Computerized adaptive testing in Dutch mental health care: a new tool to assess depression and anxiety

Slok-Flens, G.

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.

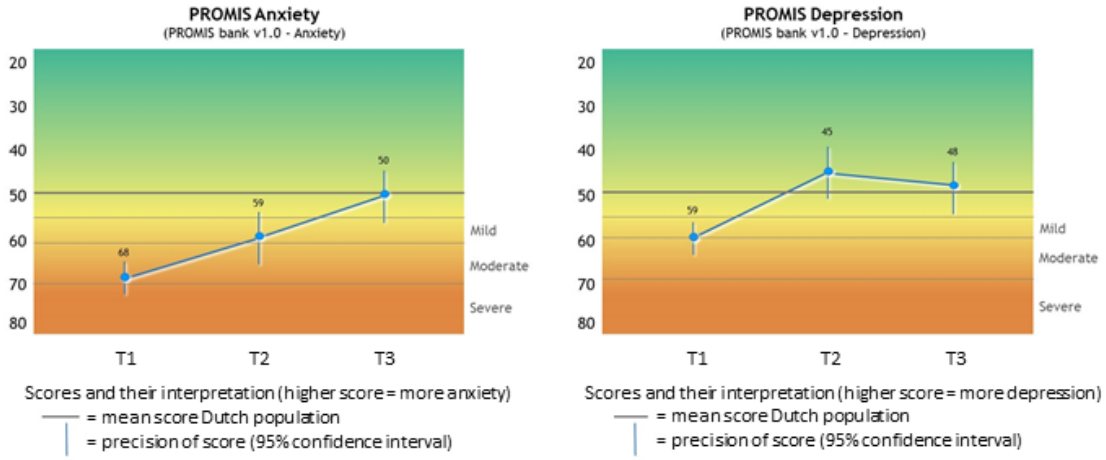
7.3.4 Developing useful feedback tools

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.