



Universiteit
Leiden
The Netherlands

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

Slok-Flens, G.

Citation

Slok-Flens, G. (2022, October 5). *Computerized adaptive testing in Dutch mental health care: a new tool to assess depression and anxiety*. Retrieved from <https://hdl.handle.net/1887/3466118>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3466118>

Note: To cite this publication please use the final published version (if applicable).

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, & Sijsma, 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 States (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., Bevens 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.