	0,004	0,006	-0,007	-0,002	0,002	-0,011	0,010	0,010	-0,009
EDDEP29	-0,004	0,005	0,020	-0,024	0,028	-0,029	-0,029	0,033	-0,014	0,014
EDDEP30	-0,002	0,003	0,030	-0,032	-0,020	0,022	-0,012	0,012	0,005	-0,004
EDDEP31	-0,004	0,005	0,008	-0,010	0,001	-0,001	-0,006	0,007	0,001	-0,001
EDDEP35	-0,005	0,005	0,006	-0,007	-0,018	0,016	0,013	-0,012	0,003	-0,003
EDDEP36	-0,003	0,004	0,000	0,000	0,022	-0,025	-0,007	0,009	-0,012	0,013
EDDEP39	-0,019	0,021	0,010	-0,011	0,004	-0,005	-0,001	0,001	0,007	-0,006
EDDEP41	-0,009	0,011	-0,008	0,009	0,003	-0,005	0,010	-0,011	0,004	-0,004
EDDEP42	-0,007	0,008	0,008	-0,008	-0,022	0,020	0,012	-0,012	0,010	-0,007
EDDEP44	-0,006	0,006	0,024	-0,024	-0,014	0,015	-0,007	0,006	0,003	-0,003
EDDEP45	-0,005	0,006	0,017	-0,019	-0,014	0,015	-0,004	0,004	0,006	-0,005
EDDEP46	-0,004	0,005	0,011	-0,012	0,011	-0,010	-0,033	0,031	0,016	-0,013
EDDEP48	-0,007	0,010	0,003	-0,005	0,004	-0,005	-0,001	0,001	0,002	-0,002
EDDEP50	-0,004	0,005	-0,009	0,012	0,031	-0,037	-0,007	0,008	-0,010	0,012
EDDEP54	-0,004	0,004	0,010	-0,011	0,031	-0,032	-0,031	0,033	-0,007	0,006

Note. T1 = pretest; T2 = posttest; each difference is based on the model-predicted probability of the threshold invariance model minus the model-predicted probability of the loading invariance model.

show sufficient invariance for the Depression item bank, and items EDANX01, EDANX03, EDANX05, EDANX07, EDANX08, EDANX18, EDANX26, EDANX30, EDANX51, and EDANX53 did not show sufficient invariance for the Anxiety item bank. However, the relative mean change between the full threshold invariance model and the partial threshold invariance model did not exceed the cutoff value of 10% for both item banks (although that of the Anxiety item bank came close to 10%). For the Depression item bank, the mean latent change score was -0.81 for the full threshold invariance model and -0.76 for the partial threshold invariance model, resulting in a relative mean change of 6.82%. For the Anxiety item bank, the mean latent change score was -0.61 for the full threshold invariance model and -0.55 for the partial threshold invariance model, resulting in a relative mean change of 9.58%. These results indicate that the bias caused by the threshold invariance violation on the mean latent change score was not substantial for both item banks. Consequently, we decided to continue the modeling sequence for both item banks using the full threshold invariance model.

Table 5.6 Differences between the model-predicted probabilities of choosing specific response categories on specific items at specific test-occasions based on the loading invariance and the threshold models for the Anxiety item bank.

	Never		Rarely		Sometimes		Often		Always	
Item code	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
EDANX01	0.001	-0.002	0.015	-0.015	0.018	-0.017	-0.030	0.032	-0.004	0.004
EDANX02	-0.007	0.008	0.021	-0.022	-0.018	0.018	0.013	-0.013	-0.008	0.009
EDANX03	-0.032	0.038	0.014	-0.020	0.011	-0.011	0.007	-0.007	-0.001	0.001
EDANX05	-0.016	0.022	0.009	-0.012	0.037	-0.041	-0.024	0.025	-0.006	0.006
EDANX07	0.000	0.002	0.061	-0.056	0.010	-0.010	-0.050	0.045	-0.021	0.019
EDANX08	0.000	0.002	0.018	-0.019	0.020	-0.017	-0.034	0.031	-0.003	0.002
EDANX12	-0.001	0.000	0.006	-0.005	-0.001	0.002	-0.002	0.002	-0.002	0.002
EDANX13	-0.003	0.002	-0.008	0.009	0.002	-0.002	0.013	-0.012	-0.004	0.003
EDANX16	-0.004	0.003	-0.017	0.017	0.010	-0.010	0.003	-0.003	0.008	-0.006
EDANX18	-0.007	0.007	0.011	-0.011	0.026	-0.026	-0.023	0.023	-0.007	0.007
EDANX20	-0.009	0.008	-0.003	0.003	0.012	-0.011	0.003	-0.003	-0.003	0.003
EDANX21	-0.013	0.015	-0.004	0.001	-0.005	0.005	0.011	-0.012	0.010	-0.009
EDANX24	-0.016	0.015	0.006	-0.005	0.003	-0.003	0.001	-0.001	0.007	-0.006
EDANX26	0.001	-0.004	0.014	-0.012	-0.008	0.008	0.000	0.001	-0.007	0.007
EDANX27	-0.016	0.016	0.000	-0.001	0.013	-0.012	0.000	0.000	0.003	-0.003
EDANX30	0.002	-0.003	0.012	-0.012	0.021	-0.021	-0.025	0.025	-0.011	0.010
EDANX33	-0.029	0.031	0.020	-0.023	0.006	-0.006	0.010	-0.010	-0.007	0.008
EDANX37	0.003	-0.001	-0.009	0.007	0.015	-0.014	-0.018	0.015	0.010	-0.008
EDANX40	-0.003	0.002	-0.015	0.018	0.022	-0.025	-0.005	0.004	0.001	-0.001
EDANX41	-0.009	0.011	0.016	-0.019	-0.007	0.007	-0.006	0.005	0.006	-0.005
EDANX44	-0.019	0.016	-0.003	0.004	0.009	-0.010	0.013	-0.012	-0.001	0.001
EDANX46	-0.005	0.005	-0.003	0.003	0.008	-0.009	0.000	0.000	0.000	0.000
EDANX47	0.000	0.000	0.003	-0.003	0.008	-0.007	-0.023	0.019	0.012	-0.009
EDANX48	-0.001	0.003	0.009	-0.010	-0.009	0.009	-0.007	0.005	0.008	-0.006
EDANX49	0.000	0.000	0.004	-0.004	-0.013	0.012	0.001	-0.002	0.009	-0.007
EDANX51	0.002	-0.001	0.026	-0.026	-0.028	0.026	-0.009	0.008	0.009	-0.007
EDANX53	0.001	0.001	-0.002	0.001	-0.006	0.005	-0.019	0.011	0.027	-0.018
EDANX54	0.001	-0.003	0.015	-0.014	-0.020	0.020	-0.003	0.002	0.006	-0.005
EDANX55	-0.003	0.004	0.009	-0.009	-0.015	0.014	-0.005	0.002	0.014	-0.010

Note. T1 = pretest; T2 = posttest; each difference is based on the model-predicted probability of the threshold invariance model minus the model-predicted probability of the loading invariance model.

5.4.6.2 Unique factor invariance

For the unique factor invariance violation, all differences between the model-predicted probabilities of the threshold invariance model and the unique factor invariance model did not exceed the cutoff value of 5% for both item banks. The differences were found to be between -2.7% and 2.6% for the Depression item bank, and between -3.7% and 3.2% for the Anxiety item bank. Consequently, the overall results indicate that the rejection of unique factor invariance does not substantially affect the endorsement of a specific response category of a specific item administered at a specific test-occasion.

In addition, we selected items EDDEP19, EDDEP29, EDDEP30, EDDEP41, EDDEP42, and EDDEP54 as anchor set for the Depression item bank, and items EDANX12, EDANX13, EDANX24, EDANX26, EDANX37, and EDANX41 as anchor set for the Anxiety item bank. The item set of the Depression item bank showed sufficient invariance according to the χ^2 scaled difference test. For the Anxiety item bank, however, we had to remove the equality constraints of item EDANX37 and EDANX41 before the anchor item set was sufficiently invariant. When we used these item sets to evaluate the other items on unique factor invariance, items EDDEP04, EDDEP06, EDDEP09, EDDEP17, EDDEP23, EDDEP27, EDDEP28, EDDEP35, EDDEP44, EDDEP45, and EDDEP50 did not show sufficient invariance for the Depression item bank, and items EDANX03, EDANX07, EDANX08, EDANX27, EDANX46, EDANX47, EDANX48, EDANX51, EDANX53, EDANX54, and EDANX55 did not show sufficient invariance for the Anxiety item bank. However, the relative mean change between the full unique factor invariance model and the partial unique factor invariance model did not exceed the cutoff value of 10% for both item banks. For the Depression item bank, the mean latent change score was -0.84 for the full unique factor invariance model and -0.85 for the partial unique factor invariance model, resulting in a relative mean change of -1.88%. For the Anxiety item bank, the mean latent change score was -0.65 for the full unique factor invariance model and -0.64 for the partial unique factor invariance model, resulting in a relative mean change of 2.04%. These results indicate that the bias caused by the unique factor invariance violation was not substantial for the mean latent change score of both item banks.

5.5 Discussion

Until now, none of the PROMIS item banks were evaluated on LMI. In the present study, LMI was investigated in the Dutch-Flemish PROMIS adult v1.0 item banks for Depression and Anxiety using two clinical samples with mood and anxiety disorders. To study LMI, we used factor analysis to evaluate whether (a) the item banks were sufficiently unidimensional at two test-occasions, and (b) the measured constructs remained the same over time. Moreover, we assessed two effect sizes relevant for test users to evaluate the practical significance of the found invariance violations. Specifically, we investigated when (i.e., which test-occasion) and where (i.e., which item and response category) the LMI violations had a substantial impact (Liu et al., 2017), and to what degree changes in test scores were affected (Liu & West, 2018).

Both EFA and one-factor CFA indicated that the item banks were sufficiently unidimensional. The measured constructs, however, became more homogeneous over time, indicating some change within the constructs. Longitudinal CFA models confirmed this change in the constructs as equality constraints on thresholds and residual variances were shown to be untenable. These results indicate that the item banks may lead to biased pretest to posttest change scores. Similar results were found by Fokkema et al. (2013) and Fried et al. (2016) for other instruments measuring depression.

We performed two analyses to gauge the practical significance of the invariance violations using tentatively determined rules of thumb. In the first analysis, we found that none of the response categories of each item at each test-occasion was substantially affected by the violations. Only the Anxiety item bank showed that the number of respondents predicted to endorse response Category 2 (i.e., *rarely*) on item EDANX07 (i.e., *I felt like I needed help for my anxiety*) at the pretest was 6.1% higher in the threshold invariance model than in the loading invariance model, while at the posttest it was 5.6% lower. This item is included in two out of four of the PROMIS short-forms (i.e., short-form 6a and 8a), but because the differences can be considered somewhat small, the impact on scores will likely be small. In addition, the second practical significance analysis showed that none of the relative mean changes between the estimated mean latent change scores of the pretest and posttest exceeded our cutoff value for substantial bias. These results suggest that the item banks provide sufficiently invariant latent factor scores for use in clinical practice. We should stress, however, that the practical significance analysis of Liu et al. (2017) still needs to be investigated further to confirm that it is equally sensitive to invariance violations of factor loadings, thresholds, and residual variances. Moreover, the detection of individual (non)invariant items, performed in the practical significance analysis of Liu and West (2018), is complex and many procedures are, to some extent, conceptually or statistically flawed (Bechger & Maris, 2015; Borsboom, 2016). Therefore, we cannot rule out that the Dutch-Flemish PROMIS item banks for Depression and Anxiety lack LMI to at least some degree for patients with a mood and anxiety disorder. In particular, the Anxiety item bank may be vulnerable for LMI, as the relative mean change for the threshold invariance violation came close to the proposed cutoff value for substantial bias. Thus, the mean latent change score may not entirely represent actual changes in the constructs over time as measured through the item banks.

Assuming at least some invariance violations, Fried et al. (2016) argued that possible problems with LMI do not imply that test scores are not useful in clinical practice or that they should not be interpreted, as we can safely assume that the sum of symptoms does provide information about the general psychopathological burden people carry. This means that when an instrument shows practically significant invariance violations, it may still be used to assess clinical subjects meaningfully, albeit with somewhat more caution. Furthermore, in the case of assessing individuals, a test user should be aware that an instrument is a tool designed to help practitioners as a complement to their clinical expertise and not as an objective decision tool (i.e., each test-score includes measurement error; Greenhalgh et al., 2018). Therefore, professionals should not only discuss (changes in) test scores with their patients, but also question them on the development of specific symptoms and the progress towards their treatment goals. In addition, when assessing groups, researchers should decide whether the possible bias due to invariance violations is acceptable for their research question(s) and discuss the possible consequences when reporting their findings (Borsboom, 2006).

For further research, we have the following suggestions. First, we suggest to investigate whether the degree of LMI differs between specific subgroups, which may help explain the results. For example, Fokkema et al. (2013) found that LMI in the Beck Depression Inventory (Beck & Beamesderfer, 1974) was weaker for patients who received psychotherapy than for those who only received medication and additional clinical management. The authors suggested that less invariant measurement may be found in patients undergoing psychological treatments for depression due to a larger focus on the psychoeducation of patients. Thus, by studying specific subgroups, the authors found differences in the degree of LMI, and generated a hypothesis that may be studied further to possibly explain these differences. For more information on possible explanations for a lack of LMI, see Fried et al. (2016).

Second, it may be recommended to investigate whether modifications of the item banks will increase the degree of LMI. Specifically, it may be recommended to investigate the removal of items as rewriting or replacing them would be more complicated considering the comprehensive process of PROMIS to establish their item banks (Pilkonis et al., 2011). However, we should again stress that detecting individual noninvariant items is complex and many procedures are, to some extent, conceptually or statistically flawed. For example, Borsboom (2006) showed that using different methods for detecting noninvariant items can lead to different results. Also, researchers should realize that modifying an item bank, even when it concerns only one item, may lead to changes in the construct it measures. As a result, the set of items that shows invariance violations may change too (i.e., items that first showed sufficient invariance may found to be noninvariant for the modified item bank, and vice versa; Bechger & Maris, 2015). Furthermore, removing items could adversely affect content validity, and it can even result in more biased change scores because the equilibrium of biasing effects needed for cancellation to occur is disturbed (Borsboom, 2006). For these reasons, caution is warranted when item banks are modified. Alternatively, detecting individual noninvariant items may help to generate hypotheses about the origin of noninvariance. For example, it can be noted in our study that the individual items that showed the largest LMI violations assess anxiety very broadly (e.g., item EDANX05, *I felt anxious* or EDANX07 *I felt like I needed help for my anxiety*). This might imply that the anxiety construct as measured by the item bank actually consists of multiple constructs (e.g., generalized anxiety, social anxiety, and panic). In this case, bias may occur because patients think of different types of anxiety at separate test-occasions.

Third, we suggest studying LMI in patients with primary diagnoses other than anxiety or depression (e.g., attention deficit disorder, somatoform disorder or personality disorder), as the item banks also bear relevance for these patients. The reason for this is that depression and anxiety are often comorbid conditions (e.g., Löwe et al., 2008). Furthermore, anxiety and depression constitute a prime element of the distress that causes patients to seek help from mental health care professionals, also when their primary diagnosis is for instance a personality disorder (Leyro, Zvolensky, & Bernstein, 2010). In addition, we suggest that LMI is studied in populations without mental health problems, populations not in treatment, and general populations. Although changes in the observed item responses are expected to be low in these populations, it is still fairly unclear what causes a lack of LMI (Fried et al., 2016). Therefore, the assumption of sufficient LMI in populations that do not show a substantial change in severity level over time should be studied.

Fourth, although the current study used a methodology that is the state of the art, additional new methods and software implementations would be welcome to study LMI in more detail. For example, LMI was evaluated in this study within the framework of factor analysis. In this framework, new methodology is available to investigate LMI for multiple group models that may also be extended to longitudinal models (Wu & Estabrook, 2016). Furthermore, although we investigated LMI with factor analysis because all new methodologies used in this study were primarily developed for this framework (Liu et al., 2017; Liu & West, 2018), PROMIS instruments are commonly calibrated using IRT, as it allows for the implementation of CAT. Studying equivalent longitudinal methods based on IRT (Meade & Lautenschlager, 2004; Wang, 2016) would allow for relating LMI violations to the metric used in clinical practice and the established properties of the item banks (Flens et al., 2017, 2019). A third example of new methodology concerns missing data. In the used version of the R package lavaan (i.e., 0.5-18), missing data handling is not available for CFA with ordered-categorical data (i.e., it uses listwise deletion). As missing data is common in longitudinal data, developing new methods that can handle missing data may result in improved parameter estimates.

In addition, the effect sizes used in this study were selected because, together, they provide highly practical information about the indicators of interest for test users (Liu & West, 2018). Specifically, they do not only provide information about the impact of invariance violations on change scores, but also on specific test-occasions, items, and response categories. However, the used rules of thumb for these effect sizes need to be verified in a (simulation) study to assess whether they correspond sufficiently to the proposed degree of bias. Furthermore, other effect sizes may provide additional useful information for test users (e.g., Choi, Gibbons, & Crane, 2011; Kim, Cohen, Alagoz, & Kim, 2007; Liu & West, 2018; Meade, 2010). A comparative (simulation) study on effect sizes and their rules of thumb used to quantify LMI with ordered-categorical indicators and for different applications of the item banks (e.g., full item bank administration, short-form administration, or CAT administration; Reeve et al., 2007) could provide new insights on the matter. In such a study, it could also be assessed whether the effect sizes could be further developed for evaluating LMI in individuals as compared with groups. Borsboom (2006) argued that when instruments are used for assessing individuals, LMI should conform to higher standards because of the increased danger of bias.

Fifth, we suggest to compare the degree of LMI between (a) the item banks and other instruments measuring Depression or Anxiety (e.g., the Center for Epidemiological Studies Depression scale or the Patient Health Questionnaire; Pilkonis et al., 2011) and (b) different languages (e.g., English and Dutch). By performing a comparative LMI study between instruments, test users have more available information to decide which instrument they want to use. Furthermore, it may provide new insights in the type of items that influence the degree of LMI. In addition, by performing a comparative LMI study between different languages, it could be assessed whether the lack of LMI may (also) be a translation problem.

In addition to the evaluation of LMI, the PROMIS item banks need to be studied on their responsiveness. According to the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) terminology (Mokkink et al., 2018), responsiveness (also known as sensitivity to change) refers to the ability to detect change in the measured construct over time (Mokkink et al., 2010), usually assessed by comparing changes in PROMIS scores to changes in one or several legacy instruments. Preferably, responsiveness should be studied for CAT administration rather than full item bank administration, as CAT will likely be the primary mode of administration in Dutch clinical practice. Moreover, we suggest to consider the results of the present study when comparing the responsiveness of the CAT administrations to that of other instruments. With CAT, the number of administered items is generally lower than with a full item bank administration. As a result, bias may be larger than in a full item bank administration as the items have a larger weight in the final test scores, and cancellation of biasing effects is less likely to occur (Borsboom, 2006).

In addition to responsiveness, we suggest to study whether multidimensional computerized adaptive testing (MCAT; Paap, Born, & Braeken, 2019) with the Depression and Anxiety item banks can be more efficient and precise than CAT based on separate unidimensional item banks. In the current study, the item banks were treated as measurements of separate unidimensional constructs because PROMIS deliberately chose to develop their instruments in this way (Cella et al., 2007). Numerous studies, however, show that the constructs depression and anxiety are highly correlated (e.g., de Beurs et al., 2007). Therefore, a logical next step with the PROMIS item banks could be to assess whether MCAT can be applied to the item banks. If this is the case, then LMI should once more be assessed for the multidimensional construct.

A strength of the current study is that the ordered-categorial data of the PROMIS item banks were explicitly treated as ordered-categorial instead of continuous, the latter being usually the case in LMI studies (Liu et al., 2017). Consequently, the item parameters may be more accurate (Rhemtulla et al., 2012). We also used two analyses to study the practical significance of the invariance violations, meaning that we gained information on (a) when (i.e., which test-occasion) and where (which item and response category) the problem occurred and (b) the magnitude of the problem for the parameter of interest in clinical practice (i.e., the mean latent change score; Liu & West, 2018). Finally, the patients' diagnoses were based on a standardized diagnostic interview (i.e., the MINI-plus; Sheehan et al., 1998), which will likely have increased the accuracy of the diagnoses compared with merely using the clinician's point of view (Aboraya, Rankin, France, El-Missiry, & John, 2006). However, although the MINI(-plus) has adequate diagnostic properties, studies did not show sufficient interrater agreement with other diagnostic instruments on detecting a generalized anxiety disorder and a simple phobia (Lecrubier et al., 1997; Sheehan et al., 1998). This may lead to underestimation or overestimation of these diagnoses. Overestimation may be unlikely, as the condition of each patient was deemed sufficiently severe to receive treatment. Underestimation may lead to these disorders being somewhat underrepresented in the present study sample.

In addition, there are several other reasons why the used samples of this study might lack representativeness for the Dutch clinical population. First, although we found that the included patients did not differ substantially from the nonincluded patients in terms of gender, age distributions, and pretest score distributions, we could not evaluate the representativeness of the samples in terms of other variables that may affect LMI, such as type of treatment, comorbidity, or personality traits (e.g., agreeableness). We suggest to include these variables in future LMI studies. Moreover, the data should preferably be collected using stratified sampling (e.g., using stratification variables such as gender, age, education, ethnicity, and region; Flens et al., 2017). Second, we could not assess whether the change score distributions of the final samples were representative for the Dutch clinical population. It may be, for example, that patients with small change scores were more likely to refuse the posttest invitation because they did not respond to treatment. However, if such selection would be at play, it would hard if not impossible to overcome as participation in research is always voluntary. Third, the group who responded to the pretest, but were not invited for the posttest may appear large (Depression item bank, *n* = 1,470; Anxiety item bank, *n* = 951). According to Parnassia Psychiatric Institute (i.e., the mental health care provider that collected the data), reasons for this are diverse. For example, respondents could have dropped out of treatment (e.g., due to long waiting times or spontaneous remission), respondents' diagnoses could have changed during treatment, or treatment could have been terminated before the posttest was administered. As we did not know the specific reason for each individual that was not invited for the posttest, it is difficult to elaborate on how these reasons may have affected the representativeness of the samples for the Dutch clinical population. It may therefore be recommended that future studies administrate more specifically why respondents are not included in the study, but that may require a substantial investment.

The lack of uniformity in pretest to posttest interval could also have affected the results. To investigate the impact of this lack of uniformity on LMI to at least some extent, we repeated our analyses (not shown herein) on a more homogeneous subsample with additional inclusion criteria: (a) the pretest was administered before or on the day of the first treatment session and (b) the pretest and posttest were separated no longer than 12 months (Depression, *n* = 488; Anxiety, *n* = 414). We found that the results were highly similar, which can be seen as some evidence that the pretest to posttest interval is not a highly relevant factor in the degree of LMI. We should note, however, that these findings do not imply that the results would also have been highly similar when uniformity in the pretest to posttest interval was even larger (e.g., the pretest was administered at the first treatment session and the posttest exactly six months later). Unfortunately, we could not apply this larger extent of uniformity because the diminished sample size may result in data that is prone to nonconvergence, improper factor solutions, large standard errors, biased estimates of factor loadings and thresholds, and problematic goodness-of-fit tests (Liu et al., 2017). For future longitudinal studies, we suggest aiming for a higher degree of uniformity in the pretest to posttest interval to investigate more specific hypotheses about the length of the retest interval and LMI.

In addition to this, we suggest to evaluate LMI in more than two test-occasions. By investigating more test-occasions, the results may lead to a better understanding of the causes of invariance violations (e.g., by studying hypotheses concerning the impact of the degree of change on LMI). Extending the analyses of this study to more test-occasions is fairly straightforward. For an illustration of the analyses concerning the tenability of equality constraints and the source of the invariance violations (i.e., which test-occasion, item, and response category), see Liu et al. (2017; 4 test-occasions). For an illustration of the analyses concerning the degree of impact on change scores, see Liu and West (2018; 4 test-occasions). Alternatively, the data sets used in this study could have been split into separate samples (e.g., a short-term and a long-term test-retest interval sample) to study LMI hypotheses (e.g., the effect of remembering items on LMI). However, we did not apply this approach because, again, the diminished sample size may result in data that is prone to nonconvergence, improper factor solutions, large standard errors, biased estimates of factor loadings and thresholds, and problematic goodness-of-fit tests (Liu et al., 2017).

Finally, the order of administered questionnaires at the pretest may have influenced the degree of LMI. This measurement was, in contrast to the posttest, preceded by several other questionnaires depending on a patient's disorder and age. Consequently, patients may have responded differently to items than they would have done when the PROMIS measurements were administered first (e.g., because of tiredness, or context effects; Windle, 1954).

In this study, we evaluated LMI in the Dutch-Flemish PROMIS item banks for Depression and Anxiety. Using tentatively determined rules of thumb, the results suggest that, even though some statistically significant violations of LMI were found, the item banks provide sufficiently invariant latent factor scores for use in clinical practice. This conclusion is often assumed for other (PROMIS) measurements. By assuming sufficient LMI, however, test users may have to deal with biased change scores without being aware of it. We therefore urge other researchers to study LMI in their own measurements.

Chapter 6

Construct validity, Responsiveness, and Utility of Change Indicators of the Dutch-Flemish PROMIS Item banks for Depression and Anxiety Administered as Computerized Adaptive Test (CAT): A Comparison with the Brief Symptom Inventory (BSI).

Published as:

Flens, G., Terwee, C. B., Smits, N., Williams, G., Spinhoven, P., Roorda, L. D., & de Beurs, E. (2022). Construct validity, responsiveness, and utility of change indicators of the Dutch-Flemish PROMIS item banks for depression and anxiety administered as computerized adaptive test (CAT): A comparison with the Brief Symptom Inventory (BSI). *Psychological Assessment*, *34*(1), 58-69.

6.1 Abstract

We evaluated construct validity, responsiveness, and utility of change indicators of the Dutch-Flemish PROMIS adult v1.0 item banks for Depression and Anxiety administered as computerized adaptive test (CAT). Specifically, the CATs were compared to the Brief Symptom Inventory (BSI) using pretest and retest data of adult patients treated for common mental disorders (*N* = 400; median pretest to retest interval = 215 days). Construct validity was evaluated with Pearson's correlations and Cohen's *d*s; responsiveness with Pearson's correlations and pre-post effect sizes (*ES*); utility of change indicators with kappa coefficients and percentages of (dis)agreement. The results showed that the PROMIS CATs measure similar constructs as matching BSI scales. Under the assumption of measuring similar constructs, the CAT and BSI Depression scales were similarly responsive. For the Anxiety scales, we found a higher responsiveness for CAT (*ES* = 0.64) compared to the BSI (*ES* = 0.50). Finally, both CATs categorized the change scores of more patients as changed compared to matching BSI scales, indicating that the PROMIS CATs may be more able to detect actual change than the BSI. Based on these findings, the PROMIS CATs may be considered a modest improvement over matching BSI scales as tools for reviewing treatment progress with patients. We discuss several additional differences between the PROMIS CATs and the BSI to help test users choose instruments. These differences include the adopted measurement theory (Item Response Theory vs. Classical Test Theory), the mode of administration (CAT vs. fixed items), and the area of application (universal vs. predominantly clinical).

Keywords: clinical assessment, depression, anxiety, PROMIS CAT, psychometric properties

6.2 Background

In Dutch health care, computerized adaptive tests (CATs) are gradually being implemented to evaluate self-reported health in clinical subjects (e.g., depression, physical function, and ability to participate in social roles and activities; Terwee et al., 2014). A CAT is a computer-based test in which items are administered from an item bank (i.e., a set of items that measure a specific construct) according to the answers to previous selected items, and that terminates when a stopping rule is met (e.g., a specific measurement precision). As a result, patient burden can be reduced with a shorter measurement and a negligible loss of precision (Fliege et al., 2005).

The first item banks that were psychometrically evaluated for CAT administration in Dutch *mental* health care were the Patient-Reported Outcomes Measurement Information System (PROMIS$^{®}$) adult v1.0 item banks for Depression and Anxiety. In previous studies, these item banks were translated into Dutch-Flemish (DF; Terwee et al., 2014) and psychometrically evaluated for cross-sectional (Flens et al., 2017, 2019) and longitudinal applications (Flens et al., 2021). The cross-sectional studies showed that both item banks have good quality item parameters according to the PROMIS standards (Reeve et al., 2007). Moreover, post hoc CAT simulations showed that both item banks, when administered adaptively, can be highly precise as well as efficient in both the general population and clinical ambulatory populations at the start of treatment. In addition, the longitudinal study showed that, using tentative rules of thumb, the Depression and Anxiety item banks were sufficiently invariant over time in clinical samples with mood and anxiety disorders, respectively. In other words, the item banks appear to provide (change) scores that reflect single depression and anxiety constructs.

The results of these earlier studies indicate that the DF PROMIS adult v1.0 item banks for Depression and Anxiety have adequate psychometric properties for both cross-sectional and longitudinal applications. However, the item banks still need to be validated with actual CAT administrations, and compared to an established Dutch legacy instrument before introducing them in routine assessment of clinical subjects. After all, we want to ensure that the psychometric properties of the PROMIS CATs are at least as good as those of legacy instruments to convince users that changing instruments results in similar (and preferably even better) assessment of patients.

Psychometric properties that demand additional attention are construct validity and responsiveness (Maruyama & Ryan, 2014; Mokkink et al., 2010; Pilkonis et al., 2014). Furthermore, the utility of reliability-based indicators of clinical significant change need to be evaluated to facilitate the use of the PROMIS CATs in clinical practice (Jacobson & Truax, 1991). These aspects are seen as relevant because they reflect an instrument’s ability to aid professionals in planning treatments, evaluating therapeutic interventions, and anticipating and planning timely termination (de Beurs et al., 2018). Furthermore, regular or continuous monitoring of progress with appropriate and psychometrically sound instruments may help to prevent treatment failure (Lambert, 2010).

In previous studies, using clinical samples, it was demonstrated that the Unites States (US) PROMIS instruments for Depression and Anxiety (i.e., CATs and short-forms) measure similar constructs as legacy instruments, and are similarly responsive (Kroenke et al., 2019; Pilkonis et al., 2014). These results were shown for the PROMIS Depression instruments compared to the legacy instruments Center for Epidemiological Studies Depression scale (CESD) and Patient Health Questionnaire (PHQ-9), and for the PROMIS Anxiety instruments compared to the legacy instruments Generalized Anxiety Disorder (GAD-7), Symptom Checklist (SCL), Posttraumatic Stress disorder checklist (PCL), Short Form (SF)-36, and SF-12 Mental Component Summary (MCS). We therefore expect that the DF PROMIS CATs for Depression and Anxiety also measure similar constructs as Dutch legacy instruments, and are at least as responsive. In addition, Pilkonis et al. (2014) showed that the US PROMIS CAT for Depression measures more reliably than the legacy instruments CESD and PHQ-9, probably because CAT ensures that each administration meets the minimally required measurement precision, by which the number of administered items is allowed to vary among respondents. The legacy instruments, on the other hand, fix the number of items, by which the measurement precision will vary among respondents. Based on these measurement properties, we expect that reliability-based indicators of clinical significant change categorize more patients as actually changed for the DF PROMIS CATs compared to fixed-item legacy instruments.

This study was the first in the Netherlands in which PROMIS CATs were administered. We aimed to assess construct validity, responsiveness, and utility of change indicators of the DF PROMIS adult v1.0 item banks for Depression and Anxiety administered as CAT in a clinical sample. Specifically, the PROMIS CATs were compared to the nine subscales of the Dutch Brief Symptom Inventory (BSI; de Beurs & Zitman, 2005; Derogatis et al., 1973) using pretest and retest data. We chose the BSI as legacy instrument because two of its subscales aim to measure the same constructs as the CATs; it is often used as outcome measure in routine assessment of patients internationally and in the Netherlands; and it has been claimed to have adequate psychometric properties for both cross-sectional and longitudinal applications (Carlier et al., 2017; de Beurs & Zitman, 2005; van Noorden et al., 2010). More specifically, it has been demonstrated that the BSI is sufficiently reliable, valid and responsive compared to a large number of legacy instruments. These include the Symptom Questionnaire-48 (SQ-48), the Outcome Questionnaire-45 (OQ-45), and several disorder-specific instruments (e.g., the Montgomery Äsberg Depression Rating Scale [MADRS], Beck Depression Inventory [BDI-II], Padua Inventory [PI], Yale Brown Obsessive Compulsive Scale [Y-BOCS], and Panic Disorder Severity Scale [PDSS].

6.3 Methods

6.3.1 Participants

Data were collected between September 2017 and June 2019 in a clinical population of adult patients who started outpatient treatment for common mental disorders. Patients were invited by the Dimence Group, which is a large mental health institute offering inpatient and outpatient treatment in the eastern part of the Netherlands. The patient’s diagnosis (*Diagnostic and*

Statistical Manual of Mental Disorders, 5th ed.; *DSM-5*; American Psychiatric Association, 2013) was assessed by a therapist in a clinical face-to-face assessment (i.e., the intake of treatment).

This study has not been submitted to a research ethics committee because, according to Dutch law, data collected as part of clinical practice may be used in anonymized form for scientific research (de Beurs et al., 2011). Consequently, all data were coded before they were released to the first author for analysis, and could not be traced back to a person by the authors. This was approved by the privacy and information security officer of the Dimence Group. In addition, patients were informed upon their referral and registration for treatment that their data might be used for research, and that an opt-out procedure was available if they did not consent to this. Data from patients with an opt-out registration were not released to the first author.

In accordance with similar studies, we aimed to include at least 200 patients (Pilkonis et al., 2014; Schalet et al., 2016). A patient was included when (a) a pretest and retest score were available for all measures to perform the analyses in a straightforward fashion (i.e., without missing cases), (b) the measures were completed on the same day for both pretest and retest to establish a set of instruments that was administered under similar conditions as much as possible, (c) the retest was administered at least one month after the pretest to increase the possibility that at least some change had occurred between measurements, and (d) the retest was administered after the first treatment session to ensure that at least some treatment was provided.

6.3.2 Measures

The measures were part of a larger battery of instruments to be completed by the patients, and consisted of the DF PROMIS adult v1.0 item banks for Depression (Flens et al., 2017) and Anxiety (Flens et al., 2019) administered as CAT, and the Dutch BSI (de Beurs & Zitman, 2005). For each patient, the measures were administered digitally through an automated process. In this process, the PROMIS CATs were assigned in alternating order for both pretest and retest: the CAT Anxiety was administered first at even weeks, the CAT Depression was administered first at odd weeks. The BSI was always administered directly after the PROMIS CATs. According to Dimence Group’s policy, the invitation for the pretest was sent before or during the intake session.

6.3.2.1 PROMIS CATs

The content of the DF PROMIS adult v1.0 item banks for Depression and Anxiety item banks reflects a wide range of depression and anxiety symptoms, problems, or negative affective states (e.g., Depression item bank, EDDEP04 *I felt worthless*; Anxiety item bank, EDANX01 *I felt fearful*). Respondents were asked by computer to indicate on a 5-point scale how frequently they experienced the symptoms, problems or negative states in the past 7 days (1 = *never*, 2 = *rarely*, 3 = *sometimes*, 4 = *often*, and 5 = *always*), a higher score indicating more severe depression or anxiety.

For the CAT item selection algorithm, we followed the recommendations of other studies (Flens et al., 2017, 2019), using Fisher's information function calculated with the DF item parameters. The initial item was selected as the item with the greatest Fisher's information at the value of the estimated mean (*M*) of the latent trait for the Dutch general population. For the CAT Depression this item was EDDEP36 *I felt unhappy*; for the CAT Anxiety EDANX54 *I felt tense* was selected. After each item, the maximum likelihood estimate (MLE) of the test taker's latent trait was calculated. Each sequential item was then selected as the item with the greatest Fisher's information at the value of the MLE. The CAT was terminated when either the measurement precision fell below a predefined threshold or the upper limit of administered items was reached. The measurement precision threshold was set to a *SE*(θ) below .22, with the *SE*(θ) approximated as the reciprocal of the square root of the information function. The threshold of .22 was selected to be comparable to a marginal reliability of .95 (Green et al., 1984), which is considered a high standard for the precision of assessments that provide scores to individuals (Bernstein & Nunnally, 1994). The upper limit of administered items was set to nine for the CAT Depression, and to 12 for the CAT Anxiety (Flens et al., 2019).

According to PROMIS convention, we used the item responses, the US item parameters, and the expected a posteriori (EAP) estimator to calculate PROMIS T-scores and their accompanying measurement precision through the HealthMeasures provided Scoring Service, powered by Assessment Center (www.assessmentcenter.net/ac_scoringservice). PROMIS T-scores are represented on a scale with a *M* of 50 and a standard deviation (*SD*) of 10 in the general US population, a higher score indicating more severe depression or anxiety.

6.3.2.2 BSI

The BSI is a multidimensional self-report instrument that evaluates the severity of symptoms of psychopathology. The instrument consists of an 53-item overall scale, divided into nine subscales (i.e., Depression, 6 items; Anxiety, 6 items; Somatization, 7 items; Obsessive-Compulsive, 6 items; Interpersonal Sensitivity, 4 items; Hostility, 5 items; Phobic Anxiety, 5 items; Paranoid Ideation, 5 items; Psychoticism; 5 items) and four remaining items. For this study, we used the Depression subscale (e.g., item 18 *feeling no interest in things*) and the Anxiety subscale (e.g., item 38 *feeling tense or keyed up*) to evaluate the CATs on their relation with scales measuring matching constructs. The other subscales were used to evaluate the CATs on their relation with scales measuring other constructs. For all subscales, respondents were asked by computer to indicate on a 5-point scale to what extent they were bothered by the symptoms, problems or negative states in the past 7 days (0 = *not at all*, 1 = *a little bit*, 2 = *moderately*, 3 = *quite a bit*, and 4 = *extremely*). Average scores were calculated for each subscale (ranging from 0 – 4), a higher score meaning more distress.

6.3.3 Statistical analyses

We performed analyses to report on descriptive statistics, construct validity, responsiveness, and utility of change indicators. A hypothesis was formulated for each analysis to compare the instruments. As rule of thumb, we considered a psychometric property as sufficiently supported when at least 75% of the hypotheses were confirmed (Prinsen et al., 2018). For indicators of change between pretest and retest scores, we did not correct for pretest severity (O'Connell et al., 2017). All statistical analyses were performed in the statistical environment R (R Core Team, 2018).

6.3.3.1 Descriptive statistics

Based on the inclusion criteria, we assessed the gender- and age distribution of the study sample. Furthermore, we evaluated whether the composition of the study sample was representative for the mental health provider that collected the data. To accomplish this, it was assessed whether the included patients were similar to the nonincluded patients regarding the distribution of gender, age, and pretest score. For gender, we investigated Pearson's residuals as measure of effect size, following the suggestion of 2.00 as cutoff value for indicating a substantial difference between the observed respondents and the expected number of respondents under the model (Agresti & Kateri, 2011). For age and pretest score, we investigated Cohen's *d* as measure of effect size (i.e., the *M* difference divided by the pooled *SD*), following the guideline proposed by Cohen (1988) to interpret the size of the effect: 0.20 = small effect, 0.50 = medium effect, and 0.80 = large effect.

In addition, we assessed the mean number of administered items for both pretest and retest of the Depression and Anxiety scales. Furthermore, we evaluated the variation in pretest to retest interval by calculating quantiles of the days between the tests.

6.3.3.2 Construct validity

A classic definition of construct validity is the degree to which a test measures the concept it is supposed to measure (Cook & Campbell, 1979). We investigated this psychometric property by collecting multiple sources of empirical evidence commonly claimed as indicative of validity (Newton & Shaw, 2014).

First, we studied convergent and divergent validity by evaluating whether the measured constructs of the PROMIS CATs are related to those of matching BSI scales, and unrelated to those of other BSI scales (Cook & Campbell, 1959). For convergent validity, it was hypothesized that Pearson's correlation coefficients between the CATs and matching BSI scales were above 0.50 (Prinsen et al., 2018) for both pretest (*Hypothesis 1*) and retest (*Hypothesis 2*). For divergent validity, it was hypothesized for both pretest (*Hypothesis 3*) and retest (*Hypothesis 4*) that Pearson's correlation coefficients between the CATs and other BSI scales were at least 0.10 points lower than those between the CATs and matching BSI scales (Prinsen et al., 2018).

Next, we studied concurrent validity by evaluating whether the PROMIS CATs are at least as able as matching BSI scales to distinguish between distinct groups based on the patient's primary diagnosis (i.e., the condition that causes the patient the most problems or discomfort, as assessed at the intake of treatment; American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1974). Consequently, the study sample was divided into patients with and without a primary depression diagnosis to compare the Depression scales, and into patients with and without a primary anxiety diagnosis to compare the Anxiety scales. We then compared Cohen's *d* measure of effect size with a 95% confidence interval (*CI*) between the CATs and matching BSI scales (Hedges & Olkin, 2014). Cohen's *d* was calculated as the *M* score difference between patients with and without a primary diagnosis divided by the pooled *SD* of these subsamples. We suggest that a difference in *d*-values of at most 0.10 points indicates sufficient similarity in the ability to discriminate between patients with and without a specific disorder. Consequently, it was hypothesized that the *d*-values of the CATs were at most 0.10 points lower than those of matching BSI scales (*Hypothesis 5*). This was evaluated for the pretest only because the primary diagnosis was assessed around this test.

Finally, we studied stability by evaluating whether the pretest to retest associations of the PROMIS CATs are sufficiently similar to those of matching BSI scales (Drenth & Sijtsma, 2005). To study stability, we suggest that a difference in Pearson's pretest to retest correlation coefficients of at most 0.10 points indicates sufficient similarity in stability. Consequently, it was hypothesized that the pretest to retest correlation coefficients of the CATs differed at most 0.10 points from those of matching BSI scales (*Hypothesis 6*).

6.3.3.3 Responsiveness

Responsiveness is defined by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) as the ability of an instrument to detect change over time in the construct to be measured (Mokkink et al., 2010). To study this psychometric property, we evaluated whether the change scores of the PROMIS CATs are related to those of matching BSI scales, and unrelated to those of other BSI scales. It was hypothesized that Pearson's correlation coefficients between the change scores of the CATs and matching BSI scales were above 0.50 (*Hypothesis 7*; Prinsen et al., 2018). Furthermore, it was hypothesized that Pearson's correlation coefficients between the change scores of the CATs and other BSI scales were at least 0.10 points lower than those between the CATs and matching BSI scales (*Hypothesis 8*; Prinsen et al., 2018).

In addition, pre-post effect sizes (*ES*) with a 95% *CI* were used to evaluate whether the PROMIS CATs are at least as responsive as matching BSI scales (Seidel et al., 2014). *ES* is calculated as the *M* change score of a scale divided by the *SD* of that scale's pretest scores. We suggest that a difference in *ES* values of at most 0.10 points indicates sufficient similarity in responsiveness. Consequently, it was hypothesized that the *ES* values of the CATs were at most 0.10 points lower than those of matching BSI scales (*Hypothesis 9*).

6.3.3.4 Utility of change indicators

To evaluate whether patients improve or deteriorate, often-used indicators are reliable change and clinically significant change (CSC; Jacobson & Truax, 1991). Reliable change is defined as a change in scores that may not have occurred due to random measurement error alone. *CSC* is defined as a change from a clinical population score to a general population score. We combined reliable change and *CSC* to evaluate whether the PROMIS CATs categorize more patients as actually changed than matching BSI scales (de Beurs et al., 2019).

Reliable change was evaluated with the *Z*-test for the CATs (Brouwer et al., 2013) and with the reliable change index (*RCI*) for the BSI scales (Jacobson & Truax, 1991). Different methods were used because the CATs and the BSI assume different measurement theories (i.e., item response theory [IRT] and classical test theory [CTT], respectively). To assess reliable change, we used the SEs of a patient's pretest and retest for the *Z*-test, and the test-retest reliability as determined by de Beurs and Zitman (2005) for the *RCI*. A *Z*- or *RCI* value larger than 1.96 reflects with a 95% *CI* that the change in pretest to retest scores may not have occurred due to random measurement error alone (Brouwer et al., 2013).

The cutoff for *CSC* was calculated as the point halfway the general- and clinical population. To determine this value for each Depression and Anxiety scale, we used the samples of previous psychometric studies for the general population (CAT Depression, Flens et al., 2017; CAT Anxiety, Flens et al., 2019; BSI, de Beurs & Zitman, 2005), and the pretest sample of the current study for the clinical population. Subsequently, we used the cutoff values for both *CSC* and reliable change to categorize the patients into four groups: recovered (CATs, $Z \geq 1.96$; BSI, $RCI \geq 1.96$; pretest score > *CSC*; retest score ≤ *CSC*), improved (CATs, $Z \geq 1.96$; BSI, $RCI \geq 1.96$), unchanged (CATs, $-1.96 \leq Z < 1.96$; BSI, $-1.96 \leq RCI < 1.96$), and deteriorated (CATs, Z < -1.96; BSI, $RCI < -1.96$).

We used the modified Fleiss kappa statistic for ordinal variables (i.e., the s^* statistic) with linear weights and a 95% *CI* (Marasini et al., 2016) as well as the percentage of agreement to assess whether the PROMIS CATs showed a substantial disagreement with matching BSI scales in categorizations, and, if so, whether the CATs categorized the change scores of less patients as unchanged. We considered this to be the case when three criteria where met: the s^* statistic was smaller than 0.60 (McHugh, 2012), the percentage of agreement was smaller than 80% (McHugh, 2012), and the percentage of patients that were categorized as unchanged was smaller for the CATs than for matching BSI scales (*Hypothesis 10*).

The data of this study are not publicly available because they were used under license from the Dimence Group. However, the data can be made available from the first author upon reasonable request and with permission of the Dimence Group. The study analysis code can be requested from the first author. This study was not preregistered.

6.4 Results

6.4.1 Descriptive statistics

The eligible sample (i.e., the patients that were invited for the pretest and retest) consisted of 549 respondents. Of these respondents, 544 responded to the pretest (response rate = 99.1%) and 504 also responded to the retest (response rate = 91.8%). Furthermore, we excluded 104 respondents for not meeting the remaining inclusion criteria. Consequently, our final sample consisted of $N = 400$ (response rate = 72.9%; 64.0% female; age $M = 37.4$ years, $SD = 12.2$, range 18–66). For this sample, 46% of the patients had a mood disorder as the primary reason for seeking treatment, 39% had an anxiety disorder, and 15% had another disorder (e.g., attention deficit disorder, somatoform disorder, personality disorder). In addition, the pretest and retest did not include missing item responses. Consequently, the analyses were performed in a straightforward fashion.

Next, the comparison between the included and nonincluded patients showed that Pearson's residuals were all below 2.00 for gender, and Cohen's *d*s were all below 0.20 for age and pretest score. These results indicate that the included patients were sufficiently similar to the nonincluded patients for the variables gender, age, and pretest score.

Finally, the *M* (*SD*) number of administered items was 6.7 (1.0) for the CAT Depression pretest (7% responded to all 9 items), 6.6 (1.0) for the CAT Depression retest (9% responded to all 9 items), 8.7 (1.1) for the CAT Anxiety pretest (3% responded to all 12 items), and 8.5 (1.1) for the CAT Anxiety retest (3% responded to all 12 items). For the BSI, all patients responded to the six fixed items of the Depression scale and the six fixed items of the Anxiety scale. In addition, the median of the pretest to retest interval was 215 days (range = 32–505, interquartile range = 145–281), indicating a substantial variation in intervals.

6.4.2 Construct validity

Table 6.1 displays Pearson's correlation coefficients between the PROMIS CATs and BSI subscales for the pretest and retest. In support of *Hypothesis 1* (pretest) and *Hypothesis* 2 (retest), the correlation coefficients between the CATs and matching BSI scales were above .50 (Depression, pretest *r* = .83, retest *r* = .87; Anxiety, pretest *r* = .76, retest *r* = .81). Furthermore, in support of *Hypothesis 3* (pretest) and *Hypothesis 4* (retest), the correlation coefficients between the CATs and other BSI scales were at least 0.10 points below those between the CATs and matching BSI scales. Note that the CATs correlated lower with most of the other BSI scales than the BSI Depression and Anxiety scales did.

In support of *Hypothesis 5*, the *d*-values between patients with and without a specific primary diagnosis (i.e., depression or anxiety) were not more than 0.10 points lower for the PROMIS CATs relative to those between matching BSI scales. The comparison between the depression (*n* = 184; CAT *M* = 65.7, *SD* = 6.4; BSI *M* = 2.21, *SD* = 0.98) and not-depression subsamples (*n* = 216; CAT *M* = 63.5, *SD* = 7.0; BSI *M* = 1.95, *SD* = 0.96) resulted in Cohen's *d* = 0.33, 95% *CI* [0.13, 0.53] for the CAT Depression, and *d* = 0.26, 95% *CI* [0.06, 0.46] for the BSI Depression scale. The comparison between the anxiety (*n* = 157; CAT *M* = 68.1, *SD* = 6.0; BSI *M* = 1.97, *SD* = 0.93) and not-anxiety subsamples (*n* = 243; CAT *M* = 66.6, *SD* = 6.6; BSI *M* = 1.66, *SD* = 0.91) resulted in *d* = 0.24, 95% *CI* [0.04, 0.44] for the CAT Anxiety and *d* = 0.33, 95% *CI* [0.13, 0.53] for the BSI Anxiety scale. Note that although *Hypothesis 5* was supported for both CATs, Cohen's *d* suggested that the CAT Anxiety was somewhat less able than the BSI Anxiety scale to discriminate between patients with and without a primary anxiety diagnosis. For the Depression scales, however, we found the opposite: the CAT Depression was somewhat better able than the BSI Depression scale to distinguish between patients with and without a primary depression diagnosis.

In support of *Hypothesis 6*, Pearson's pretest to retest correlation coefficients differed less than 0.10 points between the CAT Depression (*r* = 0.54) and BSI Depression scales (*r* = 0.53). For the Anxiety scales, however, *Hypothesis 6* was rejected because the correlation coefficient for CAT (*r* = 0.40) was more than 0.10 points lower than that for the BSI (*r* = 0.56).

Overall, *Hypotheses 1-6* were supported for the CAT Depression. For the CAT Anxiety, *Hypotheses 1-5* were supported and *Hypothesis 6* was rejected. Consequently, construct validity was considered sufficient for both PROMIS CATs as more than 75% of the hypotheses were supported.

Table 6.1 Pearson's correlation coefficients between the PROMIS CATs and BSI subscales for the pretest and retest scores.

		Dep pre		Dep re		Anx pre		Anx re	
Scale	Instrument	CAT	BSI	CAT	BSI	CAT	BSI	CAT	BSI
Dep pre	CAT	1.00							
	BSI	**0.83**	1.00						
Dep re	CAT	0.54	0.46	1.00					
	BSI	0.51	0.53	**0.87**	1.00				
Anx pre	CAT	0.66	0.58	0.33	0.31	1.00			
	BSI	0.48	0.55	0.27	0.30	**0.76**	1.00		
Anx re	CAT	0.42	0.34	0.78	0.71	0.40	0.38	1.00	
	BSI	0.38	0.37	0.64	0.69	0.45	0.56	**0.81**	1.00
Som pre	BSI	**0.44**	0.48	0.32	0.32	**0.53**	0.63	0.33	0.45
Som re	BSI	0.35	0.35	**0.55**	0.59	0.37	0.41	**0.59**	0.70
Obs pre	BSI	**0.56**	0.66	0.39	0.42	**0.60**	0.62	0.37	0.46
Obs re	BSI	0.42	0.42	**0.64**	0.71	0.33	0.35	**0.67**	0.73
Hos pre	BSI	**0.39**	0.45	0.28	0.32	**0.38**	0.43	0.28	0.37
Hos re	BSI	0.28	0.27	**0.47**	0.53	0.21	0.25	**0.50**	0.56
Pho pre	BSI	**0.48**	0.53	0.37	0.38	**0.58**	0.66	0.41	0.50
Pho re	BSI	0.37	0.36	**0.60**	0.64	0.41	0.45	**0.66**	0.77
Par pre	BSI	**0.46**	0.52	0.28	0.32	**0.44**	0.47	0.28	0.37
Par re	BSI	0.34	0.37	**0.53**	0.59	0.27	0.29	**0.54**	0.60
Psy pre	BSI	**0.62**	0.73	0.37	0.43	**0.56**	0.56	0.32	0.39
Psy re	BSI	0.45	0.49	**0.72**	0.83	0.33	0.34	**0.68**	0.71
Int pre	BSI	**0.51**	0.62	0.34	0.41	**0.47**	0.52	0.31	0.42
Int re	BSI	0.36	0.38	**0.61**	0.69	0.30	0.32	**0.62**	0.70

Note. pre = pretest; re = retest; Dep = depression; Anx = anxiety; Som = somatization; Obs = obsessive-compulsive; Hos = hostility; Pho = phobic anxiety; Par = paranoid ideation; Psy = psychoticism; Int = interpersonal sensitivity; all correlations deviate statistically significantly from zero; correlations used to assess construct validity are presented bold-faced.

6.4.3 Responsiveness

Table 6.2 displays Pearson's correlation coefficients between the change scores of the PROMIS CATs and BSI subscales. In support of *Hypothesis 7*, the correlation coefficients between the CATs and matching BSI scales were above 0.50 for both Depression (r = .78) and Anxiety scales (r = .72). Furthermore, in support of *Hypothesis 8*, the correlation coefficients between the CATs and other BSI scales were at least 0.10 points below those between the CATs and matching BSI scales. Note that, similarly to the pretest and retest scores, the CATs correlated lower with the other BSI scales than the BSI Depression and Anxiety scales did.

Table 6.2 Pearson's correlation coefficients between the change scores of the PROMIS CATs and BSI subscales.

	Depression		Anxiety	
Scale	CAT	BSI	CAT	BSI
CAT Dep	1,00			
BSI Dep	**0,78**	1,00		
CAT Anx	0,67	0,61	1,00	
BSI Anx	0,55	0,64	**0,72**	1,00
BSI Som	**0,37**	0,49	**0,45**	0,59
BSI Obs	**0,46**	0,63	**0,60**	0,66
BSI Hos	**0,35**	0,46	**0,41**	0,44
BSI Pho	**0,44**	0,55	**0,48**	0,64
BSI Par	**0,46**	0,52	**0,47**	0,53
BSI Psy	**0,59**	0,71	**0,60**	0,62
BSI Int	**0,51**	0,64	**0,54**	0,60

Note. Dep = depression; Anx = anxiety; Som = somatization; Obs = obsessive-compulsive; Hos = hostility; Pho = phobic Anxiety; Par = paranoid Ideation; Psy = psychoticism; Int = interpersonal sensitivity; all correlations deviate statistically significantly from zero; correlations used to assess responsiveness are bold faced.

In support of *Hypothesis 9*, the *ES* value for the CAT Depression (pretest, *M* = 64.5, *SD* = 6.8; retest, *M* = 60.8, *SD* = 8.1; *ES* = 0.55, 95% *CI* [0.41 – 0.69]) was not more than 0.10 points lower than that for the BSI Depression scale (pretest, *M* = 2.07, *SD* = 0.98; retest, *M* = 1.54, *SD* = 1.06; *ES* = 0.54, 95% *CI* [0.40 – 0.68]). The *ES* value for the CAT Anxiety (pretest, *M* = 67.2, *SD* = 6.4; retest, *M* = 63.1, *SD* = 7.5; *ES* = 0.64, 95% *CI* [0.50 – 0.78]) was more than 0.10 points *higher* than that for the BSI Anxiety scale (pretest, *M* = 1.79, *SD* = 0.93; retest, *M* = 1.32, *SD* = 0.92; *ES* = 0.50, 95% *CI* [0.36 – 0.64]), which was also in support of *Hypothesis 9*.

Overall, *Hypotheses 7–9* were supported for both PROMIS CATs, indicating sufficient responsiveness. Under the assumption of measuring similar constructs, the CAT Anxiety even showed a higher responsiveness than the BSI Anxiety scale.

6.4.4 Utility of change indicators

Table 6.3 displays the percentages of (dis)agreement between the PROMIS CATs and matching BSI subscales for the four categories based on reliable change and *CSC*. In support of *Hypothesis 10*, the *s** statistic was lower than 0.60 for both CATs (Depression, *s** = 0.53, 95% *CI* [0.46 – 0.59]; Anxiety, *s** = 0.50, 95% *CI* [0.36 – 0.64]), the percentage of agreement was lower than 80% for both CATs (Depression, 11 + 6 + 54 + 1 = 72%; Anxiety, 11 + 3 + 52 + 1 = 67%), and less patients were categorized as unchanged by the CATs (Depression, 3 + 2 + 54 + 3 = 62%; Anxiety, 1 + 3 + 52 + 2 = 58%) relative to the BSI scales (Depression, 5 + 6 + 54 + 4 = 69%; Anxiety, 10 + 8 + 52 + 6 = 76%). These results suggest that, under the assumption of measuring similar constructs, change categorizations of the PROMIS CATs are substantially different from those of matching BSI scales, and the PROMIS CATs categorize more patients as actually changed. Note that the difference between the CATs and BSI scales in the percentage of unchanged patients was larger for Anxiety than for Depression.

Table 6.3 Percentages of (dis)agreement between the PROMIS CATs and matching BSI subscales on the categories based on reliable change and CSC.

	CAT Depression				CAT Anxiety			
BSI	Recovered	Improved	Unchanged	Deteriorated	Recovered	Improved	Unchanged	Deteriorated
Recovered	11%	4%	3%	0%	11%	2%	1%	0%
Improved	1%	6%	2%	0%	2%	3%	3%	0%
Unchanged	5%	6%	54%	4%	10%	8%	52%	6%
Deteriorated	0%	0%	3%	1%	0%	0%	2%	1%

Note. The percentages add up to 101% for the Anxiety scales due to rounding.

6.5 Discussion

This was the first study in the Netherlands in which PROMIS CATs were administered. We evaluated construct validity, responsiveness, and utility of change indicators of the DF PROMIS adult v1.0 item banks for Depression and Anxiety administered as CATs in a clinical sample, by comparing them with the Dutch BSI subscales. In line with other studies that used different legacy instruments (Kroenke et al., 2019; Pilkonis et al., 2014), we found that both PROMIS CATs showed sufficient construct validity, responsiveness, and utility of change indicators. More specifically, we found that the CATs measured similar constructs as matching BSI scales. Under the assumption of measuring similar constructs, the CAT Depression also showed a similar responsiveness relative to the BSI Depression scale. For the CAT Anxiety, we even found a higher responsiveness compared to the BSI Anxiety scale, which may suggest that the CAT Anxiety is more able to detect change. Finally, both CATs showed a substantial disagreement with matching BSI scales in change categorizations; the CATs categorized the change scores of more patients as changed, which may suggest that the CATs are more able to detect actual change. Based on these findings, the PROMIS CATs may be considered an improvement over matching BSI scales as tools for reviewing treatment progress with patients.

The findings of this study are based on the assumption that the BSI is an adequate comparator for the PROMIS CATs. It should be noted, however, that comparison instruments always differ to some extent, possibly due to differences in the methods used for test construction. It has been shown that different methods may amount to very different compilations of aspects on which a test performs well (Oosterveld et al., 2019). Identifying differences between the PROMIS CATs and the BSI may therefore help to explain some of the results in this study. First, the instruments differ in their underlying measurement theory and administration method. The PROMIS CATs were developed under an IRT model (Embretson & Reise, 2000), and use item banks to select and administer items that can differ between respondents and measurement occasions. The BSI was developed under the CTT model (Lord

& Novick, 1968), and uses a fixed number of items for all respondents and measurement occasions. Second, the PROMIS CATs provide a measurement error estimate for each individual test taker while the BSI scales only provide a single estimate of the standard error of measurement for all test takers. Third, the PROMIS CATs use response categories based on frequency (*never* to *always*) while the BSI uses response categories based on severity (*not at all* to *extremely*). Fourth, the PROMIS CATs use norm-based interval T-scores based on the US general population and the EAP estimator (Cella et al., 2010) while the BSI uses ordinal Dutch raw average scores. This means that the PROMIS CATs use prior information (i.e., the standard normal distribution) and the reliability of the test to improve the estimated score, whereas the BSI uses average scores without consideration of prior information (Bock, 1997). Finally, the PROMIS CATs were primarily developed for universal application in different populations, whereas the BSI was primarily developed for clinical populations. In case of the BSI Anxiety scale, the main focus was even more specific: patients with high anxiety levels (Derogatis et al., 1973).

In the next paragraphs, we provide possible explanations for the results based on the differences between the PROMIS CATs and the BSI, and the design of this study. We start with the results that stood out most regarding our hypotheses: the lower pretest to retest stability for the CAT Anxiety which led to the rejection of *Hypothesis 6*, and the higher responsiveness of the CAT Anxiety while we expected a similar responsiveness (Kroenke et al., 2019). Actually, both findings are in fact associated because stability is the opposite of change. To clarify this, consider that a scale’s degree of stability is related to the variation in change scores of that scale: perfect stability results from all change scores being equal while low stability results from a large variation in change scores. The degree of variation in change scores may in turn be related to the degree of responsiveness. After all, higher responsiveness enables more space to be used on the scale, which can result in a larger variation of change scores. We therefore suspected that the CAT Anxiety showed a larger variation in change scores than the BSI Anxiety scale, which was confirmed by an additional analysis using *Z*-scores for both scales (not shown herein): the *SD* of the change scores was 1.20 for the CAT Anxiety and 0.94 for the BSI Anxiety scale. Consequently, it may be that the lower stability of the CAT Anxiety was to be expected, assuming a higher responsiveness.

We found two possible explanations for the unexpected finding of the CAT Anxiety having a higher responsiveness than the BSI Anxiety scale. First, the choice of item parameters influenced the results. We concluded this by recalculating the T-scores with DF item parameters (Flens et al., 2017, 2019) and re-evaluating responsiveness. The results (not shown herein) indicated that the pre-post effect size for the CAT Anxiety was somewhat smaller for DF item parameters (*ES* = 0.58) compared to US item parameters (*ES* = 0.64). Thus, would we have used the DF item parameters to calculate the T-scores, we would not have concluded that the CAT Anxiety was more responsive than the BSI Anxiety scale, but instead that they were similarly responsive. This difference was to some extent a consequence of the numerator in the *ES* formula (i.e., the *M* pretest T-score minus the *M* retest T-score; DF parameters = 4.00; US parameters = 4.06), but especially of the denominator (i.e., the *SD* of the pretest T-scores; DF parameters = 6.85; US parameters = 6.39). Apparently, DF item parameters yield a somewhat more conservative estimation of *ES* due to the larger range in pretest scores. This finding is relevant for the discussion regarding the choice of appropriate item parameters (i.e., US parameters, country-specific parameters, or international parameters; Elsman et al., 2022; Terwee et al., 2021; van Bebber et al., 2018).

Second, the degree of longitudinal measurement invariance (LMI) may have influenced the degree of responsiveness. A set of items is said to show sufficient LMI when it measures one or more constructs in the same way over time. This means that changes in respondents' scores over time can entirely be attributed to changes *within* the construct(s) measured by the set of items (Fried et al., 2016; Liu et al., 2017). A previous study using full item bank data of Dutch patients with mood and anxiety disorders showed that the degree of LMI was sufficient in both PROMIS Depression and Anxiety item banks, but also that it was somewhat smaller in the PROMIS Anxiety item bank (Flens et al., 2021). Similarly, the degree of LMI may differ between the PROMIS CATs and matching BSI scales, which may have affected the degree of responsiveness (and perhaps other results as well). To investigate this, the BSI should be studied on LMI too, which was not within the scope of this study.

In addition, there were some findings of smaller importance in this study. First, Pearson's correlation coefficients showed that the PROMIS CATs had a lower association with the other BSI scales than the BSI Depression and Anxiety scales did. We found this for the pretest (*Hypothesis 3*), retest (*Hypothesis 4*) and change scores (*Hypothesis 8*), which may be somewhat expected when considering that the BSI scales have more in common with each other than with the CATs. Additionally, the (partially fixed) order in which the instruments were administered may have led to differences in respondent behavior (e.g., due to measurement fatigue, context effects, or order effects; Windle, 1954). As the BSI was always administered last, this may even have influenced other results as well. Unfortunately, the questionnaire-software of Dimence Group did not allow for further alternation between the instruments. For future studies, it is suggested that both PROMIS CATs and legacy instruments are alternated.

Second, Cohen's *d* showed that the CAT Anxiety pretest scale was somewhat less able than the BSI Anxiety pretest scale to discriminate between patients with and without a primary anxiety diagnosis (*Hypothesis 5*). For the Depression scales, however, we found the opposite for patients with and without a primary depression diagnosis. These findings may be explained by the item content of the scales. The PROMIS Anxiety item bank includes items that may be more appropriate for specific anxiety diagnoses such as an obsessive-compulsive disorder, phobia, or social anxiety. Consequently, the CAT Anxiety may select items that are less relevant for patients with other anxiety diagnoses, possibly leading to a somewhat lower latent trait level. The BSI Anxiety scale, on the other hand, includes mostly general anxiety symptoms. In this case, scores may be somewhat less affected because the administered items are relevant for most anxiety diagnoses. As a result, the CAT Anxiety's ability to discriminate between patients with and without a primary anxiety diagnosis may be somewhat lower than that of the BSI Anxiety scale. In contrast, this explanation may not apply to the CAT Depression as mood disorders may be less diverse in their manifestation than anxiety disorders. In this case, the selection of items from a larger item bank may lead, relative to administering a small fixed item set, to a somewhat better discrimination between patients with and without a primary depression diagnosis.

Third, there may be some method effects in the assessment of utility of change indicators (*Hypothesis 10*). For example, we used test-retest reliability instead of Cronbach's α for calculating the *RCI* for the BSI to account for variance in scores over time. Fortunately, an additional analysis (not shown herein) showed that our conclusions remained the same when using Cronbach's α based on the pretest of this study. In addition, the cutoff for *CSC* was calculated as the point halfway the general and clinical population (taking into account the variance in scores as well). A possible limitation of this method is that we had to use general population statistics from different samples for the PROMIS CATs and the BSI. Consequently, the results may have been affected by the degree of representativeness of these samples. For example, the general population samples were collected with stratified sampling for both CATs and BSI, but the sample used for the CATs accounted for more demographics variables than the sample used for the BSI (i.e., gender, age group, education, ethnicity, and region vs. gender and size of the city of residence), had a larger sample size (N = 1,002 vs. N = 200) and was collected more recently (2016 vs. 2005). Based on these differences, we could have chosen another cutoff for *CSC* that is calculated using data of the current study only. In this method, *CSC* is defined as a patient moving more than 2 *SD's* from the mean of the clinical sample (Jacobson et al., 1984). Fortunately, an additional analysis (not shown herein) showed once more that our conclusions remained the same. These findings indicate that method effects were not meaningful for the assessment of utility of change indicators.

Last, the PROMIS CATs used a stopping rule that combined measurement precision and an upper limit of administered items while the BSI Depression and Anxiety scales always administered six fixed items. Consequently, we could not eliminate any concern that findings are due to different test lengths. This could have been solved by using a stopping rule that always administered six items according to the CAT algorithm, but we preferred to use a stopping rule that most likely will be used in clinical practice to provide test users with practical information to choose instruments. As a result, they can make their own trade-off between efficient measurement and reliable measurement of the PROMIS CATs and the BSI, based on the information available.

In this study, both PROMIS CATs were shown to be sufficiently efficient, valid, and responsive relative to the BSI subscales. For utility of change indicators, we found modest improvements for the PROMIS CATs compared to matching BSI scales, which is likely due to the PROMIS methodology. Both PROMIS CATs use state of the art CAT administration, resulting in a highly relevant selection of items that is tailored to each respondent's severity level. Furthermore, CAT ensures that each administration meets the minimally required measurement precision, by which the number of administered items is allowed to vary among respondents. Consequently, measurement is both efficient and reliable for a large range of severity levels (Flens et al., 2017, 2019). The BSI subscales, however, use fixed item sets with a small number of items. As a result, measurement precision can vary among respondents (Reise & Waller, 2009) and may be generally lower than that of the PROMIS CATs (Pilkonis et al., 2014). In addition, PROMIS CATs provide a measurement error estimate for each individual test taker while the BSI subscales only provide a single estimate of the standard error of measurement for all test takers. Consequently, change indicators may be more accurate for the PROMIS CATs compared to the BSI (Brouwer et al., 2013; Mancheño et al., 2018). Based on this, the PROMIS CATs may be considered an improvement over matching BSI scales as tools for reviewing treatment progress with patients.

For current BSI users, other results may also need to be considered to decide whether to change instruments. First, the responsiveness of the CAT Anxiety was somewhat higher than that of the BSI Anxiety scale, which was unexpected considering the results of previous studies (Kroenke et al., 2019; Pilkonis et al., 2014). Second, the administration efficiency of the instruments was quite similar. The CAT Anxiety even administered somewhat more items on average (i.e., 8 items) relative to the BSI Anxiety scale (i.e., 6 items). Note, however, that relative to the CAT Depression, the CAT Anxiety also categorized a larger degree of patients as changed compared to the matching BSI subscale, which may be due to the extra items. Finally, our study design may have disadvantaged one of the study measures by always administering the BSI last, increasing the uncertainty of the results. Based on these findings, it may not yet be appealing to all BSI users to make the transition to PROMIS CATs, especially considering that test users need to get used to new instruments, which may be experienced as a burden.

When BSI users are sufficiently convinced to change instruments, the PROsetta Stone® initiative offers the possibility to convert BSI Depression scores into PROMIS CAT Depression scores for an easier transition (www.prosettastone.org/new-page-1-1; Kaat et al., 2017). Using PROMIS instruments also has additional advantages for practice that are beyond the scope of this study. For example, PROMIS instruments are universally applicable in a wide range of populations whereas the BSI is mostly used in populations that primarily suffer from mental health problems (Beleckas et al., 2018; Lizzo et al., 2019; Papuga et al., 2018; Scholle et al., 2018; Wagner et al., 2015). PROMIS scores may even be compared across countries to learn from each other's practice (Elsman et al., 2022; Terwee et al., 2021; van Bebber et al., 2018). In addition, test users have access to numerous other PROMIS (CAT) instruments measuring different constructs of a large part of the health spectrum (for more details, see www.healthmeasures.net/explore-measurement-systems/promis/obtain-administer-measures). This means that PROMIS users have more flexibility in administering a set of instruments that specifically fits the patient's treatment goals, instead of being bound to BSI subscales that may not all have to be relevant for a patient.

Strengths of this study are the sample properties and the assessment procedure. The sample included only patients that completed the PROMIS CATs and the BSI on the same day for both pretest and retest, resulting in *N* = 400 while typically *N* = 200 is used for the performed analyses (e.g., Pilkonis et al., 2a014; Schalet et al., 2016). Furthermore, the response rate was substantial (i.e., 72.9%), and the composition of the sample (regarding gender, age, and pretest severity level) was representative for the mental health provider that collected the data. In contrast, the sample may lack representativeness for the Dutch clinical population because the data were not collected using stratified sampling. For example, the Dimence Group has many departments, covering urban and rural areas, albeit only in the east of the Netherlands. Consequently, few patients from other regions in the Netherlands were included, possibly affecting the representativeness of the sample (Dieperink et al., 2008). In addition, the patients of this study showed somewhat more severe symptoms at the start of treatment than the patients used for calibrating the PROMIS item banks for Depression and Anxiety (Flens et al., 2017, 2019), possibly affecting the representativeness of the sample too.

We have several suggestions for future research. The tentative rules of thumb that were used for some of the analyses need to be evaluated in a (simulation) study to assess whether they correspond sufficiently to the suggested interpretations. Also, our sample consisted mostly of patients with a depression or anxiety disorder (i.e., 85%). Because the PROMIS CATs for Depression and Anxiety may also be relevant for patients with other conditions, such as diabetes (Lloyd et al., 2000), cancer (Singer et al., 2010), cardiovascular diseases (Hare et al., 2014), and other mental health disorders (e.g., attention deficit disorder, somatoform disorder, personality disorder; Clarke & Kissane, 2002; Frank, 1974), it is suggested for future studies to re-evaluate the investigated psychometric properties for these conditions as well.

In addition, it is suggested to compare the DF PROMIS CATs to other legacy instruments, such as the PHQ-9 (Kroenke et al., 2001), the GAD-7 (Spitzer et al., 2006), and the Mood and Anxiety Symptom Questionnaire (MASQ; Flens et al., 2016; Watson & Clark, 1991). In a previous study, the US CAT Depression was compared to the PHQ-9 and the CESD (Pilkonis et al., 2014). Similar to our study, construct validity was found to be sufficient relative to the legacy instruments. One unexpected finding, however, was that the CAT Depression displayed the smallest pretest to retest effect size. The authors suggested that this was likely a consequence of the decreased variance in the legacy instruments due to floor effects. Furthermore, they argued that such a result raises the possibility that commonly used instruments may overestimate effect sizes. Fortunately, floor effects for the BSI scales were of minor importance in this study. In an additional analysis (not shown herein), we found for both BSI Depression and Anxiety scales that approximately 5% of the patients had a retest score of 0. However, floor effects may be generally larger when all retests are administered at the end of treatment, possibly affecting the responsiveness and the utility of change indicators.

Following this line of reasoning, the wide range in the pretest to retest interval may also have affected the results of this study. It may be, for example, that the results will be different for respondents with a small pretest to retest interval compared to respondents with a high pretest to retest interval (e.g., due to differences in floor effects in the BSI scales). To investigate this, we split the study sample into two equal halves based on the median pretest to retest interval, and repeated the analyses of this study (not shown herein). We found that our conclusions remained the same in both subsamples, indicating that the length of the pretest to retest interval did not have a substantial effect. However, it may be recommended for follow-up research to additionally evaluate this for patients that are reassessed over even longer time-intervals. Note, for example, that the retest scores in this study were still somewhat high, and the change scores somewhat low. Therefore, the question remains whether the results will also be similar when the change scores are larger.

In this study, we compared the DF PROMIS adult v1.0 item banks for Depression and Anxiety administered as CAT with the nine subscales of the BSI in a clinical sample. Overall, our study suggests that the PROMIS CATs measure the same constructs as matching BSI scales, were at least as responsive, and categorized the change scores of more patients as actually changed. Based on these findings, the PROMIS CATs may be considered a modest improvement over matching BSI scales as tools for reviewing treatment progress with patients.

Chapter 7

Discussion

7.1 Can the main goal of this thesis be considered achieved?

In this thesis, it was aimed to solve three limitations of the current set of measurement instruments used in Dutch mental health care to evaluate patients' treatment. First, it is unclear for many instruments whether their quality is adequate for treatment evaluation because relevant psychometric properties have been studied insufficiently. Second, the use of fixed item sets has made it challenging to develop instruments that are both highly reliable and highly efficient. Finally, the large number of available instruments measuring the same construct(s) has made it difficult for mental health providers to learn from the treatment outcomes of other mental health providers.

To work towards a possible solution, it was aimed to lay the foundation for a new set of mental health instruments using modern methodologies. Specifically, the Dutch-Flemish (DF) Patient-Reported Outcomes Measurement Information System (PROMIS®) adult v1.0 item banks for Depression and Anxiety were psychometrically evaluated for computerized adaptive test (CAT) administration. CAT instruments ensure that items are selected in such a way that the next item is the most informative for updating a person's latent trait level (i.e., the severity level of the measured construct) with a higher measurement precision. Furthermore, the administration of items continues only for as long as is necessary to assess the latent trait level with a predetermined measurement precision. As a result, the use of CAT instruments should lead to measurement that is both reliable and efficient. This, in turn, should not only lead to an increase in the completeness of information deemed relevant to evaluate patients' treatment (due to high efficiency), but combined with the PROMIS item banks, it should also lead to more high-quality information that reduces the probability of a clinician making biased inferences. Therefore, this new set of instruments may have the potential to be the new standard in the Netherlands for evaluating patients' treatment.

Based on previous studies on the United Stated (US) PROMIS item banks, it was expected that the DF PROMIS CATs for Depression and Anxiety would measure efficiently, reliably, validly, and responsively in the Dutch general and clinical population (Kroenke, Baye, & Lourens, 2019; Pilkonis et al., 2011, 2014; Schalet et al., 2016). Moreover, due to the use of CAT technology and highly informative item banks, the DF PROMIS CATs were expected to measure even more efficiently and reliably than other instruments (Pilkonis et al., 2014). In the first main section of this discussion, it is examined whether the DF PROMIS CATs meet these expectations for the Dutch clinical (i.e., patients with common mental disorders in ambulatory mental health care) and general population. In other words: can the main goal of this thesis be considered achieved? To answer this question, I will first provide a summary of the current thesis for each studied psychometric property. I will then describe the general strengths and limitations of the PROMIS CAT studies. Finally, I will reach a conclusion.

7.1.1 Summary

7.1.1.1 Efficient and reliable measurement

The items in the PROMIS item banks were specifically chosen for their discriminative ability and coverage of the depression and anxiety constructs (Cella et al., 2010). As a result, it was shown that the US PROMIS adult v1.0 item banks for Depression and Anxiety are highly informative for a wide range of latent trait levels, and administered as CAT, measure depression and anxiety both efficiently and reliably in the US clinical and general population (Pilkonis et al., 2011, 2014). Similarly, it was shown in Chapter 3 and 4 that the DF versions of the PROMIS adult v1.0 item banks for Depression and Anxiety are also highly informative for a wide range of latent trait levels in Dutch samples. Moreover, both post hoc CAT simulations (Chapter 3 and 4) and genuine CAT administrations (Chapter 6) showed that the DF PROMIS CATs measure depression and anxiety reliably and efficiently in the Dutch clinical and general population too. With the measurement precision set to a *high* precision standard for individual assessments (i.e., standard error [SE] = 0.22; Bernstein & Nunnally, 1994), the CAT Depression administered about 6.6 items on average and the CAT Anxiety administered about 8.7 items on average. These numbers can even be lowered to about 4 in situations where less precision is acceptable. This may apply, for example, to the assessment of groups or patients who do not primarily suffer from mood or anxiety disorders.

As expected, the DF PROMIS adult v1.0 item banks for Depression and Anxiety were less informative for persons with low or very high severity levels (Chapter 3 and 4). This is not a specific issue of the DF PROMIS item banks, but of mental health instruments in general: they often lack a sufficient number of items to discriminate well among the lower or higher latent trait levels because it is challenging to compose such items (Reise & Waller, 2009). To deal with this issue, the stopping rule of the CAT algorithm included an upper limit of administered items. This upper limit was established by using the criterion that at least 90% of the clinical subjects resulted in a high measurement precision standard for individual assessments, resulting in 9 items for the CAT Depression and 12 for the CAT Anxiety. Consequently, it was shown for the CAT Anxiety that many patients with low or very high latent trait levels are measured sufficiently reliable too, without sacrificing too much efficiency (Chapter 4).

In Chapter 6, the DF PROMIS CATs were compared to the subscales of the Dutch legacy instrument Brief Symptom Inventory (BSI; de Beurs & Zitman, 2005). The results showed that the number of administered items of the PROMIS CATs was highly comparable to that of matching BSI subscales, indicating a similar efficiency. Comparing the reliability of measurement, on the other hand, was more challenging because the PROMIS CATs adopt item response theory (IRT) and the BSI classical test theory (CTT) as underlying measurement theory. That being said, the PROMIS CATs did show some evidence for modest improvements in reliability. Under the assumption of measuring similar constructs, the PROMIS CATs categorized somewhat more patients as (reliably) changed compared to matching BSI subscales. This may suggest that the PROMIS CATs are more able to detect actual change, probably due to a greater reliability (Pilkonis et al., 2014). Also, the PROMIS CATs estimate a specific reliability level for each individual test taker while the BSI subscales only provide a single reliability estimate for all test takers. As a result, the PROMIS CATs may estimate patients' change categorizations more accurately compared to matching BSI subscales (Brouwer, Meijer, & Zevalkink, 2013; Mancheño et al., 2018).

Finally, it was evaluated in Chapter 3 whether the reliability and efficiency of the CAT Depression could be further improved by adding more items to the corresponding item bank.

Using post hoc CAT simulations, the original Depression 28-item bank was compared to an extended 48-item bank under several measurement precision thresholds. The results showed that both the number of administered items and Pearson's correlation coefficient between CAT scores and full item bank scores were highly similar for the item banks. For the PROMIS Anxiety item bank, similar results were found in the pre-analysis stage of Chapter 4. Consequently, it was concluded that the reliability and efficiency of the DF PROMIS CATs is not improved much further by adding additional items to the corresponding item banks.

7.1.1.2 Valid measurement

In previous studies, it was demonstrated with US clinical and general population samples that the US PROMIS adult v1.0 item banks for Depression and Anxiety are sufficiently valid for cross-sectional usage in the US (Pilkonis et al., 2011, 2014). Similarly, it was shown with Dutch clinical and general population samples that the DF versions of the PROMIS adult v1.0 item banks for Depression and Anxiety are also sufficiently valid for cross-sectional usage in the Netherlands (Chapter 3 and 4). Specifically, this was demonstrated for the sources of evidence known as unidimensionality, local independence (LI), monotonicity, absence of differential item functioning (DIF), and fit of the graded response model (GRM; Samejima, 1969). Consequently, both PROMIS item banks are said to have valid item parameters as input for the CAT algorithm (Reeve et al., 2007).

In addition, the DF PROMIS CATs were compared to the BSI to investigate several other sources of evidence for cross-sectional usage in the Dutch clinical population (Chapter 6). The results indicated that the PROMIS CATs sufficiently matched the validity of the BSI subscales regarding convergent validity, divergent validity, and concurrent validity. For the CAT Depression, this was also the case for the stability of the pretest to retest scores. The CAT Anxiety, however, was shown to be somewhat less stable than the BSI Anxiety scale, but as the difference was minor and all other sources of evidence were sufficient, this was not considered problematic for overall validity. Consequently, it was concluded that the PROMIS CATs measure similar constructs as matching BSI subscales. This conclusion was in line with previous studies that used US clinical samples to compare the US PROMIS CATs for Depression and Anxiety to several legacy instruments (Pilkonis et al., 2011, 2014), including the Center for Epidemiological Studies Depression scales (CESD), the Patient Health Questionnaire (PHQ-9), and the Mood and Anxiety Symptom Questionnaire (MASQ).

Finally, Chapter 5 describes the first study in which longitudinal measurement invariance (LMI) was investigated in any of the PROMIS item banks. An item bank is said to be longitudinally measurement invariant when it measures one or more single constructs in the same way over time. To evaluate this longitudinal validity aspect, the study included pretest and retest data of two Dutch clinical samples in treatment for mood or anxiety disorders. The results indicated that the DF PROMIS adult v1.0 item banks for Depression and Anxiety are sufficiently unidimensional at both pretest and retest, but also that two of the four invariance assumptions were violated for both item banks (i.e., threshold invariance and unique factor invariance). Further investigation, however, revealed that the impact of these invariance violations on the mean latent change score did not exceed the proposed cutoff value. Also, none of the response categories of the Depression item bank were substantially affected. For the

Anxiety item bank, only the response Category *rarely* for Item EDANX07 *I felt like I needed help for my anxiety* was somewhat affected by the threshold invariance violation. Consequently, it was concluded that the practical significance of the invariance violations is negligible for both item banks. This means that even though some violations of LMI were found, the DF PROMIS adult v1.0 item banks for Depression and Anxiety may still provide sufficiently invariant scores for treatment evaluation.

7.1.1.3 Responsive measurement

In previous studies, it was demonstrated with US clinical samples that the responsiveness of US PROMIS CATs and short-forms for Depression and Anxiety is comparable to that of multiple legacy instruments (Kroenke, Baye, & Lourens, 2019; Pilkonis et al., 2014; Schalet et al., 2016). These instruments include the CESD, PHQ-9, Generalized Anxiety Disorder (GAD-7), Symptom Checklist (SCL), Posttraumatic Stress disorder checklist (PCL), Short Form (SF)-36, and SF-12 Mental Component Summary (MCS). Similarly, it was shown with a Dutch clinical sample that the responsiveness of the DF PROMIS CAT for Depression is comparable to that of the Dutch BSI Depression scale (Chapter 6). For the DF PROMIS CAT for Anxiety, responsiveness was shown to be higher relative to the Dutch BSI Anxiety scale, which may suggest that the CAT Anxiety is more able to detect change. However, as it was expected that the responsiveness of the Anxiety instruments would have been similar based on the US findings, two explanations were provided for this unexpected result. These explanations include (a) the choice of item parameters used to calculate T-scores for the PROMIS CATs (US vs. DF), and (b) a possible difference between the instruments in the degree of LMI.

7.1.2 Strengths

In this section, three general strengths of the PROMIS CAT studies are discussed. First, a wide collection of psychometric properties was evaluated. Validity in particular was studied thoroughly by assessing convergent validity, divergent validity, concurrent validity, stability, unidimensionality, LI, monotonicity, GRM fit, and DIF between subgroups (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1974; Cook & Campbell, 1959; Drenth & Sijtsma, 2005; Reeve et al., 2007). Also, LMI was evaluated (Liu et al., 2017). This is a longitudinal validity aspect that has barely been studied for other Dutch mental health instruments (e.g., Carlier et al., 2019; Jabrayilov, Emons, de Jong, & Sijtsma, 2017; te Poel, Hartmann, Baumgartner, & Tanis, 2017). Moreover, the results of the study suggested that LMI was sufficiently supported for both of the PROMIS item banks. This can be considered highly relevant information for test users.

Second, the methodology adopted in the PROMIS CAT studies may have increased the generalizability of the results to the Dutch clinical and general population. For example, the item parameters for the PROMIS CATs were estimated with a multiple group IRT model instead of a single group IRT model, which was used for the estimations of the US PROMIS item banks. Basically, both models can scale the latent trait to the general population. However, by merely using a general population sample in a single group model for this purpose, the number of persons with average to high severity levels may be too low to estimate accurate item parameters for the entire latent trait continuum. In a multiple group model, this issue can be handled by adding a clinical sample as a separate group and fixing the item parameters to be equal across groups (McDonald, 1999; Smits, 2016). As a result, the item parameters may be more representative for the entire latent trait continuum because the IRT model is fitted on the item responses of a sufficient number of respondents for all relevant latent trait levels.

In addition, both post hoc CAT simulations (Chapter 3 and 4) and genuine CAT administrations (Chapter 6) were used to evaluate the reliability and efficiency of the DF PROMIS CATs. It was shown that both methods led to highly similar results based on different clinical samples, increasing the generalizability of the results to the Dutch clinical population. Furthermore, these results are in line with a previous study that used post hoc CAT simulations to demonstrate they are useful to assess the measurement properties of genuine CAT administrations (Kocalevent et al., 2009). Assuming this is the case then, it may be that the CAT simulation results found for the Dutch *general population* sample are also sufficiently generalizable (Chapter 4).

Finally, some of the sample properties may have increased the representativeness for the Dutch clinical and general population. For example, the item parameter estimations for the DF PROMIS CATs were based on 2,010 clinical and general population subjects, whereas 1,000 is considered to be a minimum requirement (Reise & Yu, 1990; Chapter 3 and 4). Furthermore, the aims to include at least 500 patients in the study of Chapter 5 in order to adequately examine factor structures (Comrey & Lee, 1992; Liu et al., 2017; MacCallum, Widaman, Zhang, & Hong, 1999) and to include at least 200 patients in the study of Chapter 6 based on similar studies (Pilkonis et al., 2014; Schalet et al., 2016), were also achieved. For Chapter 6, it was even managed to include 400 patients in the study.

In addition, stratified sampling was applied to optimize the representativeness of the general population sample, incorporating five stratification variables to mirror the Dutch population (i.e., gender, age, education, ethnicity, and region; Chapter 3 and 4). For the three clinical samples, stratified sampling was not applied, but the samples did show that the composition regarding gender and age was representative for the mental health providers that collected the data (Chapter 3, 4, 5, and 6). For the longitudinal studies, the clinical samples were additionally evaluated on pretest severity level, which also showed sufficient representativeness for the mental health providers that collected the data (Chapter 5 and 6). Finally, the patients included in the studies of Chapters 3, 4 and 5 were diagnosed with the Mini International Neuropsychiatric Interview (MINI-plus; Sheehan et al., 1998). This may have increased the accuracy of the diagnoses compared to merely using the clinician’s point of view (Aboraya, Rankin, France, El-Missiry, & John, 2006).

7.1.3 Limitations

In this section, four general limitations of the PROMIS CAT studies are discussed. First, the DF PROMIS CATs were only compared to a single legacy instrument: the BSI. The BSI is a popular instrument to evaluate patients’ treatment progress in the Netherlands (and internationally), but so are others. These include, for example, the MASQ (Watson & Clark, 1991), the Outcome Questionnaire (OQ-45; de Jong et al., 2007), the Symptom Questionnaire-48 (SQ-48; Carlier et al., 2012b), and the Depression and Anxiety Stress Scale (DASS;

Lovibond & Lovibond, 1995). To get a better understanding of the quality of the DF PROMIS CATs, it is therefore suggested to compare them to other instruments as well.

Second, the DF PROMIS adult v1.0 item banks for Depression and Anxiety were psychometrically evaluated for CAT administration in the Dutch clinical (i.e., patients with common mental disorders in ambulatory mental health care) and general population. Consequently, the results can only be generalized to these populations. However, as the measurement of depression and anxiety also bears relevance for patients with other conditions, such as diabetes (Lloyd et al., 2000), cancer (Singer et al., 2010) and cardiovascular diseases (Hare et al., 2014), it may be necessary to evaluate the investigated psychometric properties for these conditions as well. Fortunately, it is expected that the DF PROMIS CATs for Depression and Anxiety will also demonstrate favorable psychometric properties in other populations (e.g., Amtmann et al., 2014; Kudel et al., 2019; Schalet et al., 2016; Teresi et al., 2016a, 2016b). This is in line with PROMIS' aim to develop instruments that are universally applicable.

Third, the methodology adopted in the PROMIS CAT studies may have decreased the generalizability of the results to the Dutch clinical and general population. For example, several tentative rules of thumb were used to evaluate concurrent validity, stability, responsiveness, and the practical significance of LMI violations (Chapter 5 and 6). These rules of thumb need to be evaluated in a (simulation) study to assess whether they correspond sufficiently to the suggested interpretations. In addition, some psychometric properties could have been evaluated with alternative methodology, which may affect the conclusions. For example, LMI can be evaluated with an alternative approach that does not depend on the specific identification condition chosen for the baseline model, possibly improving the accuracy of the results (Wu & Estabrook, 2016). Also, Monte-Carlo simulations can be used to derive empirical criteria that maximize the ability to identify both uniform and nonuniform DIF, and control for the overall Type I error rate (Choi, Gibbons, & Crane, 2011; Elsman, Flens, de Beurs, Roorda, & Terwee, 2022).

Finally, some of the sample properties may have decreased the representativeness for the Dutch clinical population. For example, the composition of the clinical samples was dependent on the willingness of mental healthcare providers and their patients to participate in the studies. And even though the samples were representative for the mental healthcare providers that collected the data regarding gender, age, and pretest score, this does not imply that the samples are also representative for the entire Dutch clinical population of outpatients with common mental health disorders. For example, the mental healthcare provider Dimence Group has many departments, covering urban and rural areas, albeit only in the east of the Netherlands. Consequently, few patients from other regions were included, possibly affecting the representativeness (Dieperink, Mulder, van Os, & Drukker, 2008; Chapter 6).

In addition, the longitudinal psychometric properties LMI and responsiveness may best be evaluated with data that are representative for the entire length of patients' treatment (Chapter 5 and 6). However, information regarding the actual length of treatment was not available in the PROMIS CAT studies. This means that the results might have been different had the retest always been administered at the end of treatment. Also, the longitudinal PROMIS CAT studies showed that the pretest to retest interval varied substantially between respondents.

The results of these studies might have been different when the tests would have been administered more uniformly (e.g., always 6 months after the pretest).

7.1.4 Conclusion

In this thesis, the DF PROMIS adult v1.0 item banks for Depression and Anxiety were psychometrically evaluated for CAT administration in the Dutch clinical (i.e., patients with common mental disorders in ambulatory mental health care) and general population. The results showed that both item banks were highly informative for a large variety of Dutch latent trait levels, making them highly suitable for CAT administration. This was confirmed by the actual CAT administrations, which demonstrated to measure both reliably and efficiently in the Dutch clinical and general population. Also, the PROMIS CATs were shown to measure sufficiently valid in the Dutch clinical and general population, based on many sources of evidence commonly claimed as indicative for validity. This even includes LMI, which has barely been studied for other Dutch instruments measuring mental health constructs. Finally, the PROMIS CATs were shown to measure sufficiently responsive in a Dutch clinical sample. Based on these findings, it can be concluded that the main goal of this thesis has been sufficiently achieved: the DF PROMIS adult v1.0 item banks for Depression and Anxiety administered as CAT measure efficiently, reliably, validly, and responsively in the Dutch clinical and general population.

That being said, two findings stand out in the PROMIS CAT studies. First, the CAT instruments only showed modest improvements compared to matching subscales of the BSI. Consequently, it may not seem very appealing to test users to make the transition to PROMIS CATs, especially considering that new instruments need to get used to, which may be experienced as a burden. When test users are sufficiently convinced to change instruments, the PROsetta Stone® initiative offers the possibility to convert the scores of several depression and anxiety instruments into PROMIS scores for an easier transition (www.prosettastone.org; e.g., BSI Depression; Kaat et al., 2017). Additionally, it should be noted that the PROMIS CATs (may) have other benefits compared to the current set of available instruments. Several of these were already introduced in Chapter 6: PROMIS instruments are designed for universal application in a wide range of populations (e.g., Beleckas et al., 2018; Lizzio et al., 2019; Papuga et al., 2018; Wagner et al., 2015), whereas many other instruments measuring mental health constructs are mostly used in populations that primarily suffer from mental health problems; PROMIS scores may be compared across countries to learn from each other's practice (Elsman et al., 2022; Terwee et al., 2021; van Bebber et al., 2018), whereas this may be more unclear for other instruments; and PROMIS CAT users have access to numerous other PROMIS (CAT) instruments measuring different constructs of a large part of the health spectrum (e.g., Crins et al., 2015, 2016, 2017, Terwee et al., 2019), allowing a lot of flexibility in composing a set of instruments to evaluate treatment goals. In addition, it was demonstrated in Chapter 5 that the PROMIS CATs were sufficiently invariant over time while this is still unclear for many other Dutch instruments measuring mental health constructs. Consequently, PROMIS CAT users may have more certainty that patients' scores are sufficiently unbiased for treatment evaluation. Finally, the benefits of the PROMIS CATs will likely become more evident when compared to other instruments than the BSI. For example, it was demonstrated in

Chapter 2 that CAT versions of the three MASQ subscales may lead to a mean decrease in items of 56% up to 74% with a negligible loss of measurement precision. Consequently, the efficiency gains of the PROMIS CATs are much larger compared to the MASQ subscales, possibly convincing more test users of the benefits of CAT instruments.

Second, there were some differences between the PROMIS CATs in their demonstration of psychometric properties. While the CAT Depression consistently demonstrated good psychometric properties, the CAT Anxiety performed somewhat less well overall. For example, Chapter 5 showed sufficient LMI for both PROMIS item banks, but the impact of the threshold invariance violation on change scores was closer to the cutoff value for substantial bias (10%) for the Anxiety item bank (9.58%) than for the Depression item bank (6.82%). Consequently, it may be more likely that the Anxiety item bank lacks threshold invariance after all, considering that the used cutoff value still needs to be evaluated in a simulation study to investigate whether it corresponds to the proposed interpretation. In addition, Chapter 6 showed that the validity and responsiveness of the CAT Depression was similar to that of the BSI Depression scale, which was in line with previous US studies (Pilkonis et al., 2014; Schalet et al., 2016). Unexpected, however, was that the CAT Anxiety demonstrated less stability and a higher responsiveness compared to the Dutch BSI Anxiety scale. This was not in line with previous US studies (Kroenke, Baye, & Lourens, 2019; Schalet et al., 2016), introducing more uncertainty about the quality of the DF version of the CAT Anxiety. Finally, both post hoc CAT simulations (Chapter 3 and 4) and genuine CAT administrations (Chapter 6) showed smaller efficiency gains for the CAT Anxiety compared to the CAT Depression. Setting the measurement precision to a high precision standard for individual assessments, the CAT Anxiety administered about 2 items more on average. This means that the PROMIS Anxiety item bank is somewhat less informative than the PROMIS Depression item bank, which was confirmed by the results presented in Chapter 3 and 4. Based on these findings, it may be suggested to investigate whether the psychometric properties of the CAT Anxiety can be further improved. For example, the PROMIS Anxiety item bank includes items that may be more appropriate for specific anxiety diagnoses such as an obsessive-compulsive disorder, phobia, or social anxiety. This might imply that the generic anxiety construct as measured by the PROMIS item bank additionally consists of several subdomains. In that case, the measurement of anxiety may be somewhat more appropriate for multidimensional computerized adaptive testing (MCAT; see section 7.2). Alternatively, the PROMIS Anxiety item bank may benefit from content balancing to ensure that different subdomains of anxiety are sufficiently taken into account.

In this thesis, two CAT instruments were evaluated for the measurement of depression and anxiety in Dutch persons. The results showed that the DF PROMIS CATs for Depression and Anxiety measure efficiently, reliably, validly, and responsively in the Dutch clinical and general population. Furthermore, the DF PROMIS CATs were shown to be a modest improvement over matching subscales of the popular BSI. Add to that the additional benefits of PROMIS instruments for clinical practice, and test users may be sufficiently convinced to implement PROMIS CATs as tools for evaluating patients' treatment. Meanwhile, we can begin to work towards the next generation of CAT instruments to improve the benefits for measurement even further.

7.2 Future directions to improve CAT methodology

CAT methodology has more to offer than I have been able to show in this thesis. Therefore, I will elaborate in the following section on possible future directions to improve the CAT methodology currently used in the Netherlands. These future directions include (a) MCAT, (b) the stopping rule of the CAT algorithm, (c) the latent trait estimator of the CAT algorithm, and (d) the use of appropriate item parameters.

First, it was shown in Chapter 3, 4 and 5 that the DF PROMIS adult v1.0 item banks for Depression and Anxiety are sufficiently unidimensional. Other studies, however, have also found a good fit for models that explain the relationship between *two or more* depression and/or anxiety constructs (Kose & Demirtasli, 2012). These models, also known as multidimensional models, include a two-dimensional correlated traits model (Bass, Morris, & Neapolitan, 2015) and a bi-factor model (Gibbons et al., 2012, 2014; www.adaptivetestingtechnologies.com), both of which are introduced below. The use of multidimensional models may provide additional advantages for measurement in clinical practice, especially when applying CAT technology. This is also known as MCAT (Smits, Paap, & Böhnke, 2018).

In a two-dimensional correlated traits model, a depression construct and an anxiety construct can be treated as primary dimensions that are correlated with each other. The benefit of adopting this model in the CAT algorithm is that the latent trait estimate for one construct also provides information for the estimation of the other. As a result, the number of administered items can even be smaller in a single MCAT administration than in two separate unidimensional CAT administrations (Paap, Born, & Braeken, 2019). Exactly this was demonstrated by a previous study that performed Monte-Carlo simulations on the US PROMIS adult v1.0 item banks for Depression and Anxiety (Bass, Morris, & Neapolitan, 2015). Using several measurement precision thresholds, the authors showed that an MCAT administration based on a two-dimensional correlated traits model reduced the number of administered items by 23% to 8% when compared to two separate unidimensional CAT administrations. This means that a PROMIS MCAT for Depression and Anxiety may decrease respondent burden even further without loss of measurement precision. Obviously, this is based on the assumption that both depression and anxiety symptoms are problematic for a patient, which is actually quite common (de Beurs et al., 2007). If "only" depression *or* anxiety symptoms are problematic, the administration can simply be limited to a single unidimensional CAT administration.

In the case of a two-dimensional correlated traits model, CAT technology is already available in the Netherlands through the DF Assessment Center. For MCATs based on a bi-factor model, however, CAT technology is not yet available. In a basic bi-factor model, an item measures a primary dimension (e.g., depression or anxiety) and additionally can measure a single subdomain. As a result, a test user can use information on specific subdomains without having to administer additional instruments. In previous studies (Gibbons et al., 2012, 2014), a bi-factor model was used to create a Depression item bank of 398 items measuring five subdomains (mood, cognition, behavior, somatic, and suicide) and an Anxiety item bank of 431 items measuring four subdomains (mood, cognition, behavior, and somatic). The authors then used post hoc MCAT simulations to demonstrate that an average of 12 items need to be administered to reliably measure depression or anxiety with the accompanying subdomains. On the one hand, this means that the number of administered items was somewhat larger in the MCATs than in the DF PROMIS CATs for Depression and Anxiety. On the other hand, the MCATs may provide more useful information for test users because the instruments also include subdomains of Depression and Anxiety.

Second, the stopping rules used to terminate the CAT administrations may be improved. For the currently used stopping rules of the PROMIS CATs, a fixed measurement precision was combined with an upper limit of administered items. Alternative stopping rules, however, can also take into account whether additional items could still increase the measurement precision or change the latent trait estimate to a prespecified degree (Babcock & Weiss, 2013; Choi, Grady, & Dodd, 2011). Consequently, measurement may become even more precise or efficient. It may become more precise because the CAT algorithm keeps selecting items until the precision cannot be improved much further. In this case, the increase in measurement precision may outweigh the administration of additional items. Alternatively, measurement may become more efficient because the administration does stop when an additional item no longer improves the precision or change the latent trait estimate to a considerable degree. In this case, the administration of additional items is unnecessary because the measurement precision and/or latent trait level cannot be substantially affected anymore. Thus, by using stopping rules that are more dynamic, CAT instruments may measure respondents even more precisely and efficiently.

Third, the latent trait estimator used in the CAT algorithm may be further investigated. In Chapter 2 and 3, the Bayesian estimator maximum a posteriori (MAP) was adopted in the CAT algorithm. Due to new information, however, the maximum likelihood (ML) estimator was adopted in later chapters (for more details, see Chapter 4). This means that in both instances the estimator deviated from PROMIS convention, which is the Bayesian estimator expected a posteriori (EAP). For standardization purposes, it may be recommended to further investigate the effects of each estimator on the assessment of groups and individuals (Penfield & Bergeron, 2005; Wang & Vispoel, 1998).

Finally, one of PROMIS' ambitions is to combine and transform all existing patient-reported outcome measures into one state of the art assessment system to measure self-reported health globally (Cella et al., 2007, 2010). Specifically, this means that PROMIS aims to implement identical item banks and US item parameters in every country to increase uniformity and enhance international comparability (Paz, Spritzer, Morales, & Hays, 2013; Wahl et al., 2015). For the Dutch clinical and general population, previous studies have demonstrated that the PROMIS CATs for Depression and Anxiety can use the US item parameters for the measurement of groups and most individuals (Elsman et al., 2022; van Bebber et al., 2018). These results imply that the country-specific item parameters estimated in Chapter 3 and 4 may not be necessary to measure depression and anxiety in Dutch persons. That being said, it may be suggested to investigate the effects of US item parameters somewhat further for the DF PROMIS CATs (as compared to the full PROMIS item banks). For example, it was shown in Chapter 5 that the US parameters affected the responsiveness of the CAT Anxiety positively. Had the Dutch item parameters been used to calculate the T-scores, the conclusion would have been that the CAT Anxiety and the BSI Anxiety scale are similarly responsive. Apparently, the

DF item parameters led to a somewhat lower responsiveness due to a larger range in pretest scores. Consequently, the use of US item parameters may require additional study for (longitudinal) CAT administrations of anxiety in Dutch patients.

According to PROMIS convention, every country uses identical items banks and US parameters unless sufficient evidence is provided that the resulting T-scores are substantially biased by this approach. This means that the appropriateness of US parameters is recommended to be studied for other countries and other PROMIS instruments as well. However, PROMIS has not yet provided clear guidelines on the meaning of "substantially biased". A risk of this lack of guidance may be that different studies use different criteria to assess DIF between countries, potentially leading to conclusions that are too optimistic in some studies. It may even be possible that (some) PROMIS instruments will not be investigated on the appropriateness of US item parameters at all. That being said, it should be noted that almost all countries actively involved in PROMIS translations intend to investigate the appropriateness of US item parameters. These studies should shed more light on the validity of this methodology. Alternatively, it may be argued that PROMIS should pursue global item parameters when the goal is a globally used measurement system. This at least requires the pooling of data from various countries and the investigation of DIF between those countries to assess whether global item parameters are appropriate.

7.3 Points of attention for CAT implementation

In the previous section, I suggested several future directions to improve the CAT methodology currently used in the Netherlands. Meanwhile, CAT instruments can be implemented in Dutch clinical practice to evaluate the treatment of patients. In this section, I will elaborate on four points of attention that may help in this regard. These points include (a) adopting measurement based care, (b) increasing the availability and accessibility of CAT technology, (c) raising awareness on CAT instruments, and (d) providing useful feedback tools.

7.3.1 Adopting measurement based care

It goes without saying that a mental health provider first needs to decide to adopt measurement based care (MBC) before CAT instruments are chosen as the specific tool to evaluate patients' treatment. In MBC, measurement instruments are used to aid clinicians in clinical decision-making concerning the patient's diagnosis, treatment selection and termination, treatment of nonresponders, and relapse prevention (de Beurs et al., 2018; Greenhalgh et al., 2018; Lambert, 2010; Lewis et al., 2015; Martin-Cook et al., 2021). As a result, patients' motivation to continue treatment may be increased, and patients' treatment outcomes may be improved (de Jong et al., 2021; Fortney et al., 2018; Guo et al., 2015; Rush & Thase, 2018; Scott & Lewis, 2015). Moreover, if MBC is combined with *shared-decision-making*, in which patients are supported to participate in the decisions concerning treatment, patients' treatment outcomes may be further improved (Metz et al., 2019; van der Feltz-Cornelis et al., 2014). Finally, aggregation of an instrument's scores allows for comparisons between groups, and, when combined with data of patient characteristics and treatment process aspects, aggregated data can be used to improve the overall quality and value of care for patients (de Beurs et al., 2018; Porter, 2009).

Based on these benefits, it may be expected that many mental health providers have already implemented MBC in their practice. Unfortunately, successful implementation of MBC has been shown to be complex and highly challenging. In the US, the United Kingdom, and Australia, less than 20% of the practitioners actually make use of MBC in their daily practice (Lewis et al., 2019). In Dutch mental health care, this percentage is less clear, but it may be somewhat similar (van Sonsbeek, Hutschemaekers, Veerman, Vermulst, & Tiemens, 2021). For more information on promising Dutch initiatives regarding MBC, see for example www.hetklikt.nu and www.uitkomstgerichtezorg.nl.

Successful implementation of MBC is related to several points of attention (Martin-Cook et al., 2021; www.isoqol.org/wp-content/uploads/2019/09/2015UsersGuide-Version2.pdf). These include, for example, an active role of leadership (e.g., moving MBC enthusiastically forward and providing employees with time and resources), clinical engagement (e.g., willingness and understanding to implement MBC; feeling ownership), patient engagement (e.g., willingness and understanding to respond to the administered instruments and discuss the results), fitting the set of instruments to the patient's disorder(s) and treatment goals, and using patient-friendly interfaces and feedback tools. Also, the use of brief (reliable) self-report instruments is encouraged, as these burden patients as little as possible. This, in turn, creates the possibility to administer instruments more often during treatment. In other words, compared to many traditional instruments, CAT instruments may have positive effects on the implementation success and the benefits of MBC.

The challenges that need to be overcome to successfully implement MBC increase the risk that implementation will not, or only partially, succeed. This is not only a waste of time, costs and resources, another disadvantage may be that MBC acquires a bad reputation. For example, clinicians may come to perceive measurement instruments as burdensome for both patients and themselves because "they cannot measure a patient's complex problems", or may come to believe that the measurements' outcomes will primarily be used by management to unfairly judge clinicians on their effectiveness and efficiency. If such perceptions and believes are not paid proper attention in the implementation process of MBC, they may be harder to overcome in the future. Consequently, mental health providers should be highly aware of the investment they need to make to reap the benefits of MBC. Otherwise, there is a good chance that implementation will not succeed.

7.3.2 Increasing the availability and accessibility of CAT technology

Another basic condition to implement CAT instruments in clinical practice is that mental health providers can administer them to patients. To accomplish this, mental health providers must have implemented a digital solution with access to CAT technology. In the Netherlands, almost all mental health providers have implemented a digital solution to administer measurement instruments to patients. Moreover, by 2021, many of these solutions (i.e., ICT-providers of measurement instruments) have access to CAT technology through their connection with the DF Assessment Center. These include Vital Health (www.philips.nl/healthcare/sites/vitalhealth/products/questmanager-vragenlijstenbeheer-proms-rom), EasyROM (www.kgvp.org/nl), OnlinePROMS (www.onlineproms.nl), Datec (www.datec.nl), Qualizorg (www.qualizorg.nl), BrightFish (www.brightfish.nl), MobileCare (www.mobile-care.nl), KLIK (www.hetklikt.nu), and Fysiomanager (www.fysiomanager.nl). Other ICT-providers, however, do not yet have access to CAT technology (e.g., NETQ Healthcare; www.netqhealthcare.nl). This means that many health care providers are still not able to start using CAT instruments.

That being said, the administration of CAT instruments through the DF assessment Center, which by 2021 is the only widely available CAT solution in the Netherlands, is not free of charge (i.e., a small fee is requested per assessment). This will probably demotivate some healthcare providers to start using CAT instruments because an increasing number of traditional instruments is free of charge. Therefore, it is highly desirable that alternative CAT solutions will find their way into Dutch mental health care to make CAT technology not only available, but also accessible for mental health providers. It should be noted, however, that PROMIS instruments are managed by the DF PROMIS National Center (www.dutchflemishpromis.nl), and are only allowed to be administered through the DF Assessment Center.

7.3.3 Raising awareness on CAT instruments

The third point of attention for the implementation of CAT instruments in clinical practice is that test users are properly informed and educated about their benefits and availability. To accomplish this on a large scale, it may be helpful to develop several means of communication, such as factsheets, instruction videos, and digital presentations. These communication means can be spread through social media, websites, or e-mail to communicate how CAT methodology works, how the available CAT instruments compare with traditional instruments, and how CAT instruments can be implemented in daily clinical practice. In this way, test users can be informed relatively easy about CAT instruments.

When informing test users, it may be suggested to give special attention to possible misconceptions about CAT methodology and alternatives to CAT instruments. To start with the former, some clinicians that participated in the PROMIS CAT studies of this thesis objected that individual items cannot be used to monitor change over time because CAT usually administers a different set of items on a retest occasion. This objection, however, is based on the misconception that individual items can be properly used to reliably monitor change in patients. Unfortunately, the reliability of a single item score is often too low for this purpose (Gliem & Gliem 2003). This means that the administration of different item sets can actually be seen as a benefit of CAT instruments because they place more emphasis on the score that generally is sufficiently reliable: the final latent trait estimate. This specific information may help test users develop a more positive attitude towards CAT instruments.

In addition, it may be that alternatives to PROMIS CAT instruments are more appealing to test users. First, a new initiative has been started that also developed a novel set of (mental health) instruments using adaptive testing: the NORSE feedback system (McAleavey, Nordberg, & Moltu, 2021; www.norsefeedback.no/en). However, in contrast to PROMIS, this measurement system was only available in Norway and the UK early 2022, but not yet in the Netherlands. Second, as noted in section 7.3.2, many instruments used in Dutch mental health care are free of charge whereas CAT instruments administered through the DF Assessment Center are not. Consequently, test users should be made highly aware and quite convinced of the benefits of PROMIS CATs in order to outweigh such a drawback. Second, in addition to CAT, the PROMIS item banks can be administered as short forms. A short form is an IRT-based instrument with a small number of fixed items that are specifically chosen for their discriminative ability and coverage of the measured latent trait. PROMIS developed four short forms for the Depression item bank with 4, 6 or 8 items (i.e., PROMIS Short Form v1.0 – Depression 4a, 6a, 8a and 8b), and four short forms for the Anxiety item bank with 4, 6, 7 or 8 items (i.e., PROMIS Short Form v1.0 – Anxiety 4a, 6a, 7a and 8a). Similar to many traditional instruments, these short forms are free of charge and do not require access to CAT technology. However, they also measure somewhat less precise than CAT instruments because the administration is not tailored to the respondent's latent trait level. Finally, as noted in section 7.1.4, PROsetta Stone developed and applied methods to link PROMIS instruments with traditional instruments measuring the same construct on a common, standardized metric (Kaat, Newcomb, Ryan, & Mustanski, 2017). As a result, test users may not feel the need to start using PROMIS CATs when their goal is to learn from other PROMIS CAT users, because the currently used instruments are familiar and could "simply" be linked on the same metric. On the other hand, when test users are interested in transitioning to PROMIS CATs, this methodology can make the implementation process somewhat easier as the scores already assessed in ongoing treatments can be transformed to the T-score metric of the PROMIS CATs. That being said, a similar initiative states on its website (www.common-metrics.org) that "little is known about the validity of these common metrics and they have rarely been validated so far in external samples". Until more information is available on this topic, using a standard set of instruments may be the best available option yet to learn from each other's practice.

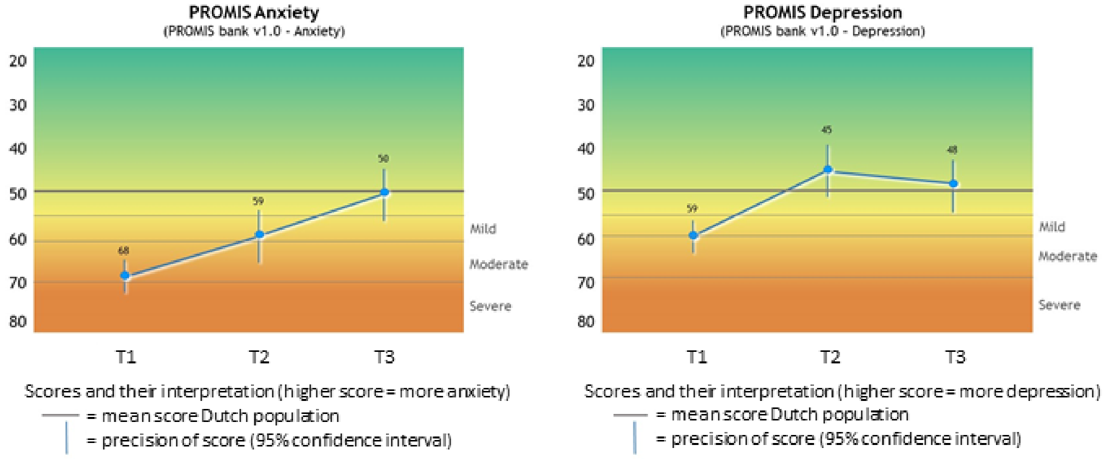
Following this line of reasoning, initiatives concerned with the standardization of measurement instruments may help raise awareness on CAT instruments. For example, the Linnean Initiative is a Dutch network of more than 350 patient representatives, healthcare providers, researchers, IT experts, and consultants who are committed to accelerating the implementation of value-based care in the Netherlands (www.linnean.nl). In their national guideline, they recommended several PROMIS instruments to evaluate treatment outcomes with patients, including DF PROMIS instruments (CATs and short forms) for Depression and Anxiety. This may help convince clinicians to start using CAT instruments in their practice. By contrast, the International Consortium for Health Outcomes Measures (ICHOM; www.ichom.org) also included several PROMIS instruments in their standard sets, but CAT instruments are not among them because many countries do not yet have access to CAT technology. ICHOM collaborates with patients and healthcare professionals to define and measure patient-reported outcomes for the improvement of quality and value of care. By 2021, their standard sets for Depression and Anxiety included the instruments World Health Organization Disability Assessment 2.0 (WHODAS 2.0), PHQ-9, GAD-7, and several other instruments for specific anxiety disorders. Consequently, this may dissuade clinicians from using CAT instruments in their practice.

7.3.4 Developing useful feedback tools

The fourth point of attention for the implementation of CAT instruments in Dutch clinical practice is that clinicians have access to useful feedback tools to discuss the scores with their patients. This requires a clear understanding of the meaning of a test result by the clinician and clear communication to the patient. The use of a common metric, such as the T score used by PROMIS instruments, may be helpful in this regard (de Beurs, Flens, & Williams, 2019).

In Figure 7.1, an example is shown of the feedback tool that the DF PROMIS National Center intends to provide to test users for their CAT instruments (Elsman et al., 2022). In this figure, the T-scores are shown on the vertical axis and the dates of assessment (T1, T2, and T3) on the horizontal axis; the blue dots and lines represent the patient's progress over time. Additionally, the feedback tool contains information that may help test users with the interpretation of the T-scores. First, the measurement precision with a 95% confidence interval is shown for each T-score using a blue vertical line. Second, Dutch T-score thresholds for mild (55), moderate (60), and severe symptoms (70) are shown with fixed grey horizontal lines as well as with a gradual change of colors on the vertical axis. Finally, the Dutch general population mean of 50 is represented with a fixed black horizontal line.

Figure 7.1 Feedback tools for the DF PROMIS CATs for Anxiety and Depression.



In addition, clinicians may benefit from several other tools to discuss the CAT scores with their patients. First, a threshold for clinically significant change may help test users to evaluate whether a patient changed from a clinical to a general population score (Jacobson & Truax, 1991). Second, an increasing number of thresholds, such as five or seven (e.g., very low, low, below average, average, above average, high, and very high), may help test users identify smaller changes that can be meaningful to patients (de Beurs, Flens, & Williams, 2019). Third, statistics that provide specific information for interpreting change scores, such as the reliable change index (RCI; Jacobson & Truax, 1991), the smallest detectable change (SDC; de Vet, Terwee, Mokkink, & Knol, 2011), and the minimal important change (MIC; Terwee et al., 2021), may help in this regard as well. Reliable change is defined as a change in scores that may not have occurred due to random measurement error alone; SDC is a measure of the variation in a scale due to measurement error, meaning that a change score is only considered to represent real change when it is larger than the SDC; MIC is defined as the smallest measured change score patients perceive as important. Finally, test users may benefit from an explanation of the results in a few lines of text. For example, the feedback tool of the US PROMIS Assessment Center explains to what degree a person’s T-score is higher or lower than that of other persons from the general population, the corresponding gender group, and the corresponding age group (i.e., the percentile score; Crawford & Garthwaite, 2009).

7.4 Closing words

In this thesis, the DF PROMIS adult v1.0 item banks for Depression and Anxiety were evaluated for CAT administration in the Dutch clinical (i.e., patients with common mental disorders in ambulatory mental health care) and general population. Based on a wide collection of psychometric properties, it was demonstrated that the CAT instruments measure efficiently, reliably, validly, and responsively in both populations. Also, the DF PROMIS CATs were shown to be a modest improvement over matching BSI scales, which was expected based on previous research regarding the US PROMIS CATs for Depression and Anxiety. Finally, it was explained that the DF PROMIS CATs have additional benefits compared to traditional instruments.

In order to implement CAT instruments in clinical practice, several points of attention were identified in this discussion. Of these points, probably the most pressing is the implementation of MBC. In the Netherlands (and many other countries), MBC is still in its infancy in Dutch mental health care, even though many studies have demonstrated positive effects on the overall quality of care. The main reason for this is that implementing MBC is complex and requires several challenges to be overcome. Only one of these challenges is the availability of brief (reliable) self-report measures, such as CAT instruments. Other challenges require sufficient attention as well (e.g., active leadership and clinical-engagement). Here may lie a great opportunity for CAT providers should they expand their services to help mental health providers implement MBC. Meanwhile, we can begin to work towards the next generation of CAT instruments that increases the benefits of measurement even further. In this discussion, several future directions were mentioned to accomplish this. As a result, CAT instruments may eventually become the new standard for evaluating patients’ treatment in the Netherlands. This, in turn, may stimulate MBC, which may result in more effective and efficient treatment of patients in general.

References

Aboraya, A., Rankin, E., France, C., El-Missiry, A., & John, C. (2006). The reliability of psychiatric diagnosis revisited: The clinician's guide to improve the reliability of psychiatric diagnosis. *Psychiatry (Edgmont), 3*(1), 41.

Agresti, A., & Kateri, M. (2011). Categorical data analysis. In *International encyclopedia of statistical science* (pp. 206-208). Berlin, Germany: Springer.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1974). *Standards for educational & psychological tests*.

Amtmann, D., Kim, J., Chung, H., Bamer, A. M., Askew, R. L., Wu, S., ... & Johnson, K. L. (2014). Comparing CESD-10, PHQ-9, and PROMIS depression instruments in individuals with multiple sclerosis. *Rehabilitation psychology, 59*(2), 220.

Arrindell, W. A. M., & Ettema, H. (1981). Dimensionele structuur, betrouwbaarheid en validiteit van de Nederlandse bewerking van de Symptom Checklist (SCL-90): Gegevens gebaseerd op een fobisch en een" normale" populatie. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden*.

Babcock, B., & Weiss, D. J. (2013). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement?. *Journal of Computerized Adaptive Testing, 1,* 1-18.

Bass, M., Morris, S., & Neapolitan, R. (2015). Utilizing multidimensional computer adaptive testing to mitigate burden with patient reported outcomes. In *AMIA Annual Symposium Proceedings* (Vol. 2015, p. 320). American Medical Informatics Association.

Baxter, A. J., Scott, K. M., Vos, T., & Whiteford, H. A. (2013). Global prevalence of anxiety disorders: a systematic review and meta-regression. *Psychological medicine, 43*(5), 897-910.

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, *80*(2), 317-340.

Beck, A., & Beamesderfer, A. (1974). Assessment of depression: The depression inventory. In P. Pichot (Ed.), *Psychological measurements in psychopharmacology: Modern problems in pharmacopsychiatry* (Vol. 7, pp. 151–169). Oxford, England: Karger.

Becker, J., Fliege, H., Kocalevent, R. D., Bjorner, J. B., Rose, M., Walter, O. B., & Klapp, B. F. (2008). Functioning and validity of A Computerized Adaptive Test to measure anxiety (A-CAT). *Depression and anxiety*, *25*(12), E182-E194.

Bedford, A. (1997). On Clark-Watson's tripartite model of anxiety and depression. *Psychological Reports*, *80*(1), 125-126.

Beleckas, C. M., Prather, H., Guattery, J., Wright, M., Kelly, M., & Calfee, R. P. (2018). Anxiety in the orthopedic patient: using PROMIS to assess mental health. *Quality of Life Research*, *27*(9), 2275-2282.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238.

Bernstein, I. H., & Nunnally, J. C. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Bevans, K. B., Gardner, W., Pajer, K. A., Becker, B., Carle, A., Tucker, C. A., & Forrest, C. B. (2018). Psychometric evaluation of the PROMIS® pediatric psychological and physical stress experiences measures. *Journal of pediatric psychology*, *43*(6), 678-692.

Bijl, R. V., Ravelli, A., & van ZESSEN, G. (1998). Prevalence of psychiatric disorder in the general population: results of The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Social psychiatry and psychiatric epidemiology*, *33*(12), 587-595.

Bijl, R. V., Van Zessen, G., Ravelli, A., De Rijk, C., & Langendoen, Y. (1998). The Netherlands mental health survey and incidence study (NEMESIS): objectives and design. *Social psychiatry and psychiatric epidemiology*, *33*(12), 581-586.

Bock, D. (1997). A brief history of item theory. *Educational measurement: issues and practice*, *16*(4), 21-33.

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.

Booth-Kewley, S., Larson, G. E., & Miyoshi, D. K. (2007). Social desirability effects on computerized and paper-and-pencil questionnaires. *Computers in human behavior*, *23*(1), 463-477.

Borsboom, D. (2006). When does measurement invariance matter?. *Medical care*, *44*(11), S176-S181.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). Measuring individual significant change on the Beck Depression Inventory-II through IRT-based statistics. *Psychotherapy Research*, *23*(5), 489-501.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Carlier, I. V., Meuldijk, D., Van Vliet, I. M., Van Fenema, E., Van der Wee, N. J., & Zitman, F. G. (2012). Routine outcome monitoring and feedback on physical or mental health status: evidence and theory. *Journal of evaluation in clinical practice*, *18*(1), 104-110.

Carlier, I. V., Kovács, V., van Noorden, M. S., van der Feltz-Cornelis, C., Mooij, N., Schulte-van Maaren, Y. W., ... & Giltay, E. J. (2017). Evaluating the responsiveness to therapeutic change with routine outcome monitoring: A Comparison of the Symptom Questionnaire-48 (SQ-48) with the Brief Symptom Inventory (BSI) and the Outcome Questionnaire-45 (OQ-45). *Clinical Psychology & Psychotherapy*, *24*(1), 61-71.

Carlier, I., Schulte-Van Maaren, Y., Wardenaar, K., Giltay, E., Van Noorden, M., Vergeer, P., & Zitman, F. (2012). Development and validation of the 48-item Symptom Questionnaire (SQ-48) in patients with depressive, anxiety and somatoform disorders. *Psychiatry research*, *200*(2-3), 904-910.

Carlier, I. V., van Eeden, W. A., de Jong, K., Giltay, E. J., van Noorden, M. S., van der Feltz-Cornelis, C., ... & van Hemert, A. M. (2019). Testing for response shift in treatment evaluation of change in self-reported psychopathology amongst secondary psychiatric care outpatients. *International Journal of Methods in Psychiatric Research*, *28*(3), e1785.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., ... & PROMIS Cooperative Group. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of clinical epidemiology*, *63*(11), 1179-1194.

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

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

Chalmers, R. P. (2015). *mirtCAT: Computerized adaptive testing with multidimensional item response theory*. Retrieved from http://CRAN.R-project.org/package=mirtCAT

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological methods & research*, *36*(4), 462-494.

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

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

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *71*(1), 37-53.

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *Journal of abnormal psychology*, *100*(3), 316-336.

Clarke, D. M., & Kissane, D. W. (2002). Demoralization: its phenomenology and importance. *Australian & New Zealand Journal of Psychiatry*, *36*(6), 733-742.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Cook, T. D., & Campbell, D. T. (1959). *Quasi-experiments: Design and analysis issues for field settings*. Rand McNally.

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. 351). Boston: Houghton Mifflin.

Cook, K. F., O'Malley, K. J., & Roddey, T. S. (2005). Dynamic assessment of health outcomes: time to let the CAT out of the bag?. *Health services research*, *40*(5 Pt 2), 1694-1711.

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical care*, *44*, S115-S123.

Crawford, J. R., & Garthwaite, P. H. (2009). Percentiles please: The case for expressing neuropsychological test scores and accompanying confidence limits as percentile ranks. *The Clinical Neuropsychologist*, *23*(2), 193-204.

Crins, M. H., Roorda, L. D., Smits, N., De Vet, H. C., Westhovens, R., Cella, D., ... & Terwee, C. B. (2015). Calibration and validation of the Dutch-Flemish PROMIS pain interference item bank in patients with chronic pain. *PloS one*, *10*(7), e0134094.

Crins, M. H., Terwee, C. B., Klausch, T., Smits, N., de Vet, H. C., Westhovens, R., ... & Roorda, L. D. (2017). The Dutch–Flemish PROMIS Physical Function item bank exhibited strong psychometric properties in patients with chronic pain. *Journal of clinical epidemiology, 87*, 47-58.

Crins, M. H. P., Roorda, L. D., Smits, N., De Vet, H. C. W., Westhovens, R., Cella, D., ... & Terwee, C. B. (2016). Calibration of the Dutch-Flemish PROMIS Pain Behavior item bank in patients with chronic pain. *European Journal of Pain*, *20*(2), 284-296.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

de Beurs, E., Barendregt, M., Flens, G., van Dijk, E., Huijbrechts, I., & Meerding, W. J. (2012). Equivalentie in responsiviteit van veel gebruikte zelfrapportage meetinstrumenten in de geestelijke gezondheidszorg [Equivalence in responsiveness of commonly used self-report questionnaires in mental health]. *Maandblad voor de Geestelijke Volksgezondheid, 67*, 259-264.

de Beurs, E., Barendregt, M., & Warmerdam, L. (2017). *Behandeluitkomsten: Bron voor Kwaliteitsbeleid in de GGZ*. Boom Uitgevers Amsterdam.

de Beurs, E., Carlier, I. V., & van Hemert, A. M. (2019). Approaches to denote treatment outcome: Clinical significance and clinical global impression compared. *International journal of methods in psychiatric research*, *28*(4), e1797.

de Beurs, E., den Hollander-Gijsman, M. E., Helmich, S., & Zitman, F. G. (2007). The tripartite model for assessing symptoms of anxiety and depression: Psychometrics of the Dutch version of the mood and anxiety symptoms questionnaire. *Behaviour Research and Therapy*, *45*(7), 1609-1617.

de Beurs, E., den Hollander-Gijsman, M. E., van Rood, Y. R., Van der Wee, N. J., Giltay, E. J., van Noorden, M. S., ... & Zitman, F. G. (2011). Routine outcome monitoring in the Netherlands: practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical psychology & psychotherapy, 18*(1), 1-12.

de Beurs, E., Flens, G., & Williams, G. (2019). Meetresultaten interpreteren in de klinische psychologie: een aantal voorstellen. *De psycholoog, 54*, 10 – 23.

de Beurs, E., Warmerdam, E. H., Oudejans, S. C. C., Spits, M., Dingemanse, P., De Graaf, S. D. D., ... & van Son, G. E. (2018). Treatment outcome, duration, and costs: A comparison of performance indicators using data from eight mental health care providers in the Netherlands. *Administration and Policy in Mental Health and Mental Health Services Research*, *45*(2), 212-223.

de Beurs, E. & Zitman, F. G. (2005). De Brief Symptom Inventory (BSI): De betrouwbaarheid en validiteit van een handzaam alternatief voor de SCL-90 [The Brief Symptom Inventory: Reliability and validity of a handy alternative for the SCL-90]. *Maandblad Geestelijke Volksgezondheid*, *61*, 120-141.

de Graaf, R., Ten Have, M., van Gool, C., & van Dorsselaer, S. (2012). Prevalence of mental disorders and trends from 1996 to 2009. Results from the Netherlands Mental Health Survey and Incidence Study-2. *Social psychiatry and psychiatric epidemiology*, *47*(2), 203-213.

de Graaf, R. O. N., Ten Have, M., & van Dorsselaer, S. (2010). The Netherlands mental health survey and incidence study-2 (NEMESIS-2): Design and methods. *International journal of methods in psychiatric research*, *19*(3), 125-141.

de Jong, K., Conijn, J. M., Gallagher, R. A., Reshetnikova, A. S., Heij, M., & Lutz, M. C. (2021). Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: A multilevel meta-analysis. *Clinical Psychology Review*, 102002.

de Jong, K., Nugter, M. A., Polak, M. G., Wagenborg, J. E., Spinhoven, P., & Heiser, W. J. (2007). The Outcome Questionnaire (OQ-45) in a Dutch population: A cross-cultural validation. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, *14*(4), 288-301.

de Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: a practical guide*. Cambridge university press.

Deng, G., Jiang, C., & Li, Y. X. (2012). Clinical utility of the mood and anxiety symptom questionnaire in a Chinese sample of patients with pancreatic cancer. *Gastroenterology Nursing*, *35*(3), 193-198.

Derogatis, L. R., Lipman, R. S., & Covi, L. (1973). SCL-90: an outpatient psychiatric rating scale–preliminary report. *Psychopharmacol Bull, 9*, 13-28.

Dieperink, C. J., Pijl, Y. J., Mulder, N., Van Os, J., & Drukker, M. (2008). Langdurig zorgafhankelijke patiënten in de ggz: samenhang met verstedelijking. *Tijdschrift voor Psychiatrie*, 761-769.

Dillman, D. A., Sinclair, M. D., & Clark, J. R. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public opinion quarterly*, *57*(3), 289-304.

Drenth, P. J. D., & Sijtsma, K. (2005). *Testtheorie: Inleiding in de theorie van de psychologische test en zijn toepassingen.* Bohn Stafleu Van Loghum.

Elsman, E. B., Flens, G., de Beurs, E., Roorda, L. D., & Terwee, C. B. (2022). Towards standardization of measuring anxiety and depression: Differential item functioning for language and Dutch reference values of PROMIS item banks. *Plos one*, *17*(8), e0273287.

Embretson, S., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Flens, G., Smits, N., Carlier, I., van Hemert, A. M., & de Beurs, E. (2016). Simulating computer adaptive testing with the Mood and Anxiety Symptom Questionnaire. *Psychological Assessment*, *28*(8), 953-962.

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., & de Beurs, E. (2017). Development of a computer adaptive test for depression based on the Dutch-Flemish version of the PROMIS item bank. *Evaluation & the health professions*, *40*(1), 79-105.

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., Spinhoven, P., & de Beurs, E. (2019). Development of a computerized adaptive test for anxiety based on the Dutch–Flemish version of the PROMIS item bank. *Assessment*, *26*(7), 1362-1374.

Flens, G., Smits, N., Terwee, C. B., Pijck, L., Spinhoven, P., & de Beurs, E. (2021). Practical significance of longitudinal measurement invariance violations in the Dutch–Flemish PROMIS item Banks for Depression and Anxiety: an illustration with ordered-categorical data. *Assessment*, *28*(1), 277-294.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of life Research*, *14*(10), 2277-2291.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466 - 491.

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: an illustration of longitudinal measurement invariance. *Psychological Assessment*, *25*(2), 520–531.

Forbey, J. D., & Ben-Porath, Y. S. (2007). A comparison of the MMPI-2 restructured clinical (RC) and clinical scales in a substance abuse treatment sample. *Psychological Services, 4*(1), 46-58.

Forkmann, T., Boecker, M., Norra, C., Eberle, N., Kircher, T., Schauerte, P., … Wirtz, M. (2009). Development of an item bank for the assessment of depression in persons with mental illnesses and physical diseases using Rasch analysis. *Rehabilitation Psychology, 54*(2), 186-197.

Forrest, C. B., Devine, J., Bevans, K. B., Becker, B. D., Carle, A. C., Teneralli, R. E., ... & Ravens-Sieberer, U. (2018). Development and psychometric evaluation of the PROMIS Pediatric Life Satisfaction item banks, child-report, and parent-proxy editions. *Quality of Life Research, 27*(1), 217-234.

Fortney, J. C., Unützer, J., Wrenn, G., Pyne, J. M., Smith, G. R., Schoenbaum, M., & Harbin, H. T. (2018). A tipping point for measurement-based care. *Focus, 16*(3), 341-350.

Frank, J. D. (1974). Psychotherapy: The restoration of morale. *American Journal of Psychiatry, 131*(3), 271-274.

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time . . . or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment, 28*(11), 1354-1367.

Fries, J., Rose, M., & Krishnan, E. (2011). The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and Internet administration. *The Journal of rheumatology, 38*(8), 1759-1764.

Fries, J. F., Krishnan, E., Rose, M., Lingala, B., & Bruce, B. (2011). Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. *Arthritis research & therapy, 13*(5), 1-8.

Gardner, W., Shear, K., Kelleher, K. J., Pajer, K. A., Mammen, O., Buysse, D., & Frank, E. (2004). Computerized adaptive measurement of depression: a simulation study. *BMC psychiatry, 4*(1), 1-11.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *Archives of general psychiatry, 69*(11), 1104-1112.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2014). Development of the CAT-ANX: a computerized adaptive test for anxiety. *American Journal of Psychiatry, 171*(2), 187-194.

Gliem J, & Gliem R. (2003). Calculating, interpreting, and reporting Cronbach’s alpha reliability coefficient for Likert-type scales. In 2003 Midwest research to practice conference in adult, continuing and community education. Columbus, OH.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational measurement*, *21*(4), 347-360.

Greenhalgh, J., Gooding, K., Gibbons, E., Dalkin, S., Wright, J., Valderas, J., & Black, N. (2018). How do patient reported outcome measures (PROMs) support clinician-patient communication and patient care? A realist synthesis. *Journal of patient-reported outcomes*, *2*(1), 1-28.

Guo, T., Xiang, Y. T., Xiao, L. E., Hu, C. Q., Chiu, H. F., Ungvari, G. S., ... & Wang, G. (2015). Measurement-based care versus standard care for major depression: a randomized controlled trial with blind raters. *American Journal of Psychiatry*, *172*(10), 1004-1013.

Hambleton, R. K. (1988). *Principles and selected applications of Item Response Theory* (3rd ed.). New York: American Council on Education.

Hare, D. L., Toukhsati, S. R., Johansson, P., & Jaarsma, T. (2014). Depression and cardiovascular disease: a clinical review. *European heart journal*, *35*(21), 1365-1372.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical Learning: Data mining, inference and prediction.* New York: Springer.

Haverman, L., Grootenhuis, M. A., Raat, H., van Rossum, M. A., van Dulmen-den Broeder, E., Hoppenbrouwers, K., ... & Terwee, C. B. (2016). Dutch–Flemish translation of nine pediatric item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS)®. *Quality of Life Research*, *25*(3), 761-765.

Hayslett, M. M., & Wildemuth, B. M. (2004). Pixels or pencils? The relative effectiveness of Web-based versus paper surveys. *Library & Information Science Research*, *26*(1), 73-93.

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

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

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

Irwin, D. E., Stucky, B., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J. S., ... & DeWalt, D. A. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research*, *19*(4), 595-607.

Jabrayilov, R., Emons, W. H., de Jong, K., & Sijtsma, K. (2017). Longitudinal measurement invariance of the Dutch Outcome Questionnaire-45 in a clinical sample. *Quality of Life Research*, *26*(6), 1473-1481.

Jabrayilov, R., Emons, W. H., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, *40*(8), 559-572.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior therapy*, *15*(4), 336-352.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12-19.

Jakob, T., Nagl, M., Gramm, L., Heyduck, K., Farin, E., & Glattacker, M. (2017). Psychometric properties of a German translation of the PROMIS® depression item bank. *Evaluation & the health professions*, *40*(1), 106-120.

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Kaat, A. J., Newcomb, M. E., Ryan, D. T., & Mustanski, B. (2017). Expanding a common metric for depression reporting: linking two scales to PROMIS® depression. *Quality of Life Research*, *26*(5), 1119-1128.

Kaplan, D. (1989). A study of the sampling variability and z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research*, *24*(1), 41-57.

Kays, K., Gathercoal, K., & Buhrow, W. (2012). Does survey format influence self-disclosure on sensitive question items?. *Computers in Human Behavior*, *28*(1), 251-256.

Keogh, E., & Reidy, J. (2000). Exploring the factor structure of the Mood and Anxiety Symptom Questionnaire (MASQ). *Journal of personality assessment*, *74*(1), 106-125.

Khanna, D., Krishnan, E., Dewitt, E. M., Khanna, P. P., Spiegel, B., & Hays, R. D. (2011). The future of measuring patient-reported outcomes in rheumatology: Patient-Reported Outcomes Measurement Information System (PROMIS). *Arthritis care & research, 63 3*(Suppl 11), S486-S490.

Kim, S., Moses, T., & Yoo, H. H. (2015). Effectiveness of item response theory (IRT) proficiency estimation methods under adaptive multistage testing. *ETS Research Report Series*, *2015*(1), 1-19.

Kim, S. H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement*, *44*(2), 93-116.

Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., ... & Klapp, B. F. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology*, *62*(3), 278-287.

Kose, I. A., & Demirtasli, N. C. (2012). Comparison of unidimensional and multidimensional models based on item response theory in terms of both variables of test length and sample size. *Procedia-Social and Behavioral Sciences*, *46*, 135-140.

Kroenke, K., Baye, F., & Lourens, S. G. (2019). Comparative responsiveness and minimally important difference of common anxiety measures. *Medical care*, *57*(11), 890-897.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, *16*(9), 606-613.

Kudel, I., Pona, A., Cox, S., Szoka, N., Tabone, L., & Brode, C. (2019). Psychometric properties of NIH PROMIS® instruments in bariatric surgery candidates. *Health Psychology*, *38*(5), 359.

Lambert, M. J. (2010). *Prevention of treatment failure: The use of measuring, monitoring, and feedback in clinical practice*. Washington, DC: American Psychological Association.

Lange, A., Schrieken, B., van de Ven, J., & Blankers, M. (2000). De Korte Klachten Lijst (KKL). [The Short Complaint List]. *Directieve therapie*, *20*(4), 181-185.

Lecrubier, Y., Sheehan, D. V., Weiller, E., Amorim, P., Bonora, I., Sheehan, K. H., ... & Dunbar, G. C. (1997). The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. *European psychiatry*, *12*(5), 224-231.

Lee, S. A., Kim, K. H., & Cho, S. M. (2015). Validation of the Mood and Anxiety Symptom Questionnaire in Korean Adolescents. *Psychiatry investigation*, *12*(2), 218-226.

Lewis, C. C., Boyd, M., Puspitasari, A., Navarro, E., Howard, J., Kassab, H., ... & Kroenke, K. (2019). Implementing measurement-based care in behavioral health: a review. *JAMA psychiatry*, *76*(3), 324-335.

Lewis, C. C., Scott, K., Marti, C. N., Marriott, B. R., Kroenke, K., Putz, J. W., ... & Rutkowski, D. (2015). Implementing measurement-based care (iMBC) for depression in community mental health: a dynamic cluster randomized trial study protocol. *Implementation Science*, *10*(1), 1-14.

Leyro, T. M., Zvolensky, M. J., & Bernstein, A. (2010). Distress tolerance and psychopathological symptoms and disorders: a review of the empirical literature among adults. *Psychological bulletin*, *136*(4), 576.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior research methods*, *48*(3), 936-949.

Lin, A., Yung, A. R., Wigman, J. T., Killackey, E., Baksheev, G., & Wardenaar, K. J. (2014). Validation of a short adaptation of the Mood and Anxiety Symptoms Questionnaire (MASQ) in adolescents and young adults. *Psychiatry research*, *215*(3), 778-783.

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological methods*, *22*(3), 486.

Liu, Y., & West, S. G. (2018). Longitudinal measurement non-invariance with ordered-categorical indicators: How are the parameters in second-order latent linear growth models affected?. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 762-777.

Lizzio, V. A., Blanchett, J., Borowsky, P., Meldau, J. E., Verma, N. N., Muh, S., ... & Makhni, E. C. (2019). Feasibility of PROMIS CAT administration in the ambulatory sports medicine clinic with respect to cost and patient compliance: a single-surgeon experience. *Orthopaedic journal of sports medicine*, *7*(1), 2325967118821875.

Lloyd, C. E., Dyer, P. H., & Barnett, A. H. (2000). Prevalence of symptoms of depression and anxiety in a diabetes clinic population. *Diabetic medicine*, *17*(3), 198-202.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Welsley Publishing Company.

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour research and therapy*, *33*(3), 335-343.

Löwe, B., Spitzer, R. L., Williams, J. B., Mussell, M., Schellberg, D., & Kroenke, K. (2008). Depression, anxiety and somatization in primary care: syndrome overlap and functional impairment. *General hospital psychiatry*, *30*(3), 191-199.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84-99.

Magasi, S., Ryan, G., Revicki, D., Lenderking, W., Hays, R. D., Brod, M., ... & Cella, D. (2012). Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. *Quality of Life Research*, *21*(5), 739-746.

Mancheño, J. J., Cupani, M., Gutiérrez-López, M., Delgado, E., Moraleda, E., Cáceres-Pachón, P., ... & Rojas, Ó. M. L. (2018). Classical test theory and item response theory produced differences on estimation of reliable clinical index in World Health Organization Disability Assessment Schedule 2.0. *Journal of clinical epidemiology, 103*, 51-59.

Marasini, D., Quatto, P., & Ripamonti, E. (2016). Assessing the inter-rater agreement for ordinal data through weighted indexes. *Statistical methods in medical research*, *25*(6), 2611-2633.

Marcus, M., Yasamy, M. T., van Ommeren, M., Chisholm, D., & Saxena, S. (2012). Depression: A global public health concern. *WHO Department of Mental Health and Substance Abuse*, *1*, 6-8.

Martin-Cook, K., Palmer, L., Thornton, L., Rush, A. J., Tamminga, C. A., & Ibrahim, H. M. (2021). Setting Measurement-Based Care in Motion: Practical Lessons in the Implementation and Integration of Measurement-Based Care in Psychiatry Clinical Practice. *Neuropsychiatric disease and treatment, 17*, 1621-1631.

Maruyama, G., & Ryan, C. S. (2014). *Research methods in social relations*. John Wiley & Sons.

McAleavey, A. A., Nordberg, S. S., & Moltu, C. (2021). Initial quantitative development of the Norse Feedback system: a novel clinical feedback system for routine mental healthcare. *Quality of Life Research*, *30*(11), 3097-3115.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*(4), 728.

Meade, A. W., & Lautenschlager, G. J. (2004). Same question, different answers: CFA and two IRT approaches to measurement invariance. In *19th Annual Conference of the Society for Industrial and Organizational Psychology*, Chicago.

Metz, M. J., Veerbeek, M. A., Twisk, J. W., van der Feltz-Cornelis, C. M., de Beurs, E., & Beekman, A. T. (2019). Shared decision-making in mental health care using routine outcome monitoring: results of a cluster randomised-controlled trial. *Social psychiatry and psychiatric epidemiology*, *54*(2), 209-219.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton; Berlin, DeGruyter.

Mokkink, L. B., De Vet, H. C., Prinsen, C. A., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*(5), 1171-1179.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., ... & de Vet, H. C. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of clinical epidemiology*, *63*(7), 737-745.

Muramatsu, K., Kamijima, K., Yoshida, M., Otsubo, T., Miyaoka, H., Muramatsu, Y., & Gejyo, F. (2007). The patient health questionnaire, Japanese version: validity according to the mini-international neuropsychiatric interview–plus. *Psychological reports*, *101*(3), 952-960.

Newton, P., & Shaw, S. (2014). *Validity in educational & psychological assessment*. London: Sage.

O'Connell, N. S., Dai, L., Jiang, Y., Speiser, J. L., Ward, R., Wei, W., ... & Gebregziabher, M. (2017). Methods for analysis of pre-post data in clinical research: a comparison of five common methods. *Journal of biometrics & biostatistics*, *8*(1), 1-8.

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, *14*(3), 587-598.

Oosterveld, P., Vorst, H. C., & Smits, N. (2019). Methods for questionnaire design: a taxonomy linking procedures to test goals. *Quality of Life Research*, *28*(9), 2501-2512.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied psychological measurement*, *24*(1), 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*(4), 289-298.

Osman, A., Freedenthal, S., Gutierrez, P. M., Wong, J. L., Emmerich, A., & Lozano, G. (2011). The Anxiety Depression Distress Inventory-27 (ADDI-27): a short version of the Mood and Anxiety Symptom Questionnaire-90. *Journal of Clinical Psychology*, *67*(6), 591-608.

Ostini, R., Finkelman, M., & Nering, M. (2015). Selecting Among Polytomous IRT Models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 285-304). New York: Routledge/Taylor & Francis Group.

Oude Voshaar, M. A., ten Klooster, P. M., Glas, C. A., Vonkeman, H. E., Taal, E., Krishnan, E., ... & van de Laar, M. A. (2014). Calibration of the PROMIS physical function item bank in Dutch patients with rheumatoid arthritis. *PloS one*, *9*(3), e92367.

Paap, M. C., Born, S., & Braeken, J. (2019). Measurement efficiency for fixed-precision multidimensional computerized adaptive tests: Comparing health measurement and educational testing using example banks. *Applied Psychological Measurement*, *43*(1), 68-83.

Papuga, M. O., Dasilva, C., McIntyre, A., Mitten, D., Kates, S., & Baumhauer, J. F. (2018). Large-scale clinical implementation of PROMIS computer adaptive testing with direct incorporation into the electronic medical record. *Health Systems*, *7*(1), 1-12.

Paz, S. H., Spritzer, K. L., Morales, L. S., & Hays, R. D. (2013). Evaluation of the patient-reported outcomes information system (PROMIS®) Spanish-language physical functioning items. *Quality of Life Research*, *22*(7), 1819-1830.

Penfield, R. D., & Bergeron, J. M. (2005). Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Applied Psychological Measurement*, *29*(3), 218-233.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., & PROMIS Cooperative Group. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. *Assessment, 18*(3), 263-283.

Pilkonis, P. A., Yu, L., Dodds, N. E., Johnston, K. L., Lawrence, S. M., & Daley, D. C. (2016). Validation of the alcohol use item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS®). *Drug and alcohol dependence, 161*, 316-322.

Pilkonis, P. A., Yu, L., Dodds, N. E., Johnston, K. L., Maihoefer, C. C., & Lawrence, S. M. (2014). Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study. *Journal of psychiatric research, 56*, 112-119.

Porter, M. E. (2009). A strategy for health care reform—toward a value-based system. *New England Journal of Medicine*, *361*(2), 109-112.

Prinsen, C. A., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., De Vet, H. C., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*(5), 1147-1157.

R Core Team (2014, 2015, 2017, 2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of educational statistics*, *4*(3), 207-230.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied psychological measurement*, *9*(4), 401-412.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., ... & Liu, H. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical care*, *45*, S22-S31.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*(1), 19-31.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual review of clinical psychology*, *5*, 27-48.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of educational Measurement*, *27*(2), 133-144.

Revelle, W. (2013). *psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psych

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods*, *17*(3), 354.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and human behavior*, *29*(5), 615-620.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses, *Journal of Statistical Software*, *17*, 1-25.

Rosseel, Y. (2012). lavaan: An R Package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36.

Rush, A. J., & Thase, M. E. (2018). Improving depression outcome by patient-centered medical management. *American Journal of Psychiatry, 175*(12), 1187-1198.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded responses. *Psychometrika Monograph Supplement, 34, 100*.

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167-180.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In *Innovations in multivariate statistical analysis* (pp. 233-247). Springer, Boston, MA.

Sawatzky, R., Kwon, J. Y., Barclay, R., Chauhan, C., Frank, L., van den Hout, W. B., ... & Sprangers, M. A. (2021). Implications of response shift for micro-, meso-, and macro-level healthcare decision-making using results of patient-reported outcome measures. *Quality of Life Research*, *30*, 3343–3357.

Schalet, B. D., Pilkonis, P. A., Yu, L., Dodds, N., Johnston, K. L., Yount, S., ... & Cella, D. (2016). Clinical validity of PROMIS depression, anxiety, and anger across diverse clinical samples. *Journal of clinical epidemiology, 73*, 119-127.

Scott, K., & Lewis, C. C. (2015). Using measurement-based care to enhance any treatment. *Cognitive and behavioral practice*, *22*(1), 49-59.

Seidel, J. A., Miller, S. D., & Chow, D. L. (2014). Effect size calculations for the clinician: Methods and comparability. *Psychotherapy Research*, *24*(4), 470-484.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., ... & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of clinical psychiatry, 59*(20), 22-33.

Simms, L. J., Grös, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and anxiety, 25*(7), E34-E46.

Simms, L. J., Prisciandaro, J. J., Krueger, R. F., & Goldberg, D. P. (2012). The structure of depression, anxiety and somatic symptoms in primary care. *Psychological medicine, 42*(1), 15-28.

Singer, S., Das-Munshi, J., & Brähler, E. (2010). Prevalence of mental health conditions in cancer patients in acute care—a meta-analysis. *Annals of oncology*, *21*(5), 925-930.

Smits, N. (2016). On the effect of adding clinical samples to validation studies of patient-reported outcome item banks: a simulation study. *Quality of Life Research*, *25*(7), 1635-1644.

Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry research*, *188*(1), 147-155.

Smits, N., Paap, M. C., & Böhnke, J. R. (2018). Some recommendations for developing multidimensional computerized adaptive tests for patient-reported outcomes. *Quality of Life Research*, *27*(4), 1055-1063.

Smits, N., Zitman, F. G., Cuijpers, P., den Hollander-Gijsman, M. E., & Carlier, I. V. (2012). A proof of principle for using adaptive testing in routine Outcome Monitoring: the efficiency of the Mood and Anxiety Symptoms Questionnaire-Anhedonic Depression CAT. *BMC Medical Research Methodology*, *12*(1), 1-10.

Soland, J. (2021). Is measurement noninvariance a threat to inferences drawn from randomized control trials? Evidence from empirical and simulation studies. *Applied Psychological Measurement*, 01466216211013102.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine*, *166*(10), 1092-1097.

te Poel, F., Hartmann, T., Baumgartner, S. E., & Tanis, M. (2017). A psychometric evaluation of the Dutch Short Health Anxiety Inventory in the general population. *Psychological assessment*, *29*(2), 186.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016). Measurement equivalence of the Patient Reported Outcomes Measurement Information System®(PROMIS®) Anxiety short forms in ethnically diverse groups. *Psychological test and assessment modeling*, *58*(1), 183-219.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016). Psychometric properties and performance of the Patient Reported Outcomes Measurement Information System®(PROMIS®) depression short forms in ethnically diverse groups. *Psychological test and assessment modeling*, *58*(1), 141–181.

Terwee, C. B., Crins, M. H., Roorda, L. D., Cook, K. F., Cella, D., Smits, N., & Schalet, B. D. (2021). International application of PROMIS computerized adaptive tests: US versus country-specific item parameters can be consequential for individual patient scores. *Journal of Clinical Epidemiology, 134*, 1-13.

Terwee, C. B., Crins, M. H. P., Boers, M., de Vet, H. C. W., & Roorda, L. D. (2019). Validation of two PROMIS item banks for measuring social participation in the Dutch general population. *Quality of Life Research, 28*(1), 211-220.

Terwee, C. B., Peipert, J. D., Chapman, R., Lai, J. S., Terluin, B., Cella, D., ... & Mokkink, L. B. (2021). Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. *Quality of Life Research*, *30*(10), 2729-2754.

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

Thissen, D. (1991). *MULTILOG user's guide*. Mooresville, IN: Scientific Software.

Torchiano, M. (2016). Package "effsize" [Software program]. Retrieved from https://cran.r-project.org/web/packages/effsize/effsize.pdf

Valentine, T. R., Weiss, D. M., Jones, J. A., & Andersen, B. L. (2019). Construct validity of PROMIS® Cognitive Function in cancer patients and noncancer controls. *Health Psychology*, *38*(5), 351.

van Bebber, J., Flens, G., Wigman, J. T., de Beurs, E., Sytema, S., Wunderink, L., & Meijer, R. R. (2018). Application of the Patient-Reported Outcomes Measurement Information System (PROMIS) item parameters for Anxiety and Depression in the Netherlands. *International journal of methods in psychiatric research*, *27*(4), e1744.

van de Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. *Frontiers in psychology, 6*, 1064.

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of statistical software*, *20*(11), 1-19.

van der Feltz-Cornelis, C. M., Andrea, H., Kessels, E., Duivenvoorden, H. J., Biemans, H., & Metz, M. (2014). Does routine outcome monitoring have a promising future? An investigation into the use of shared decision-making combined with ROM for patients with a combination of physical and psychiatric symptoms. *Tijdschrift voor psychiatrie*, *56*(6), 375-384.

van Noorden, M. S., Giltay, E. J., den Hollander-Gijsman, M. E., van der Wee, N. J., van Veen, T., & Zitman, F. G. (2010). Gender differences in clinical characteristics in a naturalistic sample of depressive outpatients: The Leiden Routine Outcome Monitoring Study. *Journal of affective disorders*, *125*(1-3), 116-123.

van Sonsbeek, M., Hutschemaekers, G. J., Veerman, J. W., Vermulst, A., & Tiemens, B. G. (2021). The results of clinician-focused implementation strategies on uptake and outcome of measurement-based care in general mental health care.

van Vliet, I. M., & De Beurs, E. (2007). The MINI-International Neuropsychiatric Interview. A brief structured diagnostic psychiatric interview for DSM-IV en ICD-10 psychiatric disorders. *Tijdschrift voor psychiatrie, 49*(6), 393-397.

van Widenfelt, B. M., Treffers, P. D., De Beurs, E., Siebelink, B. M., & Koudijs, E. (2005). Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families. *Clinical child and family psychology review*, *8*(2), 135-147.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods, 3*(1), 4-70.

Verdam, M. G. E., Van Ballegooijen, W., Holtmaat, C. J. M., Knoop, H., Lancee, J., Oort, F. J., ... & Sprangers, M. A. G. (2021). Re-evaluating randomized clinical trials of psychological interventions: Impact of response shift on the interpretation of trial results. *Plos one*, *16*(5), e0252035.

Vilagut, G., Forero, C. G., Adroher, N. D., Olariu, E., Cella, D., Alonso, J., & INSAyD Investigators. (2015). Testing the PROMIS® Depression measures for monitoring depression in a clinical sample outside the US. *Journal of psychiatric research, 68*, 140-150.

Wagner, L. I., Schink, J., Bass, M., Patel, S., Diaz, M. V., Rothrock, N., ... & Cella, D. (2015). Bringing PROMIS to practice: brief and precise symptom screening in ambulatory cancer care. *Cancer, 121*(6), 927-934.

Wahl, I., Rutsohn, J., Cella, D., Löwe, B., Rose, M., Brähler, E., ... & Schalet, B. (2015). Does anxiety mean the same in English and German language? Evaluation of the psychometric equivalence of the PROMIS® anxiety item bank and its German translation. *Journal of Psychosomatic Research, 6*(78), 629-630.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., & Steinberg, L. (2001). Computerized adaptive testing: A primer. *Quality Life Research, 10*(8), 733-734.

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for ‘Anxiety’(Anxiety-CAT). *Quality of Life Research*, *16*(1), 143-155.

Wang, M. (2016). *Longitudinal differential item functioning detection using bifactor models and the Wald test* (Doctoral dissertation, University of Kansas).

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(2), 109-135.

Wardenaar, K. J., van Veen, T., Giltay, E. J., de Beurs, E., Penninx, B. W., & Zitman, F. G. (2010). Development and validation of a 30-item short adaptation of the Mood and Anxiety Symptoms Questionnaire (MASQ). *Psychiatry research*, *179*(1), 101-106.

Watson, D., & Clark, L. A. (1991). *The Mood and Anxiety Symptom Questionnaire*. Iowa City, IA: University of Iowa.

Watson, D., Clark, L. A., Weber, K., Assenheimer, J. S., Strauss, M. E., & McCormick, R. A. (1995). Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. *Journal of abnormal Psychology*, *104*(1), 15-25.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child development perspectives*, *4*(1), 10-18.

Windle, C. (1954). Test-retest effect on personality questionnaires. Educational and *Psychological Measurement*, *14*(4), 617-633.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42-57.

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*(4), 1014-1045.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement*, *30*(3), 187-213.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 435-463.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, *39*(4), 561-577.

Dutch Summary

Computergestuurd Adaptief Testen in de Nederlandse Geestelijke Gezondheidszorg

Een nieuw hulpmiddel voor het meten van Depressie en Angst

Hoofdstuk 1: Introductie

In Nederland hebben meer dan 4 op de 10 personen in hun leven last gehad van een of meerdere psychische aandoeningen. De prevalentie van de meest voorkomende psychische aandoeningen is bovendien niet veel veranderd in de afgelopen decennia. Met andere woorden: psychische aandoeningen komen frequent voor in Nederland zonder duidelijke afname over tijd (Bijl, van Zessen, Ravelli, de Rijk, & Langendoen, 1998; Bijl, Ravelli, & van Zessen 1998; de Graaf, ten Have, & van Dorsselaer; 2010, 2012). Het is daarom relevant dat we blijven zoeken naar nieuwe oplossingen om dit maatschappelijke probleem te verkleinen.

Elk jaar wordt ongeveer 6% van de Nederlandse bevolking behandeld voor psychische problemen (de Beurs, Barendregt, & Warmerdam, 2017). Om patiënten effectief en efficiënt te behandelen kunnen clinici gebruik maken van verschillende hulpmiddelen. Een van die hulpmiddelen is het gebruik van zelfrapportage-meetinstrumenten. Een zelfrapportage-meetinstrument bestaat uit een reeks vragen (items) over een of meerdere constructen die wordt beantwoord door de patiënt, bijvoorbeeld over depressie of angst. Door relevante constructen periodiek te meten tijdens de behandeling worden clinici ondersteund bij de klinische besluitvorming over de aard en de ernst van de klachten van de patiënt, de invulling van de behandeling en de beëindiging daarvan, de behandeling van non-responders, en bij terugvalpreventie (de Beurs et al., 2018; Greenhalgh et al., 2018; Lambert, 2010; Lewis et al., 2015; Martin-Cook et al., 2021). Hierdoor kan de motivatie van patiënten toenemen om de behandeling voort te zetten en kunnen behandelresultaten verbeterd worden (de Jong et al., 2021; Fortney et al., 2017; Guo et al., 2015; Rush & Thase, 2018; Scott & Lewis, 2015). Wordt het gebruik van meetinstrumenten daarnaast gecombineerd met samen beslissen (d.w.z., het ondersteunen van patiënten om te participeren in de besluitvorming over de behandeling), dan kunnen behandelresultaten nog verder worden verbeterd (Metz et al., 2019; van der Feltz-Cornelis et al., 2014). Ten slotte maakt de aggregatie van meetinstrumentscores het mogelijk om de algehele kwaliteit en de waarde van zorg voor patiënten te verbeteren door groepen met elkaar te vergelijken en hiervan te leren (de Beurs et al., 2018; Porter, 2009).

Meetinstrumenten zijn het meest geschikt als hulpmiddel voor het evalueren van de behandeling als ze betrouwbaar zijn (d.w.z., de resultaten zijn consistent over replicaties; Crocker & Algina, 1986), valide (d.w.z., het te meten construct wordt adequaat gemeten; Cook & Campbell, 1979), responsief (d.w.z., verandering in de tijd wordt voldoende gedetecteerd in het gemeten construct; Mokkink et al., 2010), en efficiënt (d.w.z., het aantal items is zo klein mogelijk om de patiënt zo min mogelijk te belasten). Voor veel Nederlandse meetinstrumenten is het echter onduidelijk of ze voldoen aan deze criteria, omdat relevante psychometrische eigenschappen onvoldoende zijn onderzocht. Daarnaast is het een uitdaging om meetinstrumenten te ontwikkelen die zowel zeer betrouwbaar zijn als zeer efficiënt, omdat nagenoeg alle instrumenten uit een vaste set items bestaan. Hierdoor gaat een hogere betrouwbaarheid vaak ten koste van de efficiëntie en een hogere efficiënte vaak ten koste van de betrouwbaarheid. Ten slotte maakt het grote aantal beschikbare meetinstrumenten het lastig voor veel zorgaanbieders om van elkaar te leren omdat meetinstrumenten altijd wel wat verschillen in content, antwoordcategorieën en psychometrische eigenschappen.

Het doel van dit proefschrift is het neerzetten van de basis voor een nieuw type meetinstrumentarium dat de genoemde beperkingen oplost door moderne meettechnieken toe te passen. Om dit doel te bereiken wordt een uitgebreide psychometrische evaluatie uitgevoerd op twee nieuwe meetinstrumenten die zijn ontwikkeld voor het meten van generieke depressie en angst. Depressie en angst zijn gekozen omdat het wereldwijde problemen zijn (Baxter, Scott, Vos, & Whiteford, 2013; Marcus, Yasamy, van Ommeren, Chisholm, & Saxena, 2012) en de meest voorkomende geestelijke gezondheidsproblemen in Nederland (de Graaf, ten Have, van Gool, & van Dorsselaer, 2012). Bovendien zijn depressie en angst vaak aanwezig bij andere psychische aandoeningen en zijn ze meestal de belangrijkste reden om geestelijke gezondheidszorg te zoeken (Clarke & Kissane, 2002; Frank, 1974).

De geëvalueerde meetinstrumenten zijn ontwikkeld door het Patient-Reported Outcomes Measurement Information System (PROMIS®) initiatief. PROMIS is aan het begin van deze eeuw opgericht in de Verenigde Staten (VS) om een nieuw, uniform zelfrapportage-meetinstrumentarium te ontwikkelen voor het meten van een groot deel van het gezondheidsspectrum (Cella et al., 2007, 2010). Ze hebben hiervoor tientallen item banken ontwikkeld die onder andere afgenomen kunnen worden als computergestuurde adaptieve test (CAT). CAT-instrumenten selecteren nieuwe items uit een set items (de item bank), meestal op zo een manier dat het volgende item het meest informatief is om het construct met een zo hoog mogelijke betrouwbaarheid te meten. Het selecteren van nieuwe items stopt zodra de meting een vooraf ingestelde betrouwbaarheid heeft bereikt. Met deze methodiek zou de clinicus niet alleen de *gewenste* informatie voor de behandeling kunnen verzamelen omdat de meetinstrumenten efficiënt zijn, maar in combinatie met de PROMIS item banken ook *kwalitatieve* informatie omdat de meetinstrumenten betrouwbaar, valide en responsief zijn.

In 2009 is de Nederlands-Vlaamse (NV) PROMIS groep opgericht om te onderzoeken of PROMIS ook geïmplementeerd kan worden in Nederland en België. Als startpunt hiervoor vertaalden ze 17 van de PROMIS item banken voor volwassenen (Terwee et al., 2014) en 9 van de PROMIS item banken voor kinderen (Haverman et al., 2016). Voor het meten van generieke depressie en angst bij volwassenen vertaalden ze de PROMIS v1.0 item banken voor Depressie en Angst. Verwacht wordt op basis van Amerikaans onderzoek dat CAT-afnames van deze item banken depressie en angst efficiënt, betrouwbaar, valide en responsief zullen meten in de Nederlandse en Vlaamse bevolking (Kroenke, Baye, & Lourens, 2019; Pilkonis et al., 2011, 2014; Schalet et al., 2016). Daarnaast wordt ook verwacht dat de PROMIS CATs efficiënter en betrouwbaarder zullen zijn dan andere meetinstrumenten (Pilkonis et al, 2014). In dit proefschrift worden deze verwachtingen onderzocht voor de algemene Nederlandse bevolking en voor Nederlandse patiënten die worden behandeld voor veelvoorkomende, psychische stoornissen.

Hoofdstuk 2: Het CAT potentieel evalueren met de MASQ

In Hoofdstuk 2 is eerst het *potentieel* van CAT onderzocht met simulatieonderzoek. Hiervoor is gebruikt gemaakt van de drie subschalen van de Mood and Anxiety Symptom Questionnaire (MASQ; Watson & Clark, 1991) en een bestaande dataset van patiënten die zijn behandeld door de zorgaanbieder Rivierduinen (*N* = 3.597). Het CAT-potentieel is vervolgens onderzocht met behulp van een psychometrische evaluatie en meerdere post hoc CAT-simulaties. Een post hoc CAT-simulatie is geen echte CAT-afname, maar gebruikt de antwoorden op de items uit de volledige item bank en evalueert deze alsof ze adaptief zijn verzameld. Eerder onderzoek heeft laten zien dat de uitkomsten van post hoc CAT-simulaties veel lijken op die van echte CAT-afnames (Kocalevent et al., 2009).

De resultaten lieten zien dat de drie subschalen van de MASQ gunstige psychometrische eigenschappen hebben voor een CAT-afname. Alleen voor de somatische angst schaal werd een item verwijderd vanwege de grote gelijkenis met een ander item. De CAT-simulaties lieten vervolgens voor alle MASQ schalen zien dat het aantal items aanzienlijk verminderd kan worden met een minimaal verlies aan betrouwbaarheid. Bij een acceptabele betrouwbaarheidsstandaard voor individuele metingen (standaard error [*SE*] ≤ 0,30) varieerde de percentuele item afname van 74% tot 56%. Bovendien bleven de scores en de voorspellende diagnostische waarde van de gesimuleerde CATs vergelijkbaar met die van de volledige MASQ schalen. Deze resultaten suggereren dat een CAT-versie van de MASQ kan leiden tot substantieel efficiëntere metingen met behoud van gunstige psychometrische eigenschappen.

Hoofdstuk 3: Psychometrische eigenschappen van de Nederlands-Vlaamse PROMIS Depressie item bank

In Hoofdstuk 3 is de PROMIS Depressie item bank onderzocht om een valide set items vast te stellen voor CAT-afnames (Reeve et al., 2007). Als startpunt hiervoor zijn de 28 items uit de NV PROMIS v1.0 volwassene item bank voor Depressie aangevuld met 28 vertaalde items die de VS versie net niet hebben gehaald door ongunstige psychometrische eigenschappen. Deze item set is vervolgens psychometrisch geëvalueerd met behulp van een gecombineerde algemene populatiesteekproef (*n* = 1.002) en een klinische steekproef van patiënten die zijn behandeld door de zorgaanbieder Parnassia Groep (*n* = 1.008). Op basis van de resultaten zijn acht van de aanvullende items verwijderd. De overgebleven item set is ten slotte gebruikt als input voor meerdere post hoc CAT-simulaties om de uitgebreide item bank (48 items) te vergelijken met de originele item bank (28 items). De resultaten suggereerden dat beide item banken geschikt zijn om depressie efficiënt, betrouwbaar en valide te meten met CAT. Voor de rest van het proefschrift is daarom gebruikt gemaakt van de originele PROMIS item bank omdat het gebruik hiervan internationale vergelijkingen mogelijk maakt. De CAT-simulaties voor deze item bank selecteerden voor de klinische steekproef gemiddeld 8,40 items bij een hoge betrouwbaarheidsstandaard voor individuele metingen (*SE* ≤ 0,20) en gemiddeld 3,40 items bij een acceptabele betrouwbaarheidsstandaard voor individuele metingen (*SE* ≤ 0,30).

Hoofdstuk 4: Psychometrische eigenschappen van de Nederlands-Vlaamse PROMIS Angst item bank

In Hoofdstuk 4 is de PROMIS Angst item bank onderzocht om een valide set items vast te stellen voor CAT-afnames. In deze studie is direct gekozen voor het gebruik van de items uit de originele NV PROMIS v1.0 volwassene item bank voor Angst omdat een pre-analyse van de data vergelijkbare resultaten liet zien als in Hoofdstuk 3: de efficiëntie, betrouwbaarheid en validiteit van een gesimuleerde CAT verschilt nauwelijks tussen de originele item bank en een uitgebreidere variant. Opnieuw werd de gecombineerde algemene en klinische populatiesteekproef gebruikt om aan te tonen dat ook de PROMIS Angst item bank gunstige psychometrische eigenschappen heeft voor een CAT-afname (*N* = 2.010). Daarnaast lieten verschillende post hoc CAT-simulaties zien dat angst efficiënt, betrouwbaar en valide gemeten kan worden met een CAT-afname van de item bank. Bij een hoge betrouwbaarheidsstandaard voor individuele metingen (*SE* ≤ 0,22) werden gemiddeld 8,64 items geselecteerd voor de klinische steekproef en 9,48 items voor de Nederlandse bevolkingssteekproef. Werd deze betrouwbaarheidsstandaard verlaagd naar acceptabel voor individuele metingen (*SE* ≤ 0,32), dan daalde het aantal geselecteerde items zelfs naar gemiddeld 4,25 voor de klinische steekproef en gemiddeld 6,06 items voor de Nederlandse bevolkingssteekproef.

Hoofdstuk 5: Longitudinale meetinvariantie in de PROMIS item banken

In hoofdstuk 5 zijn de NV PROMIS v1.0 volwassene item banken voor Depressie en Angst onderzocht op het validiteitsaspect longitudinale meetinvariantie (LMI; Fokkema et al., 2013; Fried et al., 2016). Hiervoor is gebruikt gemaakt van begin- en tussenmetingen, afgenomen bij patiënten die zijn behandeld door de zorgaanbieder Parnassia Groep voor stemmingsstoornissen (*N* = 640) en angststoornissen (*N* = 528). Met factoranalyse is vervolgens onderzocht of de item banken voldoende unidimensioneel zijn op beide meetmomenten en de gemeten constructen niet substantieel veranderen over tijd (Liu et al., 2017; Liu & West, 2018). De resultaten suggereerden dat de item banken voldoende unidimensioneel zijn op beide meetmomenten, maar de depressie en angst constructen veranderden wel over tijd. Op basis van voorlopige criteria lieten vervolganalyses echter zien dat de invariantieschendingen geen substantieel effect hebben op de antwoorden van patiënten en hun uiteindelijke scores. Hieruit kan geconcludeerd worden dat de invariantieschendingen waarschijnlijk van weinig betekenis zijn voor de klinische praktijk. Met andere woorden: de NV PROMIS v1.0 volwassene item banken voor Depressie en Angst kunnen beschouwd worden als voldoende invariant over tijd.

Hoofdstuk 6: Psychometrische eigenschappen van de PROMIS CATs voor Depressie en Angst

In Hoofdstuk 6 zijn echte CAT-afnames van de NV PROMIS volwassene v1.0 item banken voor Depressie en Angst vergeleken met de subschalen van de Brief Symptom Inventory (BSI; de Beurs & Zitman, 2005). Hiervoor is gebruikt gemaakt van begin- en tussenmetingen, afgenomen bij patiënten die zijn behandeld door de zorgaanbieder Dimence Groep (*N* = 400). Met verschillende analyses is vervolgens de validiteit en de responsiviteit van de PROMIS CATs onderzocht en zijn twee indicatoren voor verandering geëvalueerd. De resultaten suggereerden dat de PROMIS CATs vergelijkbare constructen meten als de BSI schalen voor Depressie en Angst. Onder deze aanname werd ook aangetoond dat de CAT Depressie volgens verwachting ongeveer even responsief is als de BSI Depressie schaal. De CAT Angst daarentegen was tegen de verwachting in substantieel responsiever dan de BSI Angst schaal. Ten slotte categoriseerden beide PROMIS CATs meer patiënten als daadwerkelijk veranderd dan de overeenkomende BSI schalen. Op basis hiervan kan geconcludeerd worden dat de PROMIS CATs voor Depressie en Angst een bescheiden verbetering zijn ten opzichte van de BSI schalen voor Depressie en Angst. De CAT-instrumenten zijn mogelijk beter in staat om daadwerkelijke verandering te detecteren, waarschijnlijk door een hogere betrouwbaarheid.

Hoofdstuk 7: Discussie

In dit proefschrift zijn de NV PROMIS volwassene v1.0 item banken voor Depressie en Angst psychometrisch geëvalueerd voor CAT-afnames in de algemene Nederlandse bevolking en de Nederlandse klinische populatie (d.w.z., patiënten die worden behandeld voor veelvoorkomende, psychische stoornissen). De resultaten suggereerden dat beide item banken zeer informatief zijn voor het meten van depressie en angst bij een groot deel van de Nederlandse (klinische) populatie. Dit werd bevestigd door zowel post hoc CAT-simulaties als echte CAT-afnames, die beide lieten zien dat depressie en angst betrouwbaar en efficiënt gemeten kunnen worden. Daarnaast werd aangetoond dat de CAT-instrumenten valide zijn voor het meten van depressie en angst in de Nederlandse (klinische) populatie, en responsief voor het meten van veranderingen in depressie en angst in de klinische populatie. Op basis van deze resultaten kan worden geconcludeerd dat het hoofddoel van dit proefschrift voldoende is bereikt: CAT-versies van de DF PROMIS volwassene v1.0 item banken voor Depressie en Angst zijn efficiënt, betrouwbaar, valide en responsief voor het meten van depressie en angst in de Nederlandse (klinische) populatie.

In de PROMIS CAT-studies vielen voornamelijk twee bevindingen op. Ten eerste lieten de PROMIS CATs slechts bescheiden verbeteringen zien ten opzichte van de BSI. Het lijkt daarom misschien niet erg aantrekkelijk voor clinici om de overstap te maken naar PROMIS CATs, zeker gezien het feit dat gebruikers moeten wennen aan nieuwe meetinstrumenten, wat mogelijk als last ervaren kan worden. Hierbij moet echter opgemerkt worden dat de PROMIS CATs (mogelijk) extra voordelen hebben vergeleken met andere meetinstrumenten. Ten eerste zijn PROMIS instrumenten ontwikkeld voor een universele toepassing (Belekkas et al., 2018; Lizzio et al., 2019; Papuga et al., 2018; Scholle et al., 2018; Wagner et al., 2015), terwijl veel andere instrumenten vaak primair zijn ontwikkeld voor het meten van klinische populaties. Ten tweede kunnen PROMIS scores wellicht vergeleken worden tussen verschillende landen om van elkaar te leren (Elsman, Flens, de Beurs, Roorda, Terwee, 2022; Terwee et al., 2021; van Bebber et al., 2018), terwijl dit voor andere instrumenten onduidelijker is. Ten derde hebben PROMIS gebruikers toegang tot tal van andere PROMIS instrumenten voor het meten van een groot deel van het gezondheidsspectrum (bijv. Crins et al., 2015, 2016, 2017; Terwee et al., 2019), waardoor een set meetinstrumenten voor behandelevaluatie flexibel samen te stellen is. Ten vierde is in dit proefschrift aangetoond dat de NV PROMIS volwassene v1.0 item banken voor Depressie en Angst voldoende invariant zijn over tijd, terwijl dit nog onduidelijk is voor veel andere Nederlandse instrumenten die mentale gezondheidsconstructen meten. Ten slotte zullen de voordelen van de PROMIS CATs waarschijnlijk groter zijn vergeleken met andere meetinstrumenten. In hoofdstuk 2 is bijvoorbeeld aangetoond dat CAT-versies van de drie MASQ schalen een item afname laten zien van 74% tot 56% met een minimaal verlies aan betrouwbaarheid.

De tweede opvallende bevinding is het verschil tussen de PROMIS CATs met betrekking tot de onderzochte psychometrische eigenschappen. Voor beide meetinstrumenten zijn adequate psychometrische eigenschappen aangetoond, maar de CAT Angst presteerde net wat minder goed dan verwacht. Dit bleek voor zowel de efficiëntie, de betrouwbaarheid, de validiteit als de responsiviteit van het instrument. Op basis van deze resultaten kan het zinnig zijn om na te gaan of de psychometrische eigenschappen van de PROMIS CAT Angst nog verder verbeterd kunnen worden. Het kan bijvoorbeeld zijn dat het gemeten Angst construct beter onderverdeeld kan worden in meerdere subdomeinen, waardoor het meten van angst wellicht geschikter is voor een multidimensionele CAT (MCAT; Smits, Paap, & Böhnke, 2018). Misschien is het echter ook wel afdoende om een betere balans te realiseren in het aantal items waarmee elk subdomein wordt gemeten. Nader onderzoek moet hier uitsluitsel over geven.

De CAT-versies van de NV PROMIS volwassene v1.0 item banken voor Depressie en Angst kunnen vanaf nu geïmplementeerd worden in de klinische praktijk. Voor een succesvolle implementatie zijn verschillende aandachtspunten van belang. Voorbeelden hiervan zijn het vergroten van de beschikbaarheid en toegankelijkheid van CAT-technologie voor zorgaanbieders, het vergroten van kennis over CAT-instrumenten bij clinici, en het vergroten van de beschikbaarheid van gebruiksvriendelijke interfaces en feedbacktools. Misschien wel het meest dringende aandachtspunt is dat zorgaanbieders überhaupt gebruik gaan maken van meetinstrumenten als hulpmiddel voor het evalueren van de behandeling. In de Nederlandse geestelijke gezondheidszorg staat dit gebruik nog steeds in de kinderschoenen, ondanks dat tal van studies gunstige effecten hebben aangetoond op de algehele kwaliteit van zorg (de Jong et al., 2021; Fortney et al. 2017; Guo et al., 2015; Rush & Thase, 2018; Scott & Lewis, 2015). De belangrijkste reden hiervoor is dat een succesvolle implementatie van meetinstrument-gerichte zorg verschillende uitdagingen met zich meebrengt (Martin-Cook et al., 2021). Slechts één van deze uitdagingen is de beschikbaarheid van korte en betrouwbare zelfrapportage-meetinstrumenten zoals de PROMIS CATs voor Depressie en Angst. Andere voorbeelden zijn het tonen van actief leiderschap bij de implementatie, een actieve betrokkenheid van de clinicus en patiënt, en de afstemming van het meetinstrumentarium op de aandoeningen en behandeldoelen van patiënten.

Terwijl de implementatie van CAT-instrumenten in de klinische praktijk opgestart wordt, kan de ontwikkeling van een volgende generatie CAT-instrumenten in gang worden gezet. Enkele voorbeelden van toekomstige onderzoeksrichtingen zijn het verbreden van het CAT-aanbod, het ontwikkelen van MCATs en het verbeteren van de stopregel van het CAT-algoritme. Door de CAT-instrumenten verder te verbeteren, wordt de kans vergroot dat dit moderne meetinstrumentarium uiteindelijk de nieuwe standaard wordt voor het evalueren van behandelingen in Nederland. Dit kan het gebruik van meetinstrumenten in de behandelkamer verder stimuleren, wat weer kan gaan leiden tot effectievere en efficiëntere behandelingen in het algemeen.

About the Author

Gerard Slok - Flens was born on May 15th, 1985 in Zaandam. In 2003, he completed secondary school at SG Zaanlands Lyceum in Zaandam. Additionally, he completed a Bachelor of Science in Psychology in 2007 and a Master of Science in Psychology in 2011, both at the University of Amsterdam. In 2014, he started as an external PhD student at Leiden University (Faculty of Social and Behavioural Sciences, Institute of Psychology) under the supervision of prof. dr. Edwin de Beurs, prof. dr. Philip Spinhoven, dr. Niels Smits, and dr. Caroline B. Terwee. The work presented here is the final product of this project. In the meantime, Gerard held various positions at the Foundation for benchmarking mental health care (Stichting Benchmark GGZ; SBG) from 2011 to 2018. His current position is that of Product owner of the Data portal mental health (GGZ Dataportaal; https://ggzdataportaal.nl) at the Quality alliance mental health (Alliantie kwaliteit in de geestelijke gezondheidszorg; Akwa GGZ; https://akwaggz.nl). Gerard’s primary motivation in his work is to contribute to the improvement of mental health care. Additionally, he has a keen interest in world literature and the nature of reality.

List of Publications

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

de Beurs, E., Barendregt, M., **Flens, G.**, van Dijk, E., Huijbrechts, I., & Meerding, J. W. (2012). Vooruitgang in de behandeling meten: Een vergelijking van vragenlijsten voor zelfrapportage (A comparison of self-report questionnaires for treatment outcome). *Maandblad Geestelijke Volksgezondheid, 67*, 259-270.

de Beurs, E., & **Flens, G.** (2017). Gebruik van verschillende meetinstrumenten: de getransformeerde T-score en equivalente responsiviteit. In E. de Beurs, M. Barendregt, & L. Warmerdam (Eds.), *Behandeluitkomsten: bron voor kwaliteitsbeleid in de GGZ* (pp. 223-236). Amsterdam: Boom.

de Beurs, E., **Flens, G.**, & Williams, G. (2019). Meetresultaten interpreteren in de klinische psychologie: een aantal voorstellen [The interpretation of measurement results in clinical psychology: several proposals]. *De Psycholoog*, *54*(6), 10-23.

Elsman, E. B., **Flens, G**., de Beurs, E., Roorda, L. D., & Terwee, C. B. (2022). Towards standardization of measuring anxiety and depression: Differential item functioning for language and Dutch reference values of PROMIS item banks. *Plos one*, *17*(8), e0273287.

Flens, G., & de Beurs, E. (2017). De toekomst van ROM: computer-gestuurd adaptief testen (The future of ROM: computerised adaptive testing). *Tijdschrift voor Psychiatrie, 59*(12), 767-774.

Flens, G., & de Beurs, E. (2017). Verzamel gegevens over behandeluitkomsten uniform. In E. de Beurs, M. Barendregt, & L. Warmerdam (Eds.), *Behandeluitkomsten: bron voor kwaliteitsbeleid in de GGZ* (pp. 211-221). Amsterdam: Boom.

Flens, G., Smits, N., Carlier, I., van Hemert, A. M., & de Beurs, E. (2016). Simulating computer adaptive testing with the Mood and Anxiety Symptom Questionnaire. *Psychological Assessment, 28*(8), 953-962.

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., & de Beurs, E. (2017). Development of a computer adaptive test for depression based on the Dutch-Flemish version of the PROMIS item bank. *Evaluation & the health professions, 40*(1), 79-105.

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., Spinhoven, P., & de Beurs, E. (2019). Development of a computerized adaptive test for anxiety based on the Dutch–Flemish version of the PROMIS item bank. *Assessment, 26*(7), 1362-1374.

Flens, G., Smits, N., Terwee, C. B., Pijck, L., Spinhoven, P., & de Beurs, E. (2021). Practical Significance of Longitudinal Measurement Invariance Violations in the Dutch–Flemish PROMIS Item Banks for Depression and Anxiety: An Illustration With Ordered-Categorical Data. *Assessment, 28*(1), 277-294.

Flens, G., Terwee, C. B., Smits, N., Williams, G., Spinhoven, P., Roorda, L. D., & de Beurs, E. (2022). Construct validity, responsiveness, and utility of change indicators of the Dutch-Flemish PROMIS item banks for depression and anxiety administered as computerized adaptive test (CAT): A comparison with the Brief Symptom Inventory (BSI). *Psychological Assessment, 34*(1), 58-69.

Heesterbeek, M. R., Luijten, M. A. J., Gouw, S. C., Limperg, P. F., Fijnvandraat, K., Coppens, M., ... & Haverman, L. (2022). Measuring anxiety and depression in young adult men with haemophilia using PROMIS. *Haemophilia: the official journal of the World Federation of Hemophilia, 28*(3), e79-e82.

van Bebber, J., **Flens, G.**, Wigman, J. T., de Beurs, E., Sytema, S., Wunderink, L., & Meijer, R. R. (2018). Application of the Patient-Reported Outcomes Measurement Information System (PROMIS) item parameters for Anxiety and Depression in the Netherlands. *International journal of methods in psychiatric research, 27*(4), e1744.

van Gorp, M., Dallmeijer, A. J., van Wely, L., de Groot, V., Terwee, C. B., **Flens, G.**, ... & PERRIN DECADE Study Group. (2021). Pain, fatigue, depressive symptoms and sleep disturbance in young adults with cerebral palsy. *Disability and rehabilitation, 43*(15), 2164-2171.

Williams, G. L., de Beurs, E., Spinhoven, P., **Flens, G**., & Paap, M. C. (2021). Support for the higher-order factor structure of the WHODAS 2.0 self-report version in a Dutch outpatient psychiatric setting. *Quality of Life Research*, 1-11.

Williams, G. L., **Flens, G.**, & de Beurs, E. (2017). Computergestuurd Adaptief Testen (CAT), moderne meettechniek voor uitkomstmaten. *PsyXpert, 4*, 14–21.

Word of Gratitude

Many thanks to Edwin de Beurs for initially suggesting that I start this PhD, and for taking the initiative to get the project going. I would also like to recognize everyone involved at Leiden University and the Foundation for benchmarking mental health care (Stichting Benchmark GGZ; SBG) for accepting our proposal. Furthermore, I am grateful to SBG and the Quality alliance mental health (Alliantie Kwaliteit in de GGZ; Akwa GGZ) for supporting me to work on this thesis. In particular, I would like to thank Maarten Erenstein (SBG), Marko Barendregt (SBG), Astrid van Meeuwen (SBG), Dominique Vijverberg (Akwa GGZ), and Tim Hüfken (Akwa GGZ) for your continuous support. Maarten, I really appreciate that you let me present my work around the globe. Not to mention Betsy Jekel, who helped me with many practical arrangements around these wonderful adventures.

Edwin de Beurs, Philip Spinhoven, Niels Smits, and Caroline Terwee, I would like to express my deepest gratitude for your excellent and patient guidance. Throughout this project you continuously provided me with valuable feedback and positively challenged me to increase the quality of my work. To my other coauthors Albert van Hemert, Ingrid Carlier, Joost Dekker, Irma Huijbrechts, Liv Possen-Pijck, Guido Williams, and Leo Roorda, many thanks for your careful suggestions to improve the manuscripts. I would also like to acknowledge Ellen Elsman-Perlot, Jan van Bebber, Martine Crins, Guido Williams, Marloes van Gorp, Madelief Heesterbeek, and Edwin de Beurs for allowing me to do the same for you. A special thanks goes to Terrence Jorgensen for his assistance with the longitudinal measurement invariance analyses presented in Chapter 5. That being said, I would like to acknowledge all the researchers for their previous studies on which I have built. Should anyone feel that their work is not fully acknowledged, please let me know. Also, I would like to recognize the journals Assessment, Psychological Assessment, and Evaluation & the Health Professions for publishing the studies presented here. Finally, to everybody involved in finalizing my PhD project: thank you! To Bernet Elzinga, Rob Meijer, Anne Stiggelbout, Kim de Jong, Muirne Paap, and Marjolein Fokkema for carefully reading through the scientific content of this thesis and allowing me to defend it; to Paul Wouters for his work as dean; to Anita Nieuwenhuizen and Cordula Burtscher from the Graduate School for all their help and kind advice in the final stages of the project; and to my paranymphs Astrid Partouns, Anna Hoogenkamp and Lilian Hoonhout for their loving support.

The CAT studies would not have been possible had it not been for the efforts of many people. I would like to humbly thank each and every one of you. First, to everybody involved at the mental health providers Rivierduinen, Parnassia Group, and Dimence Group for facilitating the data collections. In particular, I would like to thank Ingrid Carlier, Irma Huijbrechts, Liv Pijck, Monique van Bueren, Erik de Groot, and Alfred Kaal for their help. Second, to all patients of the participating mental health providers for their cooperation with the studies. Third, to everybody involved at the Dutch-Flemish PROMIS group for allowing me to evaluate their item banks for CAT administration. In particular, I would like to thank Leo Roorda and Caroline Terwee for their work on the Dutch-Flemish Assessment Center, which made it possible to administer genuine CAT instruments to patients. Fourth, to everybody involved at Vital Health for believing in the added value of CAT and making it possible to set up the very first PROMIS CAT study in the Netherlands. And finally, to Karel Kroeze for developing the CAT algorithm and releasing it to the public.

To my beloved husband Bjorn Slok - Flens, I would like to express my deepest gratitude for your continuous support, and for your patience and understanding when you had to make sacrifices for my personal goals. To my wonderful sister and fellow guru Claudia Flens, my generous parents Gerda Vleeshakker and Klaas Flens, my lovely friends Susanna Gerritse, Anna Hoogenkamp, Marc van Wageningen, Marius Roothaan, and Michel Boet, my fellow CAT enthusiasts Muirne Paap, Benjamin Schalet, and Felix Fischer, and to all my (former) colleagues at SBG and Akwa GGZ, thank you for being there in times of need. And a big final shout out to my dear lifelong friend Soe Rafeek for your beautiful cover: namasté!