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

Slok-Flens, G.

# Chapter 2

## Simulating Computer Adaptive Testing with the Mood and Anxiety Symptom Questionnaire

Published as:

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

## 2.1 Abstract

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

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

## 2.2 Background

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

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

The full 90-item MASQ is an extensive questionnaire which has a unique way of assessing symptoms of the two most prevalent psychiatric syndromes, depression and anxiety disorders (according to the tripartite model), and takes into account the high comorbidity between both syndromes and high level of symptom overlap (Watson & Clark, 1991). It is used as research- and clinical assessment instrument, and has been validated in multiple countries, for multiple age groups, and for multiple disorders (e.g., de Beurs, den Hollander-Gijsman, Helmich, & Zitman, 2007; Deng, Jiang, & Li, 2012; Lee, Kim, & Cho, 2015). Ideally, for efficient measurement of clinical subjects, all subscales of the MASQ are transformed into a CAT. Previous studies have generally confirmed three subscales of the 90-item MASQ: a positive affect scale (PA), a negative affect scale (NA), and a somatic anxiety (SA) scale (Bedford, 1997; Clark & Watson, 1991; de Beurs et al., 2007, Keogh & Reidy, 2000; Watson et al., 1995). Other studies that developed shorter versions of the MASQ also applied this three

factor structure in their item design (Osman et al., 2011; Wardenaar et al., 2010). In these studies, the number of items for each MASQ scale was fixed, but by doing so, the measurement precision for test outcomes could vary among respondents with different trait levels. By contrast, CAT is more dynamic: it fixes the test outcomes' measurement precision for all trait levels and allows for the number of administered items to vary among respondents (Embretson & Reise, 2000). In other words, CAT is essentially more efficient than fixed questionnaires because CAT administers only the most informative items to each individual respondent.

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

## 2.3 Method

### 2.3.1 Participants

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

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

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

### 2.3.2 The MASQ

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

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

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

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

### 2.3.3 Psychometric evaluation of the MASQ scales

We undertook a psychometric evaluation of the three MASQ scales (Reeve et al., 2007), which is a prerequisite for applying CAT. It was evaluated whether each of the scales met the three

main item response theory (IRT) assumptions of unidimensionality, local independence (LI) and monotonicity. Violation of these assumptions may cause bias in the scaling of persons and items on a common latent trait, which could result in over- or underestimated trait scores. In addition, we evaluated differential item functioning (DIF; Embretson & Reise, 2000) among key demographic groups. Items containing DIF cause bias in latent trait scores because persons from different groups with the same latent trait score have different probabilities of selecting item response categories.

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

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

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

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

### 2.3.4 CAT Simulation

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

environment R (R Core Team, 2014) to implement these decisions into three separate CAT simulations. Below, we will present the details concerning the decisions rules.

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

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

Finally, we chose a starting level and stopping rule for the CAT. The starting level was set to the average value of the latent trait ($\theta = 0$). As a first item for all respondents, we therefore chose the MASQ item which had the highest information at this starting level: Item 86 for the PA scale (*Felt really good about myself*), Item 22 for the NA scale (*Felt hopeless*), and Item 79 for the SA scale (*Was trembling or shaking*). In addition, there are generally two types of stopping rules for a CAT: (a) a fixed number of administered items, or (b) a prespecified value of measurement precision (*SE*). Because this study was set out to find both reliable and shorter measures, we specified that the CAT simulation stopped applying new items when the latent trait estimate of a patient reached a $SE(\theta) < 0.3$, comparable to a marginal reliability of .90 (Green, Bock, Humphreys, Linn, & Reckase, 1984). This value of measurement precision is generally required for minimal reliability for individual assessments (Bernstein & Nunnally,

1994, p. 265). When a $SE(\theta) < 0.3$ was not obtained after administering all items, the CAT simulation stopped.

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

### 2.3.5 Comparing full-scale data with CAT data

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

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

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

## 2.4 Results

### 2.4.1 Psychometric qualities of the MASQ scales

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

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

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

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

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

The graphs of item mean scores conditional on rest scores showed monotonicity for all items as the underlying level of the scale was higher. This result was confirmed by the Mokken scale analysis (van der Ark, 2007). The scalability coefficient for the PA and NA scales was

high (0.53 and 0.55), and for the SA scale it was moderate (0.42). Furthermore, the scalability coefficients for all items were above the lower bound of 0.30. According to these results, we concluded that all three scales sufficiently met the monotonicity assumption.

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

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

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

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

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

### 2.4.2 Calibration of the MASQ scales

The correlations between the GRM's estimated theta's and the traditional MASQ scale scores were high for all scales (PA: $r$ = .98, NA: $r$ = .98, SA: $r$ = .96), indicating the GRM as a good

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

### 2.4.3 Characteristics of the CATs

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

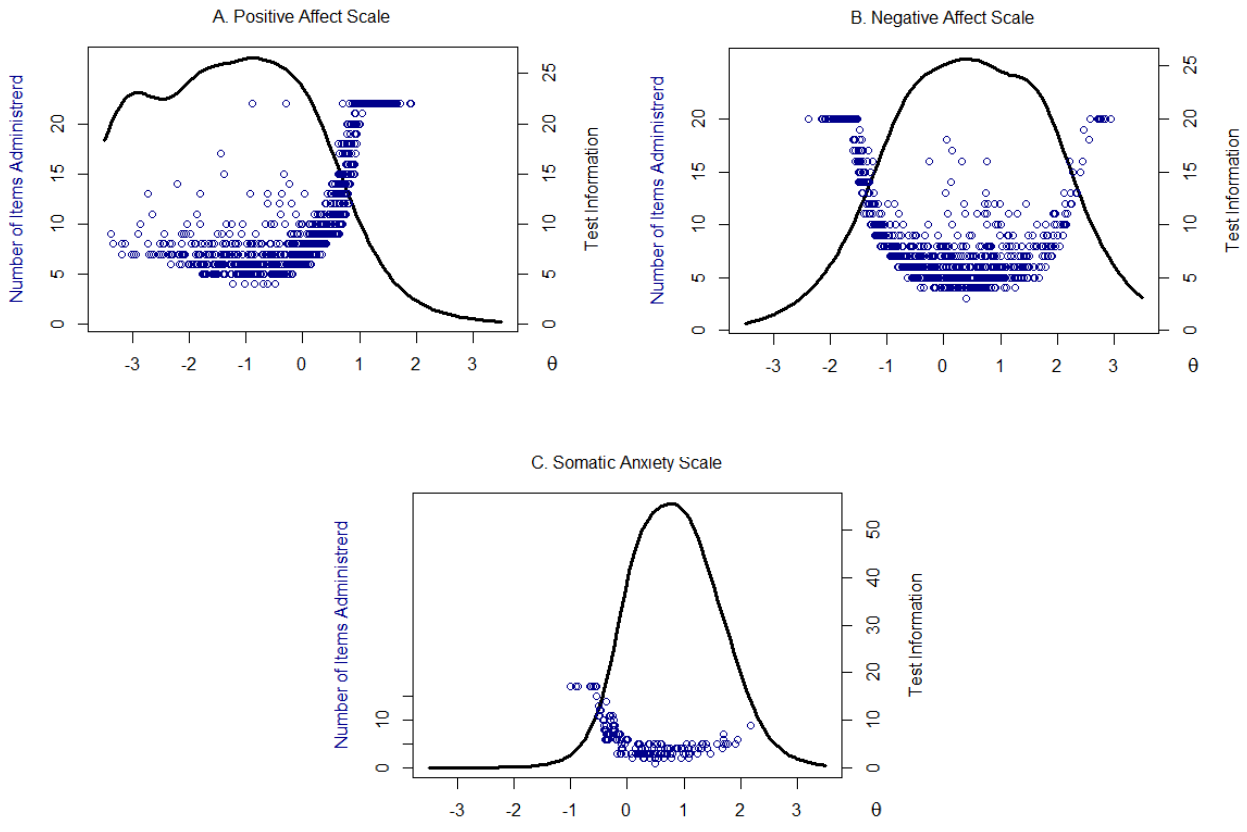
**Table 2.4** CAT simulation statistics for the three MASQ scales (PA, NA and SA) under stopping rule $SE(\theta) < 0.3$.

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

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

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

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




## 2.4.4 Validity of the CATs

Table 2.5 displays Pearson's correlations and sizes of differences (Cohen's *d*) for each subscale (PA, NA, SA) between the CAT θ estimates and the full-scale θ estimates. The correlations were high for all scales (PA: $r = .98$, NA: $r = .98$, SA: $r = .89$), indicating a high similarity between the CAT and the full-scale θ estimates. Next, we investigated the scatterplots for each

scale and did not identify notable outliers. Finally, the Cohen's *d* values were small for all scales, indicating no structural differences between the CAT θ estimates and the full-scale θ estimates.

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

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

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

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

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

## 2.5 Discussion

Until recently, most of the studies that build a CAT version for an existing clinical scale were executed with undergraduate students (Forbey & Ben-Porath, 2007; Gardner et al., 2004; Smits et al., 2011). In this study, we used data from clinical subjects to assess whether the efficiency

of the MASQ could be improved with a CAT version. For this purpose, we performed a psychometric evaluation and a CAT simulation on each of the three MASQ scales. Performing a simulation enabled us to compare the full-scale assessments and the CAT simulations within the same patient group. Thus, we could directly assess to what extent the CAT simulations reduced the number of administered items and whether the θ estimates of the CAT simulations had similar outcomes and diagnostic accuracy compared with the full-scale θ estimates.

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

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

After replication of the present study's results with an actual CAT, the MASQ CAT could be used in clinical practice for single measure purposes. When the final goal is to use the MASQ CAT in ROM, two additional requirements have to be met. First, the MASQ CAT has to measure the same three scales (PA, NA and SA) at different points in time (factorial invariance over time). When patients' values or their internal standards for measurement are

changed, comparing observed scale scores could be biased (response-shift; Fokkema, Smits, Kelderman, & Cuijpers, 2013). Second, the responsiveness to change of the MASQ CATs should be equal to the full-scale scores. When instruments' sensitivity to detect change is different, treatment outcomes could be biased (de Beurs et al., 2012). In future research, these requirements have to be investigated to assess the utility of the MASQ CAT in ROM.

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

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

Two last lines of future research, which can be pursued with the MASQ, are the investigation of clinical cut points and their sensitivity and specificity for mood and/or anxiety diagnosis, and the factor structure for patients. In clinical practice, the patient's diagnosis is usually determined by a standardized clinical interview without using the MASQ. When cut points are available for determining mood/anxiety diagnosis with the MASQ, these could be used to assess whether a CAT classification would differ from the classification according to the full MASQ score. Moreover, the MASQ could be used in clinical practice for diagnostic prediction. In addition, some hold the view that the distinction between depressive and anxiety symptoms could best be described by a hierarchical model instead of a three-factor model (Lin

et al., 2014; Simms, Gros, Watson, & O'Hara, 2008; Simms, Prisciandaro, Kruger, & Goldberg, 2012). A hierarchical model assumes that anxiety and depression are measured by a general factor and several underlying factors. When this hierarchical model also applies to MASQ data from (Dutch) clinical subjects, then the MASQ could be used to develop a CAT which takes into account the dimensional structure of the combined scales (Reckase, 1985). This type of CAT could enhance the reliability of the scores and the administration efficiency even further.

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