

From **DEPARTMENT OF CLINICAL NEUROSCIENCE**  
Karolinska Institutet, Stockholm, Sweden

# **MEASUREMENT OF MANIA AND DEPRESSION**

Mats Adler



**Karolinska  
Institutet**

Stockholm 2011

All previously published papers were reproduced with permission from the publisher.

Published by Karolinska Institutet. Printed by US-AB

© Mats Adler, 2011

ISBN 978-91-7457-485-2

## ABSTRACT

**Background:** In psychiatry, the assessment of symptom severity is being increasingly assisted by rating scales, in clinical practice as well as in research and quality control. Transforming the subjective symptoms of psychiatric disorders into valid numerical measures is subjected to numerous confounding factors. Careful evaluation of rating scales is therefore essential. This doctoral project arose from a clinical need for a useful self-rating scale for affective symptoms at an outpatient clinic for affective disorders. No existing rating scales fulfilling the clinical need were found in the literature.

**Aims:** The aims of the doctoral project were to develop and evaluate a self rating scale for measurement of severity in depressive, manic and mixed affective states and to explore if Item Response Theory (IRT) is useful for evaluation and improvement of rating scales for mania and depression. A further aim was to investigate if Randomized Controlled Studies of Antidepressants (RCT-ADs) might be biased due to measurement properties of the most commonly used rating scale, the Hamilton Depression Rating Scale (HDRS).

**Methods:** A self rating scale consisting of 18 items was developed and named the Affective Self-Rating Scale (AS-18) with separate subscales for depression and mania/hypomania. It was evaluated in two samples of patients (N=61 and N=231) and was compared to the Patient Health Questionnaire (PHQ9) and the Montgomery Åsberg Depression Rating Scale (MADRS). Data from five RCT-ADs included in a recently published meta-analysis were analyzed (N=516). Statistical methods from Classical Test Theory (CTT) and IRT were used.

**Results:** The AS-18 showed good estimates of reliability with Cronbachs alpha (CTT) of 0.89 and 0.91 for the depression and mania subscales. The ratings on the AS-18 showed strong correlation to reference scales. A factor analysis largely confirmed the predicted factor structure. Items for irritability, risk-taking and increased sleep did not, however, behave as predicted. The IRT analysis showed that the AS-18 and PHQ9 had strong capacities to rank respondents according to their scores, while the MADRS had weak such properties. Several items in the rating scales contributed little information to the measurement. There was a shortage of items covering lower levels of the depression and mania dimensions making measurement of lower levels of symptoms imprecise.

In the analysis of five RCTs it was found that the HDRS yielded decreasing amounts of information at declining levels of depression severity. In addition it was found that the items of HDRS were understood differently by the study persons of the different RCT-ADs. The conclusion of the meta-analysis, that antidepressants had negligible effect in low to moderate depression severity, was therefore found to be unsupported by data.

**Conclusions:** The AS-18 has demonstrated reliability and validity in two studies. In outpatient settings for affective disorder patients, it can be used as a time-efficient aid for clinicians in identifying patients with different affective states as well as rating their severity. IRT-methods were demonstrated to be useful for analyzing rating scales concerning the amount of information that individual items contribute to the measurement, how the precision of measurement varies over the severity spectrum and for investigating whether different study populations perceive items differently. Studies of antidepressant efficacy can be biased due to shortcomings of measurement.

## LIST OF PUBLICATIONS

- I. **Adler M**, Liberg B, Andersson S, Isacsson G, Hetta J. Development and validation of the Affective Self Rating Scale for manic, depressive, and mixed affective states. *Nord J Psychiatry*. 2008;62(2):130-5.
- II. **Adler M**, Brodin U. An IRT validation of the Affective Self Rating Scale. *Nord J Psychiatry*. 2011 May 4. [Epub ahead of print]
- III. **Adler M**, Hetta J, Isacsson G, Brodin U. An Item Response Theory evaluation of three depression questionnaires in a clinical sample. *Submitted*.
- IV. Isacsson G, **Adler M**. Randomized Clinical Trials underestimate the efficacy of antidepressants in less severe depression. *Submitted*.

# CONTENTS

1.	Introduction .....	1
1.1	Aims of the doctoral project.....	2
1.2	Overview of the project.....	2
2.	Classification and symptoms of mood disorders.....	3
2.1	Classification of mood disorders.....	3
2.2	Symptoms of depression and mania .....	4
2.3	Mixed states .....	5
2.4	Special considerations for bipolar depression .....	5
2.5	Overlap of symptoms of mania and depression .....	6
2.6	Existing rating scales for affective symptoms.....	6
2.6.1	Rating scales for depression.....	6
2.6.2	Rating scales for mania .....	7
2.6.3	Rating scales for mixed states.....	7
3.	Measurement theory.....	9
3.1	Measurement of latent variables .....	9
3.1.1	Different purposes of rating scales .....	9
3.1.2	Types of latent variables .....	9
3.1.3	Different forms of rating scales .....	10
3.1.4	Different ways of collecting information .....	11
3.1.5	Different types of data.....	11
3.1.6	Restrictions on calculations .....	11
3.2	Key concepts for evaluation of rating scales .....	12
3.2.1	Reliability .....	12
3.2.2	Validity .....	12
4.	Statistical methods for the evaluation of rating scales .....	15
4.1	Classical Test Theory (CTT).....	15
4.1.1	Reliability in CTT. ....	15
4.1.2	Validity in CTT .....	16
4.1.3	Advantages of CTT-methods.....	16
4.1.4	Shortcomings of CTT-methods .....	16
4.2	Item Response Theory.....	17
4.2.1	Criteria for a well functioning rating scale in IRT .....	18
4.2.2	Different types of IRT .....	19
4.2.3	Non-parametric IRT .....	20
4.2.4	Parametric IRT .....	22
4.2.5	Calculation of model parameters .....	25
4.2.6	Special features of IRT.....	25
4.2.7	IRT-measures .....	28
4.2.8	Advantages of IRT-methods.....	29
4.2.9	Shortcomings of IRT methods.....	30
5.	The studies.....	31
5.1	Study one: Development and validation of the Affective Self Rating Scale.....	31
5.1.1	Results .....	31
5.1.2	Discussion.....	33

5.2	Study two: An IRT validation of the Affective Self Rating Scale .....	33
	5.2.1 Results .....	34
	5.2.2 Discussion .....	35
5.3	Study three: An IRT evaluation of three depression rating scales .....	36
	5.3.1 Results .....	36
	5.3.2 Conclusions .....	38
5.4	Study four: Randomized Clinical Trials underestimate the efficacy of antidepressants in less severe depression .....	39
	5.4.1 Results .....	39
	5.4.2 Discussion .....	41
6.	Conclusions from the doctoral project .....	43
7.	Sammanfattning på svenska .....	47
8.	Acknowledgements .....	48
9.	References .....	49
10.	Appendix .....	55

## LIST OF ABBREVIATIONS

<b>ADHD</b>	Attention Deficit Hyperactivity Disorder
<b>AIS</b>	Automated item selection
<b>ANOVA</b>	Analysis of variance
<b>APA</b>	American Psychiatric Association
<b>AS-18</b>	Affective Self Rating Scale
<b>BDI</b>	Beck Depression Inventory
<b>BISS</b>	Bipolar Inventory of Symptoms Scale
<b>CGI</b>	Clinical Global Impression scale
<b>CGI-BP</b>	Clinical Global Impression scale modified for bipolar illness
<b>CTT</b>	Classical Test Theory
<b>DIF</b>	Differential Item Functioning
<b>DSM-5</b>	Diagnostic and Statistical Manual of Mental Disorders, fifth edition
<b>DSM-IV</b>	Diagnostic and Statistical Manual of Mental Disorders, fourth edition
<b>HCL-32</b>	Hypomania Checklist-32
<b>HDRS</b>	Hamilton Depression Rating Scale
<b>HIGH-C</b>	Hypomania Interview Guide
<b>ICD-10</b>	Tenth edition of the International Classification of Disorders
<b>IIO</b>	Invariant item ordering
<b>IRF</b>	Item Response Function
<b>IRT</b>	Item Response Theory
<b>MADRS</b>	Montgomery-Åsberg Depression Rating Scale
<b>MDQ</b>	Mood Disorder Questionnaire
<b>MES</b>	Bech-Rafaelsen Melancholia Scale
<b>PHQ9</b>	The depression module from the Patient Health Questionnaire

<b>QIDS-16</b>	Quick Inventory of Depressive Symptomatology
<b>RCT</b>	Randomized Controlled Trials
<b>RCT-AD</b>	Randomized Controlled Trials of Antidepressants
<b>ROC</b>	Receiver Operating Characteristics
<b>TIF</b>	Test Information Function
<b>VAS-scale</b>	Visual analogue scale
<b>WHO</b>	World Health Organization
<b>YMRS</b>	Young Mania Rating Scale
<b>1-PL</b>	One parameter model of IRT
<b>2-PL</b>	Two parameter model of IRT
<b>3-PL</b>	Three parameter model of IRT

# 1 INTRODUCTION

In psychiatry, clinical assessment of symptom severity is being increasingly supplemented by systematic measurement using rating scales. Such scales are used for many purposes including assessment of patients, obtaining data for research and quality control by health authorities.

Transforming the subjective symptoms of psychiatric disorders into meaningful and valid numerical measures is a complex process, subjected to numerous confounding factors. Ratings on items may be influenced by other factors than the intended. Items might be ambiguous, different social groups often comprehend items differently and respondents or raters might even intentionally distort the measurement.

Misleading measurements can have serious consequences. The conditions of patients could be misinterpreted, leading to erroneous decisions concerning treatment. Imprecise or distorted data will mislead scientific studies. If data from rating scales are systematically biased, health authorities will base decisions on faulty assumptions. It is therefore of vital importance that psychiatric rating scales provide reliable and valid measurement of psychiatric disorders. Careful evaluation of rating scales is therefore essential for psychiatric practice, evaluations of health care, and for research.

This doctoral project arose from a clinical need for a useful self rating scale for affective symptoms at the outpatient clinic for affective disorders situated at Psychiatry Southwest, Karolinska University Hospital Huddinge in Stockholm, Sweden. The clinic predominantly serves patients with bipolar disorder diagnoses. At appointments the patients can be in any mood state: normal, minor depressive, major depressive, hypomanic, manic or a state with where manic and depressive type symptoms are intermingled (mixed state). There was therefore a need for a self rating scale for measurement of symptom severity, useful in all these mood states, as a tool for the routine assessment at clinical follow up. A literature review indicated that there were no existing rating scales that could fulfil the clinical need. Initially in the project, the traditional methods for the evaluation of rating scales, called Classical Test Theory (CTT), were used. During the project modern methods for the construction and evaluation of rating scales, called Item Response Theory (IRT), were introduced.

## **1.1 AIMS OF THE DOCTORAL PROJECT**

The aims of the project were:

1. to develop and evaluate a self rating scale for simultaneous measurement of severity in depressive, manic and mixed affective states.
2. to explore if IRT-methods are useful for evaluation and improvement of rating scales for mania and depression.
3. to compare the psychometric properties of the depression subscale of the new scale with the depression module from the Patient Health Questionnaire (PHQ9)<sup>1</sup> and the Montgomery-Åsberg Depression Rating Scale (MADRS).<sup>2</sup>
4. to investigate if weak measurement properties of the Hamilton Depression Rating Scale<sup>3</sup> (HDRS) might explain the findings of low or absent efficacy of antidepressant medication in less severe depression in Randomized Controlled Trials (RCT).

## **1.2 OVERVIEW OF THE PROJECT**

A self rating scale consisting of 18 items was developed and named the Affective Self Rating Scale (AS-18, see appendix). It was evaluated in two patient samples. The first sample consisted of 61 patients, and the analysis was based on statistical methods from CTT (study 1). It was re-evaluated in a second sample of 231 patients, using IRT-methods (study 2).

In the third study the depression subscale from AS-18 and PHQ9 were evaluated using IRT-methods and compared to the MADRS in the same patient sample as the first study. The fourth study reanalyzed data from five Randomized Controlled Trials of antidepressants included in a meta-analysis (N=516).

## 2 CLASSIFICATION AND SYMPTOMS OF MOOD DISORDERS

A primary consideration regarding the measurement of a disorder is the clinical presentation of the disorder. The clinical presentation indicates which symptoms are to be measured. The aim of this study was to measure symptoms of mania and depression, the main syndromes of mood disorders.

### 2.1 CLASSIFICATION OF MOOD DISORDERS

Mood disorders have modern and authoritative definitions in the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) by the American Psychiatric Association (APA) and the International Classification of Disorders, tenth edition (ICD-10) by the World Health Organisation (WHO).<sup>4</sup> This group of disorders is characterized by episodes of pathological alternations of mood accompanied by different combinations of altered cognition, behavior and bodily functions. These episodes are typically interspersed with periods of normal mood and functioning.<sup>5</sup> The two main types of mood disorders are *major depressive disorder* and *bipolar disorder*.

In major depressive disorder (or unipolar depression) the patient suffers from *depressive episodes*, defined as episodes of depressed mood of sufficient duration, number of symptoms and consequences in terms of distress or impairment. In bipolar disorder the patient suffers from episodes of elevated mood and usually also depressive episodes. If the episodes of elation are severe, they may fulfill criteria for a *manic episode* and the disorder is classified as *bipolar disorder type I*. If the episodes of elation are less severe, without obvious negative consequences or psychotic features, they are called *hypomanic* and the disorder is classified as *bipolar disorder type II*, in which the depressive episodes are the main clinical problem. In the manuals the mood disorders are further subclassified and given different specifications according to symptom type, course and relation to different situational contexts. Other types of mood disorders are also described.

According to the National Comorbidity Survey Replication from 2005 the lifetime prevalence of mood disorders was 20.8 %, of which 3.9 % were of a bipolar nature.<sup>6</sup> A majority of patients with affective disorder will suffer relapses in their disorder.<sup>7-8</sup> There is evidence that the risk of recurrence in affective episodes increases with the previous number of episodes.<sup>9</sup> The psychosocial impairment associated with mania and major depression extends to most areas of functioning and persists for many years.<sup>10</sup> Increased numbers of episodes increases the risk of cognitive impairment.<sup>11-12</sup> The risk of suicide is at least 20 times higher in bipolar patients compared to the

general population.<sup>13</sup> Studies on inpatients suffering from major depressive disorder have shown a 15% rate of suicide.<sup>14</sup>

## 2.2 SYMPTOMS OF DEPRESSION AND MANIA

The definitions of bipolar disorder in both the DSM-IV and the ICD-10 (research version) include symptom criteria with descriptions of symptoms considered typical for the disorder. As can be seen from table 1 and 2 some symptom criteria are “double-barrelled”, describing more than one symptom. The number of symptoms is therefore higher than the number of symptom criteria in the diagnostic manuals.

**Table 1. Symptoms of depression included in DSM-IV and ICD-10 criteria**

	DSM-IV	ICD-10	No of symptoms
1. Depressed mood, or diminished interest or pleasure	X	X	3
2. Loss of energy or fatigue	X	X	2
3. Loss of confidence or self-esteem	X	X	2
4. Unreasonable feelings of self-reproach or inappropriate guilt	X	X	2
5. Recurrent thoughts of death, or suicide or any suicidal behavior	X	X	3
6. Diminished ability to think or concentrate, or indecisiveness	X	X	3
7. Psychomotor agitation or retardation	X	X	2
8. Insomnia or hypersomnia	X	X	2
9. Change in appetite (decrease or increase)	X	X	2
<b>Total</b>			<b>21</b>

**Table 2. Symptoms of mania and hypomania included in DSM-IV and ICD-10 criteria**

	DSM-IV	ICD-10	No of symptoms
1. Elevated mood, expansive or irritable mood	X	X	3
2. Inflated self-esteem or grandiosity	X	X	2
3. Decreased need for sleep	X	X	1
4. Increased talkativeness	X	X	1
5. Flight of ideas or subjective racing thoughts	X	X	2
6. Distractibility	X	X	1
7. Increase in goal-directed activity or psychomotor agitation	X	X	2
8. Excessive involvement in pleasurable activities that have a high potential for painful consequences	X	X	1
9. Increased sexual activities		X	1
<b>Total</b>			<b>14</b>

## **2.3 MIXED STATES**

A special form of affective state, defined in both the DSM-IV and ICD-10, is mixed episode (or mixed state), where symptoms of mania and depression are intermixed during the same episode. Mixed states are given a very restricted definition in the DSM-IV, requiring that the criteria for both a manic and depressive episode must be fulfilled during the same week, making this type of episode uncommon.

Mixtures of fewer symptoms from one affective state into the other are common, however. This type of mixed states has been reported in 5-73% of patients with mania, depending on varying definitions of mixed state and properties of the populations studied.<sup>15</sup> In a long time follow up study of patients with bipolar disorder, 57% of the hypomanic episodes recorded were of mixed type.<sup>16</sup> In a study of depression the prevalence of mixed depression ranged from 47% to 72% in bipolar II patients, and from 8% to 42% in patients with unipolar depression, depending on the definition of depressive mixed state.<sup>17</sup>

This type of mixed states has high clinical relevance. Mixed depression is associated with high risk for manic transition among both antidepressant-treated and antidepressant-untreated individuals.<sup>18</sup> The mixed form of mania has been associated with high grades of suicidality compared to pure mania.<sup>19</sup> Depression with mixed features has an illness course marked by longer duration and a lower grade of remission than pure depression in bipolar disorder.<sup>20</sup>

The diagnosis of mixed states is more difficult than the diagnosis of “pure” affective states. Mixed states may easily be mistaken for other types of affective states in mental disorders such as Attention Deficit Hyperactivity Disorder (ADHD), personality disorders and schizophrenia.<sup>21</sup> Almost all existing rating scales for affective disorders are developed for the assessment of either depression or mania. Mixed states can therefore easily be overlooked and the patient consequently misdiagnosed.

## **2.4 SPECIAL CONSIDERATIONS FOR BIPOLAR DEPRESSION**

There is evidence for differences in symptom expression between depression in bipolar and unipolar disorder. Symptoms of hypersomnia, increased appetite, psychomotor disturbances, mood lability and psychotic features have repeatedly been found more frequently in the bipolar form.<sup>22-23</sup> As a consequence it has been questioned whether rating scales constructed for the measurement of unipolar depression are valid tools for the measurement of bipolar depression.<sup>24</sup> A recent study of a very large and representative sample (N = 13058), using IRT-methods, concluded that the differences in symptom presentation between depression in bipolar and major depressive disorder are small and subtle.<sup>25</sup> Clinically, this implies that symptom profiles will be at most a suggestive

clue to the differentiation of bipolar from unipolar depression. It also suggests that the same rating scales might be useful for both bipolar and unipolar depression. In a study of the depression rating scale Quick Inventory of Depressive Symptomatology (QIDS-C16), using both CTT and IRT methods, no differences in scale functioning was found between unipolar and bipolar patients.<sup>26</sup> The authors questioned the need for a separate scale for bipolar patients if core symptoms of depression were used as base for measurement.

The question about the relevance of the differences in clinical presentation between bipolar and unipolar depression cannot be said to be completely settled.

## **2.5 OVERLAP OF SYMPTOMS OF MANIA AND DEPRESSION**

Another complicating circumstance for the measurement of symptom severity in bipolar disorder is the overlap of the symptoms of mania and depression. Decreased amount of sleep and decreased appetite are common symptoms in mania as well as in depression. Symptoms of manic distractibility and depressive concentration difficulties are not identical, but difficult to distinguish both in interview and self-rated assessment. Irritability is a main symptom in the DSM-IV definition of mania and hypomania, but is also common in unipolar depression.<sup>27</sup> If symptoms common to both depression and mania are included in a rating scale, scores can be affected by symptoms from the opposite syndrome, even in non-mixed states.

## **2.6 EXISTING RATING SCALES FOR AFFECTIVE SYMPTOMS**

Many of the most commonly used rating scales for affective disorder were developed many decades ago, before the development of modern psychometric techniques and before modern definitions of depression and mania were introduced.

### **2.6.1 Rating scales for depression**

The two most commonly used interview based rating scales for depression severity in scientific use are the Hamilton Depression Rating Scale (HDRS) and the Montgomery-Åsberg Depression Rating Scale (MADRS).<sup>2-3</sup> Another commonly used rating scale for depression is the Bech-Rafaelsen Melancholia Scale (MES).<sup>28</sup>

There are many self rating scales for depression, of which the Beck Depression Inventory (BDI) is probably the most widely used.<sup>29</sup> Another self rating scale is PHQ9, a modern rating scale built on the definition of depressive episode in DSM-IV.<sup>1, 30</sup> PHQ9 has been proposed as a general measure

of depression severity in unipolar as well as in bipolar disorder by the APA task force for the development of the forthcoming fifth edition of the DSM (DSM-5).<sup>31</sup> If the proposal is accepted by the APA, PHQ9 would obtain a central role in the measurement of depression severity worldwide. The Bipolar Depression Rating Scale (BDRS)<sup>24</sup> is a interview based rating scale developed with the aim to capture the specific symptom profile of bipolar depression and depression with mixed features. It has been validated in a study of depressed bipolar patients, by means of CTT-methods, indicating a good internal validity and strong correlations with the MADRS and the HDRS. The subscale for depression with mixed features correlated with the Young Mania Rating Scale (YMRS).<sup>24</sup>

### **2.6.2 Rating scales for mania**

Two commonly used rating scales for mania are the Young Mania Rating Scale (YMRS) and the Mania Rating Scale (Bech and Rafaelsen).<sup>32-33</sup> A seldom used alternative is the Hypomania Interview Guide (HIGH-C), especially developed for less severe manic type symptoms.<sup>34</sup>

There is a shortage of self rating scales for manic symptoms, partly due to doubts that a patient with mania can have sufficient self-awareness for a credible self rating. An attempt to create such a self rating scale was the Altman self rating Mania Scale.<sup>35</sup> Two other self rating scales for manic symptoms considered in this doctoral project were the Mood Disorder Questionnaire (MDQ) and the Hypomania Checklist-32 (HCL-32).<sup>36-37</sup> However, these scales were developed for the purpose of screening for manic and hypomanic episodes, not for measurement of symptom severity. They were therefore judged less relevant for the purpose of this project.

### **2.6.3 Rating scales for mixed states**

The Bipolar Inventory of Symptoms Scale (BISS)<sup>38</sup> is an interview based rating scale developed with the purpose to capture the full spectrum of bipolar symptoms, including mixed states. It has been validated in a bipolar sample, using CTT-methods. The study indicated that BISS fulfilled criteria for convergent validity, discriminant validity and correlated well to the MADRS and the Global Assessment of Functioning Scale (GAF).<sup>38</sup> The only self rating scale with the aim to measure both depression and mania type symptoms together which was found during the preparatory work for this doctoral project was the Chinese polarity scale.<sup>39</sup> The polarity type of the items used in this rating scale was, however, found unsuitable in our preparatory evaluation, due to the flexing nature of mania and mixed states which made our patients chose two or more points on the bipolar type of

response formats. The decision was instead taken to develop and validate a self rating scale that could distinguish and measure the intensity of depressive, manic as well as mixed states.

### 3 MEASUREMENT THEORY

Measurement is fundamental to science. Measurement means that the quantity of a phenomenon is transformed into meaningful numbers according to rules.<sup>40,41</sup> Many physical sciences have developed highly accurate and replicable measurement. Examples are measurement of weight, length and temperature.

In contrast to many physical phenomena the symptoms of mania and depression, like most symptoms of interest in psychiatry, are subjective and not directly observable. In measurement theory such unobservable phenomena are called *latent variables* or *constructs*.

#### 3.1 MEASUREMENT OF LATENT VARIABLES

Although a latent variable cannot be observed, it can be inferred by observation of indirect *manifest variables* that are observable. For example, the level of anxiety in a person is inferred by observations of manifest variables like tremor, motor restlessness or sweating. A psychological approach to the measurement of psychological phenomena is to use a *rating scale* where the patients report their symptoms. A rating scale consists of a set of questions or statements (usually called “items”) where the responses can be categorized with the purpose of measuring a latent variable.

##### 3.1.1 Different purposes of rating scales

Rating scales are designed for different purposes:

- Some rating scales are designed for *screening* of psychiatric disorders, that is to say that they aim to detect cases where a diagnosis can be suspected.
- Other scales are designed for *diagnostic purposes*, namely to assess whether the criteria for a diagnosis are met.
- A third type of rating scales are intended for the *measurement of symptom severity*.

This doctoral project concerns the measurement of symptom severity in mania and depression.

##### 3.1.2 Types of latent variables

Latent variables can be conceptualized as either *categorical* or *continuous*. A *categorical variable* implies that there are qualitative differences in the phenomena so that patients would differ in kind on the latent variable. Such differences in kind are often referred to as *latent classes*. The

classification of psychiatric disorders is an example of a system intended to define latent classes. Rating scales for screening purposes are designed with the intention to discover suspected cases of a latent class. Rating scales for diagnostic purposes are designed to verify whether a case fulfills criteria for inclusion in a latent class or not. *Continuous variables* can be conceptualized as a *latent dimensions* from low to high. Rating scales of symptom severity are tools for measurement of latent dimensions. For example, PHQ9 is an instrument for measurement of the latent dimension of depression.

### **3.1.3 Different forms of rating scales**

One of the most common rating scale formats is the Likert scale. Each item in a Likert scale is a statement or a question which the respondent is asked to evaluate according to some sort of criteria. A common format is the level of agreement or disagreement to a statement. Often a response format with five options is used:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

There are other formats of Likert-type response options; for example, how often a phenomenon is experienced or the intensity of a phenomenon. Most Likert scales use predetermined sets of response categories identical for all items included in the rating scale, like the disagree/agree example above. Other rating scales include different texts for the response options in the items, so called “anchor points”. Two of the rating scales used in the studies of this doctoral project, the MADRS and the HIGH-C, use anchor points in their Likert-type response formats.

There are other rating scale formats. A scale can have a series of open questions where the answers are assessed according to a model for correction and given ratings, like correct (1) or incorrect (0). The visual analogue scale (VAS-scale), typically a 10 centimeter line where the extreme positions are described at the end-points, is another alternative. In the VAS-scale, the respondent is asked to choose a position between these extremes that is usually described numerically by the distance in centimeters or millimeters from the lower end point.

### 3.1.4 Different ways of collecting information

Rating scale types can also be classified according to the source of information. The gold standard in most research is the assessment by a health professional (interview or observation rating), who ideally should be trained in the use of the rating scale. Ratings can also be done by the individual being assessed (self rating) or by an informant (informant rating), usually a relative or a close friend.

### 3.1.5 Different types of data

Different types of measurement produce different types of data. According to a classical definition data can be classified into one of four types:<sup>40</sup>

1. *Nominal level data* (or categorical) uses numbers to classify objects or phenomena in classes. This means that the numbers are used as labels and lack mathematical meaning. Nominal data are often used for labelling of latent classes. An example is when psychiatric diagnoses are given ICD- or DSM-codes.
2. *Ordinal level data* describes the order, but not the relative size or degree of difference between the phenomena being measured. In an ordinal measurement we know that the number of two represents a higher level than the number of one, but not the distance between the numbers. Likert-type rating scales produce numbers that represent this type of ordered categories.
3. *Interval level data* does not only reveal the ordering, but also defines the distances between positions in the rating scale. The measurement of temperature by the Celcius scale is one example of interval data.
4. *Ratio level data* indicates both that there are equal distances between positions and that there is an absolute zero point. Examples of ratio level data are age and distances.

### 3.1.6 Restrictions on calculations

The type of data puts restrictions on the type of calculations that should be used.<sup>40</sup> Nominal data can be described by the numbers of different classes or the most common class (the mode). Ordinal data allows more mathematical operations. The data can be described by measures of central tendency, like *mode* (the most common item) or *median* (the middle-ranked item). Special statistical methods for ordinal data exists, called *non-parametric*, implying that they do not assume equal distances between numbers.

Interval data allows most statistical operations to be performed including mean, standard deviation, regression, factor analysis and analysis of variation (ANOVA). While interval data cannot be multiplied or divided, such operations are possible for the ratio type of data. The mathematical procedures aimed at interval and ratio data are called *parametric*.

Usually the measurement of symptom severity in a rating scale is obtained by the sum of the numbers assigned to the response options of the items in the scale. This practice relies on implicit assumptions that often are overlooked. First, since a change in score in one item can be nullified by a change in any other item, the use of summed score relies on the assumption of equidistance's between numbers. The summed score method is therefore, in itself, a method where interval assumptions are applied to ordinal data. Second, when the summed score of the items of a rating scale is used as a single score for a person it implies that the sum is intended to measure a single latent variable, often referred to as a *unidimensional variable*.

Strictly, since rating scales are of ordinal type, only non-parametric statistical operations should be performed. It is however not uncommon that parametric statistics are used for rating scale data. In many cases such calculations yield the same results as non-parametrical statistics but might in other cases create problems in the calculations.<sup>42</sup>

## **3.2 KEY CONCEPTS FOR EVALUATION OF RATING SCALES**

Before being used in clinical work or in research, a rating scale should be evaluated. Key concepts in the evaluation are *reliability* and *validity*.

### **3.2.1 Reliability**

The concept of *reliability* refers to the correlation of an item or a full rating scale with a hypothetical measurement method which would truly measure the phenomenon of interest.<sup>43</sup> By definition no instruments can directly measure latent variables, so the reliability of a rating scale is a theoretical concept that cannot be fully verified. Consequently such measures of reliability should be regarded as *estimates*. In practice the concept of reliability refers to the extent to which a measurement gives consistent results.

### **3.2.2 Validity**

Validity of a rating scale is, in general terms, the degree to which the scale measures what it claims to measure. The word has its origins in the Latin word *validus*, meaning strong. A traditional view is to see validity as a three-faced concept: content validity, criterion validity and construct validity.<sup>44</sup>

An updated authoritative view of the concept of validity is *The Standards for Educational and Psychological Testing* published by the American Psychological Association and the National Council on Measurement in Education.<sup>45</sup> In the perspective of these standards, *construct validity* is the overall concept, defined as: "the degree to which evidence and theory support the interpretations of test scores".<sup>45</sup> In *The Standards for Educational and Psychological Testing* five types of evidence for construct validity are listed.

1. *Test content*: whether the test includes all the important aspects of the construct. A test may include elements that are actually irrelevant to what it is supposed to measure or lack important facets of the construct. This aspect of construct validity is often established by reviews of expert panels.
2. *Internal structure of the test*: an investigation of the way the elements (often items) are related to each other. For example can the assumption of unidimensionality be tested by statistical methods.
3. *Response processes*: the extent to which test takers demonstrate that they understand the construct in the same way as it is defined and follow the intended procedure of the test. This assumption can be evaluated by interviews of respondents concerning their understanding of the test. There are also statistical procedures for the evaluation of the structure of responses that can reveal unexpected response patterns.
4. *Associations with other variables*: the degree to which the results of a test (for example a rating scale) associate to other variables as predicted by the theory behind the construct. *Convergent evidence* refers to the degree to which a measure from a rating scale is correlated to related phenomena. It is, for example, to be expected that ratings of depression are correlated to the risk of suicide. A common way to evaluate a new rating scale is to investigate its correlation to an established rating scale. *Discriminant evidence* is the degree to which test results are uncorrelated to such constructs it theoretically should not be similar to. Associations can be tested statistically by calculation of correlation coefficients between two variables.
5. *Consequences of testing*: a modern addition to the validity concept, meaning that the social consequences of using a test should be included in a full validity evaluation. This addition has been criticized as being a political intrusion into science, but can also be viewed as a safeguard against biases in the test content that would discriminate for example women or a particular ethnic group.

In summary, the concept of validity concerns the inferences made from the measurement and is not a property of the rating scale itself. From this point of view, establishment of construct validity of a rating scale is not an all or none decision, but rather a process where different types of information are used. Statistical methods are very important in evaluations of reliability and many aspects of validity. Some aspects of validity, like test content, response processes, and the consequences of testing, are to a large extent evaluated with other methods like interviews, expert reviews, or analyses of social aspects of testing.

## 4 STATISTICAL METHODS FOR THE EVALUATION OF RATING SCALES

In this doctoral project two groups of statistical methods for the evaluation of rating scales are utilized: Classical Test Theory (CTT) and Item Response Theory (IRT).

### 4.1 CLASSICAL TEST THEORY (CTT)

CTT is an umbrella term for the traditional statistical methods for evaluation of rating scales. CTT is also called “true score theory”, referring to its basic assumption that the true score on a test is solely dependent on the summed score of the rating scale plus measurement error.

#### 4.1.1 Reliability in CTT

Reliability in CTT refers to the consistency of measurement.<sup>44</sup> This is usually estimated in one of four ways.

1. *Internal consistency*: correlation among items in the scale. *Cronbach's alpha* is the most commonly used measure of internal consistency. Alpha measures the extent to which item responses obtained at the same time correlate highly with each other. More specifically, alpha is a measure of the mean intercorrelation between items weighted by variances.<sup>1</sup>
2. *Split-half reliability*: a procedure where a test is split in two and the scores for each half of the test are compared with one another.
3. *Test-retest reliability (or stability)*: the correlation between ratings with the same rating scale at different occasions. This type of reliability test presupposes that the properties of the populations measured do not change between the occasions. Measures of concepts like personality are expected to change little over time and the test-retest stability can be expected to be high. In mood disorders the condition often changes, sometimes even from day to day, and the stability in such cases are expected to be lower.
4. *Inter-rater reliability*: the correlation between raters on the same scale.

---

i)  $\alpha = k/(k-1) [1 - \sum(s_i^2) / s_{\text{sum}}^2]$  where  $K$  is the number of components,  $s_{\text{sum}}^2$  is the variance of the observed total test scores and  $s_i^2$  the variance of component  $i$  for the current sample of persons.

The different methods for estimation of reliability may yield different results as they represent different aspects of reliability.

#### **4.1.2 Validity in CTT**

In CTT, the assumption of unidimensionality can be tested by a statistical process called *factor analysis*. Such an analysis searches for joint variations in a set of variables with the aim of reducing the number of variables to a lower number of unobserved variables called *factors*. If a factor analysis results in one (dominant) factor, this is regarded as evidence of unidimensionality. Factor analysis can also be used for assessing discriminant validity. If two sets of variables are intended for the measurement of different latent variables, a factor analysis can give support for the assumption that a rating scale can separate these variables.

Correlation is another method used in CTT in order to measure the degree to which two sets of variables, supposed to be associated, actually relate. A high correlation is evidence of concurrent validity. Conversely, low correlation can be used as evidence of discriminant validity. There are methods specifically constructed for estimating the correlation between ordinal data within the CTT tradition (Spearman's rho).

#### **4.1.3 Advantages of CTT-methods**

CTT-methods give overall measures of important aspects of the reliability and validity of a rating scale including its factor structure, internal consistency and test-retest stability. These estimates are easily accessible in most statistical programs and are familiar to researchers and editors of scientific journals.

#### **4.1.4 Shortcomings of CTT-methods**

The methods used for evaluation in CTT have shortcomings.<sup>46</sup> Most CTT-methods are developed for interval data, while rating scales produce data of ordinal type. This may cause problems in the analyses. Furthermore, since most CTT evaluations of rating scales rely on the summed score, they give only limited information about the performance of the individual items in the scale.

There are several potential confounders to CTT-evaluations of rating scales.

- In a CTT-analysis, a scale may have good reliability estimates but still fail to meet the criteria for unidimensionality.<sup>47</sup> An example of this is that a rating scale can show a high Cronbach's alpha and still be multidimensional when there are separate clusters of items reflecting separate latent variables which intercorrelate highly, even though the clusters themselves do not show high intercorrelation.<sup>43</sup>

- Since the number of items are included in the formula for alpha, the reliability estimate increases if more items are added to the rating scale, even if the new items are redundant and in essence ask the same questions to the respondent.
- Factor analysis may falsely imply multidimensionality in a scale even if the scale is truly unidimensional if the positions of the items on the dimension are widely separated.<sup>48</sup>

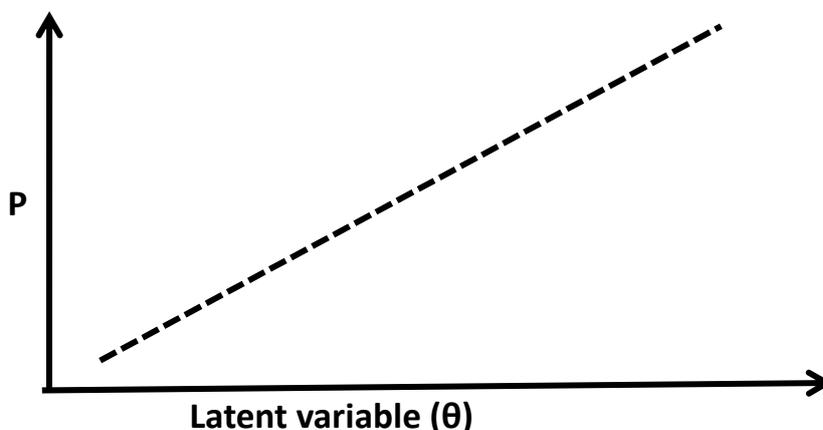
## 4.2 ITEM RESPONSE THEORY

During the last decades a new group of methods, called Item response Theory (IRT), has been used increasingly, especially in educational measurement but also in health measurement. IRT-methods were especially developed for the construction and evaluation of rating scales for latent variables of nominal and ordinal type.

While CTT-methods rely on the summed score, the basis for IRT-methods is the relationship between the latent variable and the probability of endorsing a higher response option of an item. For example the probability of endorsing items like “I feel down” and “I am considering suicide” would be expected to increase with increasing levels of depression. An IRT-analysis starts by an investigation of the probabilities for endorsing the response options of the items of a rating scale in a relevant sample.

In IRT the latent variable often is called theta ( $\theta$ ) and the probability of endorsing a higher response option of an item rather than a lower option is called P. The relationship between the latent variable  $\theta$  and the probability P can be displayed in a graph called the Item Response Function (IRF). If an item has a capacity for measurement of a latent variable, the probability of endorsing higher response options increases with an increasing latent variable (figure 1).

**Figure 1**



*The Item Response Function (IRF) is a curve describing the relationship between the latent variable ( $\theta$ ) and the probability (P) of endorsing a higher response option of an item over a lower response option.*

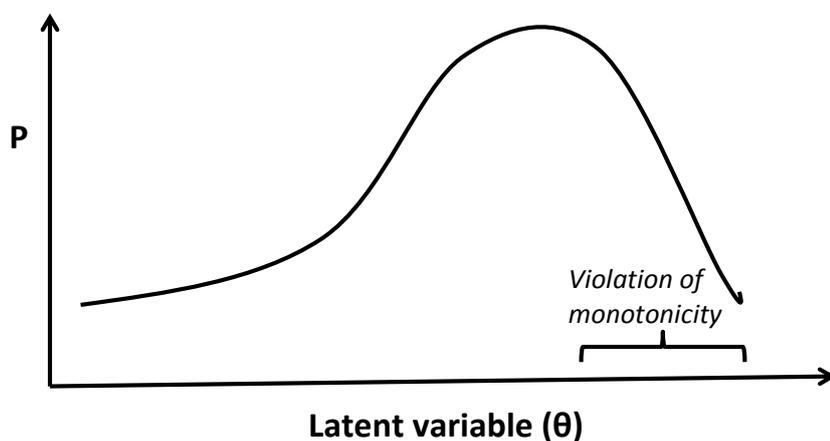
### 4.2.1 Criteria for a well functioning rating scale in IRT

A well functioning rating scale should fulfill certain characteristics that can be tested with IRT-methods.

First, a rating scale usually is intended to measure only one underlying latent variable. If it does, it meets the requirement of *unidimensionality*. Full unidimensionality is an unattainable ideal, but a questionnaire should have as little influence as possible from other latent variables.

Second, *monotonicity* is of vital importance for a well functioning item in a rating scale. This means that the probabilities of endorsing the ordered categories of a particular item are continuously increasing as the latent variable increases, like in figure 1. An item with decreasing probability with increasing latent variable would counteract measurement (figure 2).

Figure 2



The graph shows the IRF of an item that violates the assumption of monotonicity in the upper part of the latent variable, since the probability of endorsement of the item decreases when the latent variable increases.

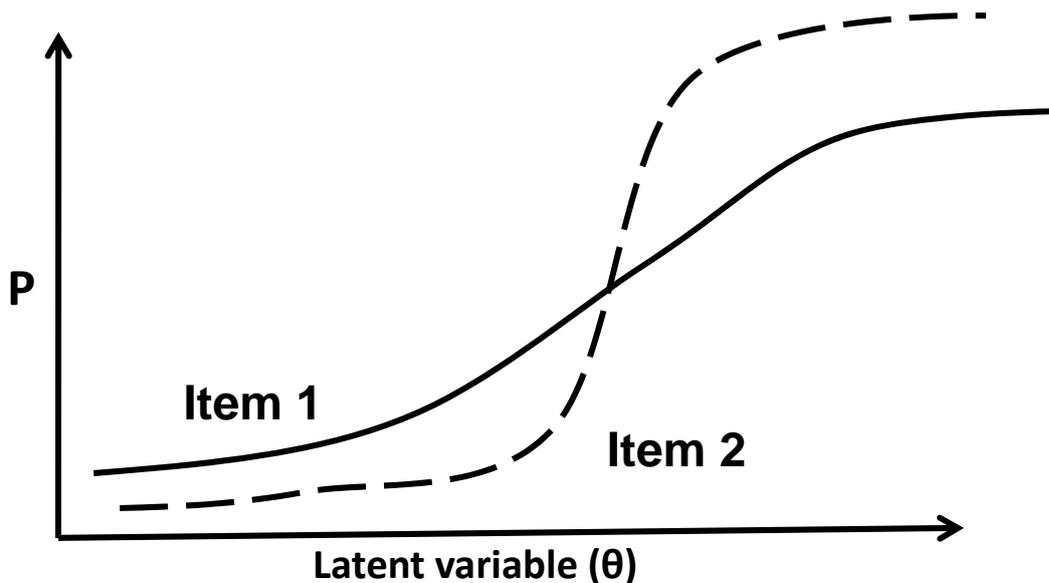
Third, a high degree of *precision* is required for the rating scale to be able to reliably separate respondents at different levels of severity on the latent dimension. This is achieved if the items in the scale are subject to low degrees of random variation.

Fourth, the rating scale should provide *equal reliability* over the full range of severity of the latent dimension. This property is optimized if the items in the rating scale have their optimal measurement level evenly spread out over the dimension.

Fifth, the items in the rating scale should be *locally independent*, meaning that the responses to items are independent of responses to other items after controlling for the latent variable. If two items are perceived as very similar, the responses will be dependent on each other and violate the assumption of local independence.

Sixth, and finally, *invariant item ordering* (IIO) is a desirable property of a rating scale. IIO means that the ordering of item locations on the latent variable are the same for respondents at all levels of the latent dimension (figure 3). IIO is a requirement for a summed score to be valid on all levels of severity of the underlying dimension and for comparisons between different patient groups.

Figure 3



Since the IRFs of the items intersect, the two items violate the assumption of Invariant Item Ordering (IIO). This means that the order of the items is not the same for respondents on different levels of severity.

#### 4.2.2 Different types of IRT

There are several different IRT-methods.<sup>49</sup> An important dividing line in IRT is between non-parametric and parametric IRT-methods.

In non-parametric IRT the primary data are analyzed in their ordinal form. Non-parametric IRT-methods are often used as a first step in the evaluation of rating scales since such methods provide valuable overall indicators for the quality of a rating scale and that they have the capacity to adequately accommodate most datasets.

In parametric forms of IRT a mathematical formula (a “model”) is created that describes the relationship between the latent variable and the probability of endorsing the set of items in the rating scale. This formula can be used for evaluations of the performance of a rating scale and for measurement of respondents (patients, examinees etc.). Parametric IRT-models are more restrictive in relation to data than non-parametric IRT, since they demand that data show an acceptable

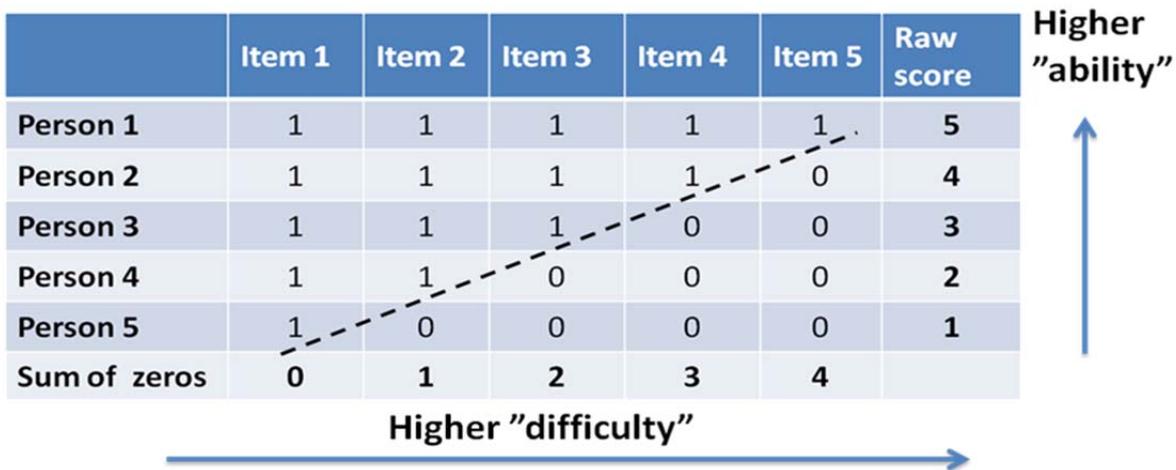
conformity to the mathematical model. The conformity of the data to the model, and the applicability of a chosen parametric model, can be tested by so called fit statistics.

### 4.2.3 Non-parametric IRT

Non-parametric methods put few restrictions on the data. This flexibility is graphically manifested in that the IRF-curves created by non-parametric methods are allowed to take any form.

The Mokken method of non-parametric IRT uses covariances between ordered categorical variables, to yield a set of *scalability coefficients* for the items in a rating scale as well as for the full rating scale. The overall scalability coefficient H is a measure of whether the response pattern of the items in the rating scale is sufficiently predictable to support the assumption that the scale is adequately measuring one underlying latent variable (figure 4 and 5).

Figure 4



The response pattern of a rating scale can be arranged in a scalogram. This scalogram shows responses of five persons on five items with the response options 0 or 1. The scalogram is arranged with items with a lower rate of endorsement horizontally (higher "difficulty") and persons with higher scores vertically (higher "ability"). In this example endorsement (=1) of a given item predicts endorsement to all previous items in the series. This is called a "perfect Guttman pattern", meaning a completely regular pattern of responses, illustrated by the dotted line. Below the line all items are scored 0 and above all items are scored 1.

Figure 5

	Item 1	Item 2	Item 3	Item 4	Item 5
Person 1	1	1	0	1	1
Person 2	1	1	1	1	0
Person 3	1	1	1	0	0
Person 4	1	0	1	0	0
Person 5	1	0	0	0	0

This scalogram shows a more probable pattern with deviations from the perfect Guttman pattern. In the Mokken method, the scalability coefficient  $H$  is a measure of how well the data conforms to the perfect Guttman pattern.  $H = 1$  means a perfect Guttman pattern and  $H = 0$  a totally random pattern. Guidelines for interpretation of scalability  $H$ :  $\geq 0.5$  indicates a strong scale,  $0.4 \leq H < 0.5$  a medium scale and  $0.3 \leq H < 0.4$  a weak scale. Below 0.3 the credibility of ranking of persons according to their total score on the rating scale becomes increasingly doubtful. An item scalability coefficient  $H_i \geq 0.3$  indicates that the item is contributing adequately to the measurement.<sup>50</sup>

The analysis of the scalability indexes starts by calculating the covariances between all item pairs ( $H_{ij}$ ) weighted against the maximal possible covariances given existing differences in rates of endorsement (“difficulties”). In the next step the covariances of each individual item to all other items are calculated, producing item scalability coefficients ( $H_i$ ). Finally the scalability for the combined set of items in the intended scale is analyzed, forming the overall scalability measure ( $H$ ).  $H$  is a measure of the quality of the measurement and reflects the capacity of the full scale to rank persons according to their total score on a latent variable.<sup>51</sup> The item scalability coefficients reflect each items contribution to the measurement. The scalability coefficients are negatively influenced by *high random variance, multidimensionality and deficiencies in monotonicity*. There are recommended values for interpretation of the scalability coefficients (see legend of figure 5).<sup>51</sup>

Monotonicity and IIO can be specifically tested within the Mokken framework. The Mokken family of methods also includes a procedure called “automated item selection”, which brings items with high covariances together, a process with a purpose similar to traditional factor analysis.<sup>52</sup>

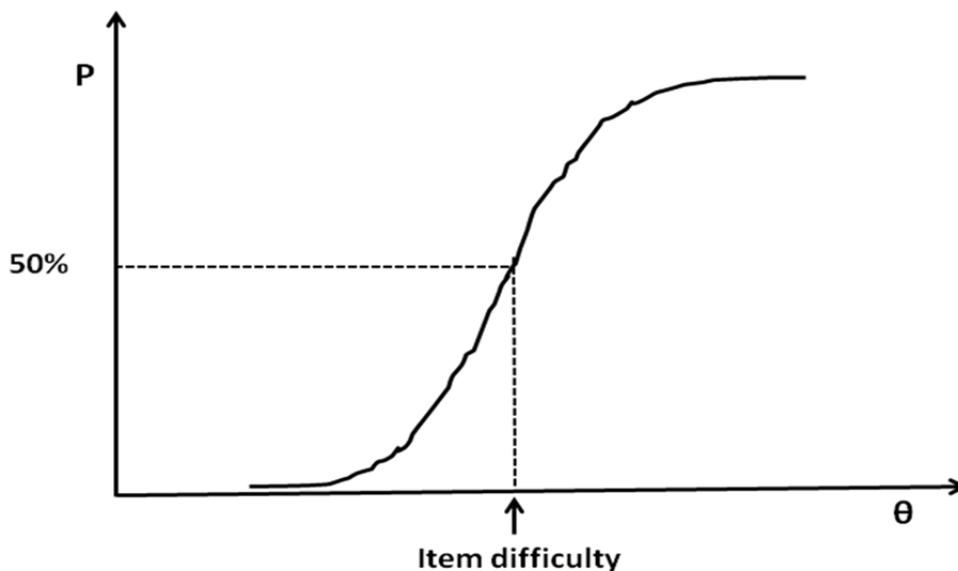
The Mokken methods are suitable first steps in the analysis of a rating scale since they provide robust indicators of the overall quality of the measurement. The Mokken analysis can also indicate whether parametric IRT-methods are suitable for further analysis. It is applicable in situations where parametric IRT-models misfit the data. Parametric models may in such cases wrongly indicate a

poor performance of an item, while the non-parametric analysis demonstrates that the item is useful for measurement anyway.<sup>50, 53-56</sup>

#### 4.2.4 Parametric IRT

Parametric models are built on the premise that it is possible to formulate a mathematical function that adequately describes the probability of respondents, at different levels of the dimension, to endorse a response option in the rating scale. In most parametric models the probability of endorsement is expressed as a logistic function, causing the IRF-curve to take a typical “s-shape”. In a rating scale the items usually will have IRF-curves at different levels, indicating that the items are sensitive to respondents at different levels of the latent dimension. The position of an item is called “difficulty”, defined as the position on the latent scale where there is a 50 % probability of endorsing a higher response option over a lower (figure 6).

Figure 6



*In a parametric IRT, the IRF-curve usually takes an logarithmic "s-form". Item difficulty is the point on the latent scale  $\theta$  where a person has a 50% chance of responding positively to the item. At this position the measurement capacity of the item is best. With increasing distance from this position, the measurement capacity of the item decreases.*

In the IRT-models the position of items on the latent dimension usually are called “difficulty”. The measurement process will also yield a position for the *respondents* on the latent variable, usually called “ability”.

There are different parametric IRT-models of increasing complexity. The most common are the one-, two- and three-parameter models.

In the one-parameter model (1-PL) the probability of endorsing a higher response option is only dependent on the positions of persons and items on the latent dimension. This means that the 1-PL

makes the assumption that all items in the rating scale have the same capacity to discriminate between the respondents. This is manifested graphically in that the IRF-curves of items become parallel and do not intersect. Consequently all items in the mathematical formula for the 1-PL have the same weight.<sup>ii</sup> The desired property of Invariant Item Ordering (IIO) is built into the 1-PL by the assumption of equal discrimination. This makes the 1-PL useful for comparisons between different populations and scores on different item sets.

The 1-PL yields a measure of the proportion of information in the data that are explained by the measures (“variance explained by measures”). This is an indication of the precision of the measurement.

The assumption of unidimensionality can be tested by a principal component analysis of the variance that is unexplained by the model (the “residuals”). If the assumption of unidimensionality is correct, the variance of the residuals should be random. A finding of significant systematic components in the residuals is a sign of noticeable influence of secondary latent variables in the measurement.

The 1-PL can produce a measure called *person reliability*, indicating the overall capacity of the rating scale to reliably separate persons at different levels on the latent variable.

If rating scale items have more than two response options, each step of the response options will get an IRF-curve of its own. In the 1-PL the fit of the response options to the data can be evaluated. Ideally each step of the response options should be connected to a meaningful increase in the underlying latent dimension. If the steps are reversed or very narrow they will distort measurement or contain very little information.

The Rasch model is a special variant of the 1-PL. Most IRT-models try to find a model that fits the data. If a dataset does not fit, usually a more complex model is tested (for example a two parameter model, see below). The Rasch approach to measurement is the other way around: data should fit the model. If the actual data do not match the expectations from the model, the Rasch approach would be to search for the reason for the misfit, which may lead to modifications of the item set. Misfit may also be caused by respondents with misfitting response patterns, which could be explained by

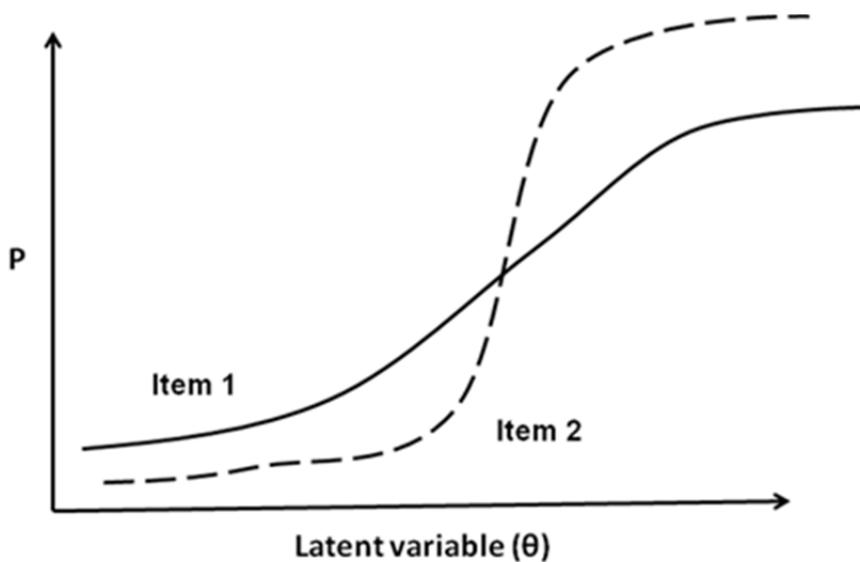
---

ii) In the one parameter model the probability of endorsing a specific category of an item can be written as  $\text{Prob}(\text{Pat}_n \text{ responds in cat. } j \text{ on item } i) = F[\sum(\theta_n - (\delta_i + \tau_j))]$  where  $\theta_n$  is the estimated patient severity,  $\delta_i$  is the item location,  $i= 1,2,\dots,I$  (the number of items) and  $\tau_j$  are the category thresholds,  $1,2,\dots,j$ . The summation is from category 1 up to  $j$ .  $F$  is the transformation function.  $\sum\tau_j = 0$  for  $j= 1,\dots,m$ ; in an  $m+1$  category item.

differences in respondent properties, for example if some respondents had a different psychiatric diagnosis than the one intended for measurement. In the Rasch perspective a rating scale should be tested and modified until the scale produces measurement fitting the model, which is considered as evidence of both the validity and the reliability of the rating scale.<sup>57</sup>

While the 1-PL assumes equal discrimination to all items, the two parameter models (2-PL) allows each item to have a discrimination of its own. Discrimination is included as a new parameter in the model, called “parameter a”. The IRF-curves may thereby have different slopes and intersect (figure 7).

**Figure 7**



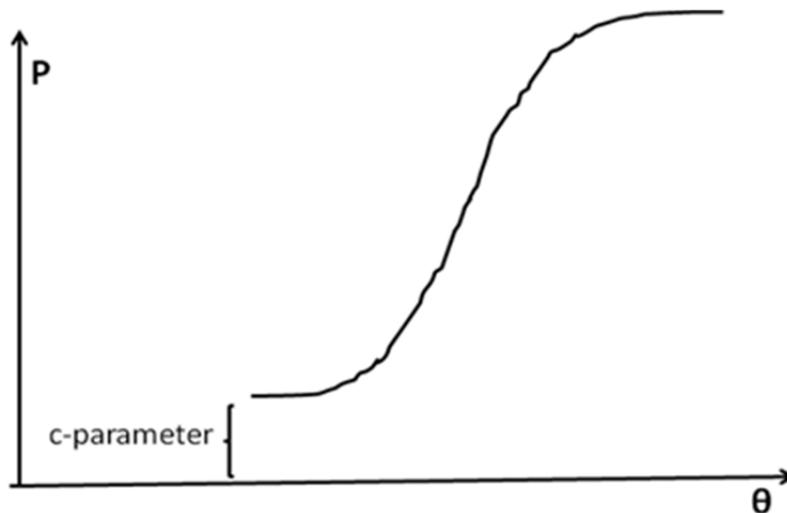
*The two items in the figure have different slopes. The item with the steep slope (item 2) performs better in differentiating individuals on the latent variable than item 1. The slope of the curve at 50 % probability of endorsement is called discrimination. Discrimination is included as a second “parameter a” in two and three parameter models of IRT.*

Items with high discrimination, having an IRF curve with a steep slope, yield higher differences in P between respondents at different levels on the latent variable than items with low discrimination. The 2-PL is suitable for situations when data does not fit the 1-PL. It can reveal items with low information content (low discrimination) and investigate if the property of IIO exists in the actual data.

In the 1- and 2-PL, the IRF-curve approaches zero as the level of the latent variable diminishes. The three parameter models (3-PL) are adapted to situations where a zero probability of a higher response alternative is unlikely. An example of this is educational applications of IRT with rating scales containing fixed response options, where guessing will yield a certain probability of getting the right answer without having a higher grade of knowledge (the latent variable). An example is an item containing four response options reflecting knowledge in a subject. Even without any

knowledge on the subject there would be a 25% chance of picking the right response option just by guessing. In such cases the IRF-curve does not approach zero at its lower end. To compensate for guessing the mathematical formula of the 3-PL adds a third parameter to the model, often called “parameter c” (figure 8).

**Figure 8**



*The three parameter model (3-PL) compensates for guessing by adding a third parameter. The lower end of the IRF-curve will approach a higher value than zero. The “guessing” parameter is usually called the c-parameter.*

#### **4.2.5 Calculation of model parameters**

In CTT the estimation of the position of a respondent on the latent dimension usually is done by a simple addition of scores on the individual items (“sum score method”). In IRT the estimation is done by a complicated mathematical search process called “the Maximum Likelihood Method”. The method calculates the level on the latent dimension for which different response patterns to the items in the scale are most probable. By repeating the calculations for different response patterns, the parameters in the models can be determined.<sup>49</sup>

The calculation process can also evaluate the consistency of the response patterns, yielding the estimates of model fit.

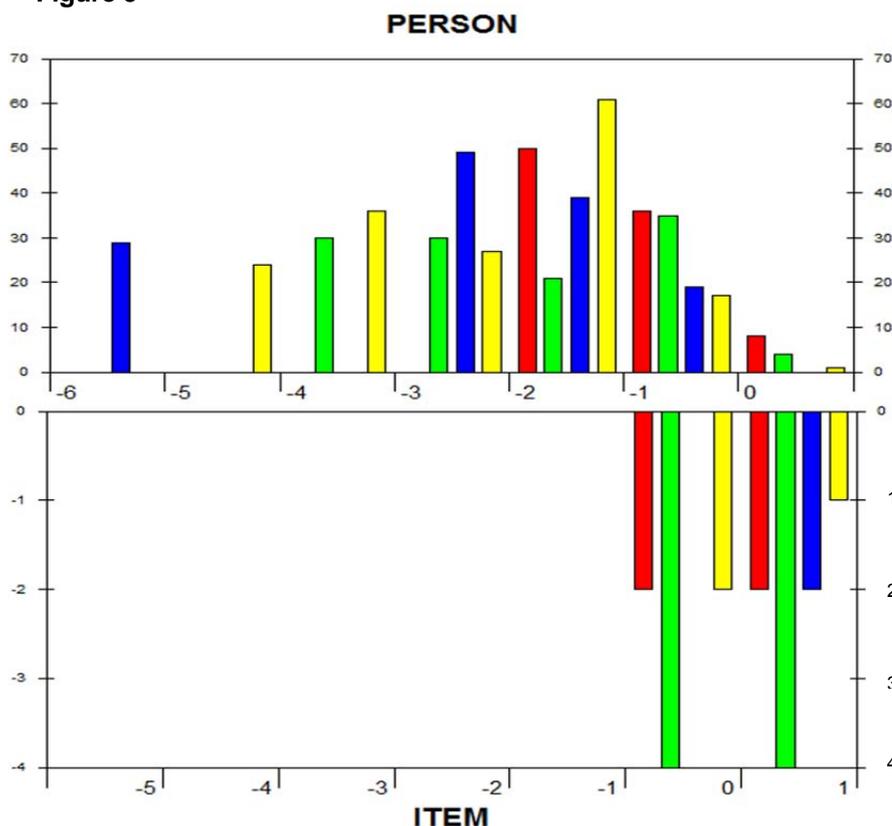
#### **4.2.6 Special features of IRT**

First, parametric IRT allows users to create an interval scale of measurement, reported in units called logits. Since logits are interval type data, parametric statistics are suitable for further analysis.

Second, in classical test theory, there is one reliability estimate for the whole test. In reality the reliability varies for different degrees of a latent variable. The reason for this is that items have their maximum capacity to discriminate between persons only at the point on the IRF-curve with the steepest slope. Persons far away from that point will be measured with lower discrimination. If persons with varying degrees of the latent variable are to be measured with equal discrimination, item difficulties must be spread out over the latent variable to cover all levels of severity in the sample. It is however very common that item difficulties are concentrated in a small part of the severity spectrum.

Third, in the 1-PL the optimal measurement level of items and the severity of respondents to a rating scale will get a measure on the same dimension and can be compared. This can graphically be displayed in a *person-item barchart* (figure 9). In this example there is a lack of items corresponding to lower levels of depression severity.

**Figure 9**

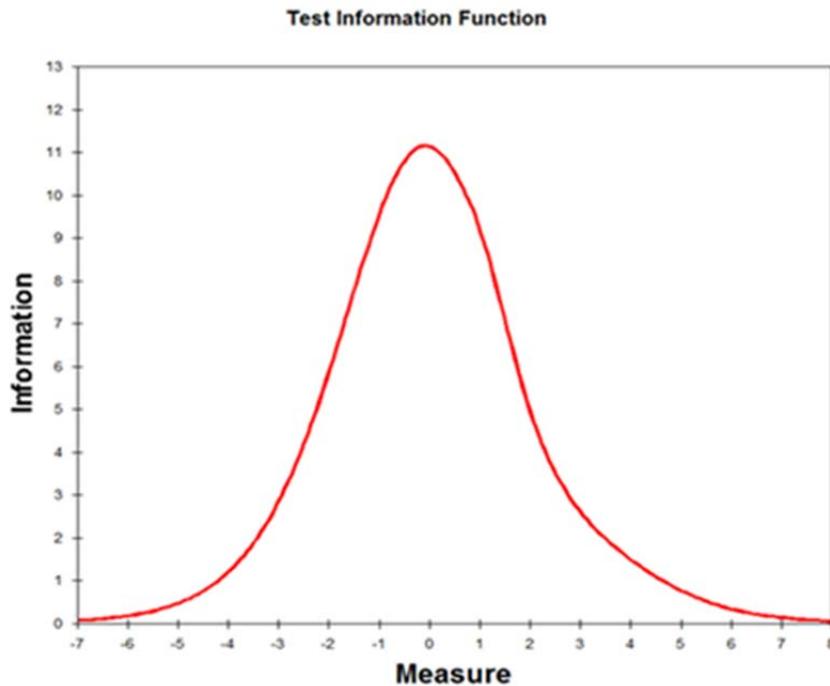


The figure reveals a mismatch between the optimal measurement level of the items and the severity of respondents (persons) on the latent variable. On the x-axis is the latent variable, increasing from left to right. The height of the bar-charts above the x-axis represents the number of persons at different levels of severity on the variable. The height of the bar-charts below the x-axis represent the numbers of items with maximum measurement capability at different levels of severity.

If there is a mismatch between persons and items, like in figure 9, the reliability of measurement at different levels of severity will vary. In the 1-PL this can be described in a graph called the Test

Information Function (TIF). The TIF is a curve describing how the information from a rating scale is distributed over different levels of severity on the latent variable (figure 10).

**Figure 10**



*The Test Information Function (TIF) is a graph describing how the relative amount of information varies over the severity spectrum.*

The measurement properties of a rating scale are optimized if items are distributed to cover the full range of the respondents' measures on the dimension. Ideally TIF should rise quickly at the beginning of the person measure range and keep approximately the same height towards the end of the range. If, however, the difficulties of items in a rating scale lie close together at a small range of severity compared to the spread of person severity, like in figure 9, the TIF-curve narrows, as in figure 10. The consequence of a narrow TIF is *low precision of ratings* where the TIF-curve is low. Also, the *sensitivity to change* falls with decreasing TIF, giving changes in the latent variable deteriorating pay-off in raw scores on the rating scale. Changes in regions with high TIF affect scores on many items while equal changes in parts of the severity spectrum with low TIF affect few items and, therefore, result in much smaller reductions of scores.<sup>58-59</sup>

Fourth, IRT allows the calculation of Differential item functioning (DIF). DIF occurs when respondents of rating scales from different groups have different probabilities of endorsing a response option in the rating scale, after matching on the underlying level on the variable that the item is intended to measure.<sup>60</sup> DIF can be used to check potential test bias between for example

gender, social groups, diagnostic subgroups or different translations of a rating scale. Studies show that differences in scores on depression and other rating scales may be caused by DIF between culturally or diagnostically dissimilar groups.<sup>61-62</sup> Such findings are indications of test bias, which is a threat to the “test consequence” aspect of validity.

Fifth, IRT allows comparisons between different samples and different tests, when the data fulfills the assumption of Invariant Item Ordering (IIO). A rating scale with IIO is a prerequisite for computerized adaptive testing, where items with difficulties matching the ability of the respondent are selected from a large item pool where the difficulty of each item has been properly established.

#### 4.2.7 IRT-measures

IRF-curves are useful for the understanding of basic principles of IRT and to get a general view of the properties of the items in a rating scale. For precise evaluations the IRT computer programs yield numeric measures describing different aspects of a rating scale that are useful for the evaluation of the reliability and validity of a rating scale (table 3).

**Table 3. Explanation of IRT-measures**

IRT-measures	
Scalability coefficient H	Capacity of the full scale to rank persons according to their total score on the latent dimension (in Mokken analysis).
Item scalability coefficient	Capacity of an item to contribute to the total score for ranking of persons on the latent dimension (in Mokken analysis).
Monotonicity	Probability of endorsing an item is non-decreasing with increasing degree of the latent variable.
Invariant Item Ordering (IIO)	Ordering of item locations on the latent variable are the same for respondents at all levels of the latent dimension.
Automated item selection (AIS)	A process within the Mokken methods that brings items with high covariances together. <sup>52</sup>
Item position (difficulty)	The modeled measure of an item on the latent dimension ( <i>parameter b</i> in all parametric IRT-models).
Person position (ability)	The modeled measure of a person on the latent dimension.
Discrimination	The capacity of an item to separate persons on the latent dimension ( <i>parameter a</i> in the 2- and 3-PL).
Guessing parameter	A parameter that compensates for the chance of a correct answer just by guessing ( <i>parameter c</i> in the 3-PL).

Fit statistics	Evaluates how well the model represents the data.
Infit	Weighted fit value, mostly sensitive to respondents close to the position of the item on the latent dimension (inlier sensitive).
Outfit	Unweighted fit value, equally sensitive to all respondents on the latent dimension (outlier sensitive).
Variance explained by measures	The proportion of the variance in the data that can be explained by the Rasch model.
Dimensionality	Analyzes if the unexplained variance in the data is systematic, which would indicate multidimensionality. Can be analyzed by the AIS-procedure in the Mokken model and by a principal component analysis of residuals in the 1-PL.
Person reliability	This is a measure in the 1-PL of the capacity of the scale to separate between persons on the latent dimension.
Item reliability	Indicates the relevance of the item for measurement and if the sample is big enough to precisely locate the items on the latent dimension.
Test Information Function (TIF)	TIF is a curve describing how information from a rating scale is distributed over different levels on the latent variable.
Differential Item Functioning (DIF)	Evaluates if items work or are perceived differently in subgroups.
Category outfit	Fit statistic evaluating how well response categories in the items conform to the Rasch model.
Monotonically increasing response categories	Evaluates if respondents at higher levels on the latent dimension prefer higher response categories.
Advancing average measures	Differences on the latent dimension between measures where one response alternative is preferred above the previous.

#### 4.2.8 Advantages of IRT-methods

IRT-methods are considered as superior to CTT-methods by proponents because they avoid the problems arising from the use of methods designed for analysis of interval data for the ordinal data produced by rating scales. From a strict statistical view a transformation of ordinal data to interval data is a prerequisite for the use of parametric statistics, like ANOVA, for rating scale data.

Scalability H is a non-parametric alternative as a reliability estimate, considered more robust than Cronbach's alpha.<sup>43, 50</sup> In the Rasch model the concept of person reliability serves a similar purpose. The Mokken method provides methods for evaluation of the dimensionality, monotonicity and IIO of a rating scale.

The scalability H and the fit statistics in parametric IRT are tests of the internal structure aspect of validity of a rating scale, concerning how well the data conforms to expectations of the models.

Dimensionality testing in the Mokken model and 1-PL can reveal influence from secondary latent variables, without the shortcomings of traditional factor analysis. Some aspects of IRT, like the Test Information Function and Differential Item functioning, have no equivalent in CTT.

#### **4.2.9 Shortcomings of IRT methods**

IRT relies on complex models and mathematics that can be difficult to understand for non-statisticians. IRT is also unfamiliar to many editors and reviewers of scientific journals, making it difficult to get results based on IRT published. IRT has so far mostly been used for cognitive testing. Clinical constructs like depression may have different properties, making application of IRT-models more complicated. It may be much more difficult to construct items with good measurement capabilities at extreme positions for psychopathological variables, like mania and depression, than for cognitive variables in ability tests.<sup>63</sup> Also, not everyone agrees that IRT-methods have yet proven their superiority over CTT-methods.<sup>44</sup> Usually IRT-model evaluations demands large samples with the intention to find a representative model and to determine parameters for use when evaluating further individuals (patients, examinees etc.). IRT-methods have however also been tested as a tool for the evaluation of smaller samples.<sup>64</sup>

## **5 THE STUDIES**

The first two studies concerned the construction and evaluation of the new rating scale AS-18. CTT-methods were used in the first study and IRT-methods in the second study. The two studies analyze two different patient samples recruited from the Affective Disorder Clinic at Psychiatry Southwest, Karolinska University Hospital Huddinge. In the third study the properties of the depression subscale of AS-18 and PHQ9 were analyzed and compared to the MADRS using IRT-methods. In the fourth study the properties of the HDRS were analyzed by IRT-methods in five randomized controlled studies and some of the consequences of the shortcomings of HDRS were discussed.

### **5.1 STUDY ONE: DEVELOPMENT AND VALIDATION OF THE AFFECTIVE SELF RATING SCALE**

In this study the development of the new self rating scale, the Affective Self Rating Scale (AS-18), was described and its performance evaluated. The paper describes how items were developed with a starting point in the DSM-IV symptom criteria for depressive and manic/hypomanic episodes. Since the aim of the new scale was to differentiate the dimensions of mania and depression, all symptoms that are shared in the definitions of both kinds of episodes were excluded. In its final form the rating scale consisted of two subscales with 9 items for depression (AS-18-D) and 9 items for mania/hypomania (AS-18-M) (see appendix).

The scale was evaluated on a clinical sample of 61 patients with ages in the range 17 to 76 years (average 44). Clinical diagnoses were Bipolar I (N=37), Bipolar II (N=8), Bipolar Not Otherwise Specified (N=8) and Major Depressive Disorder (N=8). The composition of the sample was fairly representative of the composition of diagnoses at the clinic. For evaluation, methods of CTT were used.

#### **5.1.1 Results**

Both subscales in the AS-18 showed good internal consistency, as estimated by Cronbachs alpha (table 4). The subscales showed a high and significant correlation to the predicted corresponding reference scale (convergent evidence), while the correlation to the scale measuring the opposite affective pole was low and insignificant (discriminant evidence). The correlation was also high to the clinicians overall assessment of mania and depression severity, as rated by the Clinical Global Impression scale, modified for bipolar patients (CGI-BP), with special global ratings for mania (CGI-BP-M) and depression (CGI-BP-D) symptoms (table 4).<sup>65</sup>

**Table 4. Cronbachs alpha and correlation to reference scales**

	<b>Cronbachs Alpha</b>	<b>MADRS correlation</b>	<b>HIGH-C correlation</b>	<b>CGI-BP-D correlation</b>	<b>CGI-BP-M correlation</b>
<b>AS-18-D</b>	0.89	0.74 <sup>a</sup>	0.15 <sup>b</sup>	0.68 <sup>a</sup>	-0.01 <sup>b</sup>
<b>AS-18-M</b>	0.91	0.25 <sup>b</sup>	0.80 <sup>a</sup>	0.10 <sup>b</sup>	0.73 <sup>a</sup>

*Internal consistency and correlations between depression and mania subscore of the AS-18 and the reference scales. (Spearman's rho and two-tailed significance. <sup>a</sup> =  $p < 0.01$ , <sup>b</sup> = non-significant.  $N = 61$ ).*

In a factor analysis four factors emerged (table 5). Seven out of nine mania items converged on the first factor. Items for irritability and risk-taking did however have substantial loadings on the other factors. The depression subscale was subdivided into the second and third factor. The second factor consisted of the items for hopelessness, depression, anhedonia, guilt and suicidal ideation. The items for motor retardation, low energy and slow thinking emerged as a third factor. The item for increased sleep had its highest load on a fourth factor, where irritability and risk-taking also showed substantial loadings.

**Table 5. Factor analysis of AS-18**

	<b>Components and percentage of variation explained</b>			
	<b>1 (28.6%)</b>	<b>2 (22.6%)</b>	<b>3 (14.9%)</b>	<b>4 (7.9%)</b>
1. Talkativeness	0.86	-0.05	0.12	0.14
2. Increased sleep	0.18	0.14	0.44	-0.76
3. Less need for sleep	0.69	-0.12	0.22	0.16
4. Hopelessness	0.13	0.90	0.20	0.16
5. Retardation	0.08	0.20	0.79	-0.10
6. Overactive	0.86	0.16	-0.26	-0.06
7. Agitation	0.78	0.34	-0.31	-0.09
8. Racing thoughts	0.80	0.31	-0.05	-0.02
9. Irritability	0.54	0.34	0.07	0.52
10. Depression	0.10	0.92	0.15	0.00
11. Anhedonia	0.27	0.74	0.22	-0.31
12. Low energy	-0.15	0.29	0.83	-0.06
13. Guilt	0.01	0.87	0.20	0.21
14. Slow thinking	-0.03	0.39	0.74	0.06
15. Increased self-esteem	0.84	0.08	0.11	-0.14
16. Euphoria	0.78	0.02	-0.04	0.09
17. Suicidal ideation	0.04	0.64	0.33	-0.13
18. Risk-taking	0.43	0.15	0.34	0.53

*Factor loadings for each item, together explaining 74.1% of the variation. Principal component analysis with Varimax rotation with Kaiser Normalization. Rotation converged in seven iterations. Eigenvalue >1.*

Receiver operating characteristic (ROC), or ROC-curves, is a method of analyzing suitable cut-offs for a measure, using a dichotomously classified measure as reference. The analysis suggested that scores above of 9 on the subscales indicate a depressive or manic/hypomanic state. For mixed states, a simultaneous score of 9 or more on both subscales yielded a high sensitivity (0.90), but a lower specificity (0.71). Using higher combined cut-off levels of 19 for the depression subscale and 13 for the mania subscale gave a sensitivity of 0.5 and a specificity of 0.95.

### **5.1.2 Discussion.**

Within the CTT-paradigm, the AS-18 showed excellent measures of internal consistency and positive validation data concerning associations with reference scales. The factor analysis largely confirmed the predicted factor structure, with some exceptions. Items for irritability, risk-taking and increased sleep did not behave as predicted and therefore might be considered less suitable in this format. In the DSM-IV criteria for mania, irritability is considered less specific than the classic manic symptoms of euphoria and expansivity. Our study supports this opinion.

The two separate depression factors can be interpreted as reflecting two clinical dimensions in depression: one consisting of emotional and cognitive aspects and one consisting of motor and energy aspects. The finding of the factor consisting of motor and energy aspects suggests that motor retardation can be self-rated. The scales' ability to detect mixed states was problematic. Two thresholds had to be established, the first with the aim of high sensitivity, the other with the aim of high specificity.

In conclusion, the study showed that AS-18 can be used as a time efficient aid for clinicians in outpatient settings to identify patients with different affective states and to rate symptom severity.

## **5.2 STUDY TWO: AN IRT VALIDATION OF THE AFFECTIVE SELF RATING SCALE**

The aim of the second study was to further evaluate AS-18 using IRT-methods and to analyze the potential for improvement of the scale. A new sample of 231 patients with clinical Bipolar I diagnoses was recruited at ordinary appointments at the Affective Disorder Outpatient Clinic at Psychiatry Southwest. 59.9% were female and 40.1% were male. The average age was 48.0 (range 18-87). The clinical diagnoses were reassessed by a structured interview. 24 patients were reclassified as either Bipolar II (N=13) or Schizoaffective (bipolar type) (N=11). Most patients rated low symptom levels, as could be expected at routine follow up visits. 96% of the patients were treated with medications, mostly lithium (78%). A majority (59%) used more than one mood stabilizing medication.

Mokken analysis was used as a first basic step in the study. Since the data conformed to the Rasch model of parametric IRT, this model was considered appropriate for further analysis.

### 5.2.1 Results

In the Mokken analysis, both subscales of the AS-18 showed a strong ability to rank respondents according to their total score on both subscales of AS-18 and all items contributed adequately to the measurement. This was demonstrated by item scalability coefficients above the generally accepted cut-offs of  $\geq 0.3$  for adequate items and  $\geq 0.5$  for a strong full scale.<sup>51</sup> Excluding extreme cases with scores of zero on the subscales gave somewhat lower scalability coefficients (table 6).

**Table 6. Mokken analysis**

	Coeff. H N=200	Coeff. H N=133		Coeff H N=205	Coeff H N=151
AS-18-M	.56	.49	AS-18-D	.65	.56
1. Talkativeness	.54	.47	2. Increased sleep	.46	.34
3. Less need for sleep	.51	.43	4. Hopelessness	.70	.63
6. Overactive	.68	.63	5. Retardation	.61	.53
7. Agitation.	.57	.49	10. Depression	.74	.67
8. Racing thoughts	.64	.57	11. Anhedonia	.71	.64
9. Irritability	.57	.45	12. Low energy	.72	.63
15. Increased self-esteem	.50	.44	13. Guilt	.65	.57
16. Euphoria	.56	.51	14. Slow thinking	.61	.53
18. Risk-taking	.43	.36	17. Suicidal ideation	.51	.45

*The table shows coefficient H for AS-18-M and AS-18-D and scalability coefficients for items. First column includes all cases with complete ratings. Second column includes only cases with ratings above zero.*

In the Rasch model analysis the estimate of discrimination for item 2 (increased sleep) was low. The response categories in the AS-18 worked well in most aspects. The advance in the average threshold for preferring the category step 1 over step 2 was small, however. In the analysis of dimensionality of the residuals, there were few signs of disturbing secondary dimensions. The person reliability index for the depression subscale indicated a capacity to separate the sample in two or three levels (table 7). The person reliability index for the mania subscale suggested a capacity for separation of the sample into one or two levels.<sup>66</sup>

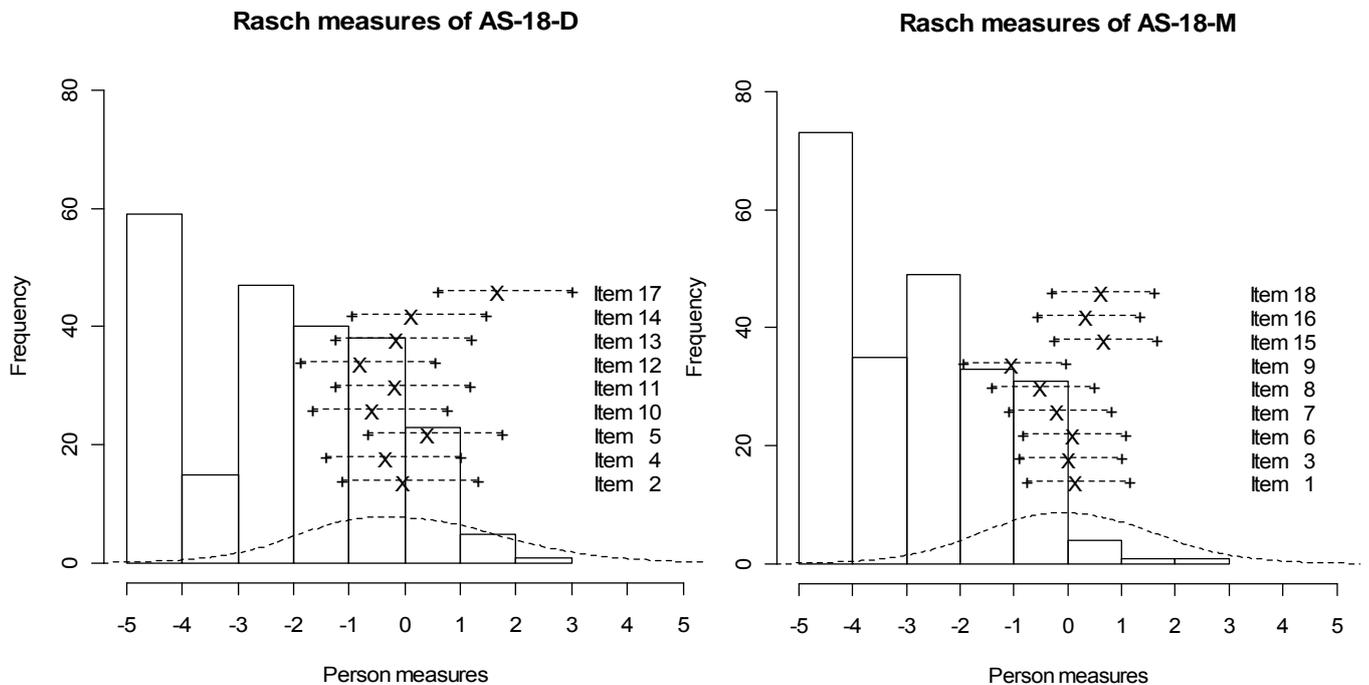
**Table 7. Person reliability**

	Person reliability
AS-18-M	.66
AS-18-D	.80

Person reliability of AS-18-M and AS-18-D (non-extreme cases). A person reliability  $\geq 0.9$  indicates a capacity to reliably separate the sample in three or four levels, 0.8-0.9 in two or three levels and 0.5-0.8 in one or two levels.<sup>66</sup>

There was a mismatch between optimal measurement levels of the items in both subscales and the symptom severity of patients. Consequently TIF was substantially lower in less severe depression and hypomania (figure 11 and 12).

**Figure 11 and 12. Coverage and TIF**



The histograms represent the number of patients at different levels of depression and mania symptom severity as estimated by the Rasch model. Horizontal lines show the range of coverage for the different items. The superimposed curve shows the 'Test information function' (TIF), a description of the relative amount of information given by the rating scale at different levels of severity.

## 5.2.2 Discussion

The IRT-models used in this study set more stringent requirements than CTT methods concerning reliability and the internal structure aspects of validity. The AS-18 subscales show strong signs of reliability in the Mokken analysis. The high scalability in addition to the dimensionality analysis

in the Rasch model indicate unidimensionality. The analysis of category function does however imply redundancy of some response categories.

The low discrimination of item two indicates that the item contributes little to the measurement and could therefore be considered redundant. This is corroborated by the factor analysis of the previous study, where item 2 ended up in a factor separated from the other depression items.

A substantial problem in the AS-18 is the limited person reliability. The explanation for this problem is the limited range of high TIF, not matching the level of symptom severity of a large proportion of the patients in the sample. The performance of AS-18 would improve if items optimal for measurement of low levels of affective symptoms were added to the scale.

In conclusion this study supports the previous study findings of reliability and validity of the AS-18.

### **5.3 STUDY THREE: AN IRT EVALUATION OF THREE DEPRESSION RATING SCALES**

In the third study IRT-methods were used to evaluate the measurement properties of AS-18-D, PHQ9 and MADRS in the same sample as study one.

A ‘3- step IRT strategy’ was used. In a first step, the rating scales were analyzed using the Mokken non-parametric approach.<sup>50</sup> In a second step, a one parameter IRT-model (1-PL) according to the Rasch definition was used.<sup>41</sup> In a third step a two parameter model (2-PL) was used and items were allowed to have different discrimination.

In a limited sample of 61 patients there might be large random variation. We therefore needed reliable confidence intervals. Conventional methods for establishing confidence intervals require distributional assumptions which are not available in this case. Bootstrapping is a group of methods for establishing confidence intervals without such assumptions and was therefore considered appropriate for this study. Bootstrapping has also been shown to work for small samples.<sup>67</sup>

#### **5.3.1 Results**

In a first step, the Mokken non-parametric analysis showed that PHQ9 and AS-18-D had strong overall scalabilities, while the scalability of MADRS was weak (table 8). MADRS item 4 was shown to degrade measurement. When item 4 was excluded, the overall measurement properties of MADRS improved and the performance of other items that previously had shown doubtful scalabilities became acceptable (MADRS item 7 and 10).

**Table 8. Scalabilities in the Mokken nonparametric analyses**

Item	scalab.	Item	scalab.	Item	scalab.	scalab*.
AS-18-D1	0.326	PHQ9:1	0.584	MADRS1	0.469	0.514
AS-18-D2	0.590	PHQ9:2	0.603	MADRS2	0.442	0.476
AS-18-D3	0.460	PHQ9:3	0.481	MADRS3	0.419	0.449
AS-18-D4	0.576	PHQ9:4	0.478	MADRS4	0.103	-
AS-18-D5	0.557	PHQ9:5	0.385	MADRS5	0.338	0.359
AS-18-D6	0.491	PHQ9:6	0.588	MADRS6	0.358	0.370
AS-18-D7	0.543	PHQ9:7	0.567	MADRS7	0.281	0.349
AS-18-D8	0.513	PHQ9:8	0.400	MADRS8	0.388	0.456
AS-18-D9	0.582	PHQ9:9	0.468	MADRS9	0.371	0.410
				MADRS10	0.257	0.332
Total H	0.513		0.510		0.339	0.415
no. of obs	57		59		56	56
Boot- strapping C.I.	[0.41, 0.63]		[0.42, 0.61]		[0.25, 0.43]	[0.31, 0.51]

\* Scalabilities calculated with MADRS4 deleted

The AS-18-D showed no violations against monotonicity and IIO and the PHQ only showed one minor violation against IIO. The MADRS however showed a number of violations against monotonicity and IIO.

In a second step, a Rasch model analysis indicated large differences concerning the item discriminating capacity and was therefore considered not suitable for the data. In a third step, applying a more flexible two parameter IRT-model, all three instruments showed large differences in item information. Several items only contributed to 5% or less in the ratings (table 9).

**Table 9. Approximate relative item information and estimated discrimination**

Item	Rel. info %	Discr.	S.E(Discr.)
AS-18-D1	<5	0.574	0.188
AS-18-D2	30	2.845	0.723
AS-18-D3	5	0.923	0.220
AS-18-D4	10	1.256	0.908
AS-18-D5	15	1.751	0.274
AS-18-D6	5	0.861	0.189
AS-18-D7	10	1.313	0.331
AS-18-D8	10	1.102	0.253
AS-18-D9	15	1.710	0.416
PHQ9:1	>25	2.425	0.390
PHQ9:2	10	1.266	0.593
PHQ9:3	5	0.918	0.183
PHQ9:4	10	1.209	0.216
PHQ9:5	5	0.704	0.209
PHQ9:6	15	1.484	0.426
PHQ9:7	10	1.245	0.269
PHQ9:8	5	0.891	0.173
PHQ9:9	10	1.150	0.278
MADRS1	25	1.927	0.719
MADRS2	15	1.425	0.481
MADRS3	15	1.175	0.239
MADRS4	<<5	0.273	0.047
MADRS5	5	0.712	0.209
MADRS6	5	0.659	0.180
MADRS7	5	0.783	0.134
MADRS8	10	1.065	0.155
MADRS9	10	0.898	0.259
MADRS10	5	0.683	0.233

The study also revealed a mismatch between the working range of items and severity of depression in the patients (poor coverage).

### 5.3.2 Conclusions

The study suggests that the PHQ9 and AS-18-D can be useful for measurement of depression severity in an outpatient clinic for affective disorder, while the usefulness of MADRS for this type of patients must be questioned and further analyzed. The study also indicates a need for improvement of depression assessment instruments, in particular in concern to reliably differentiating levels of depression, especially mild to moderate. Also, this study indicates great differences in discrimination between items. In development of new rating scales, item specific weights and the issue of coverage should be addressed.

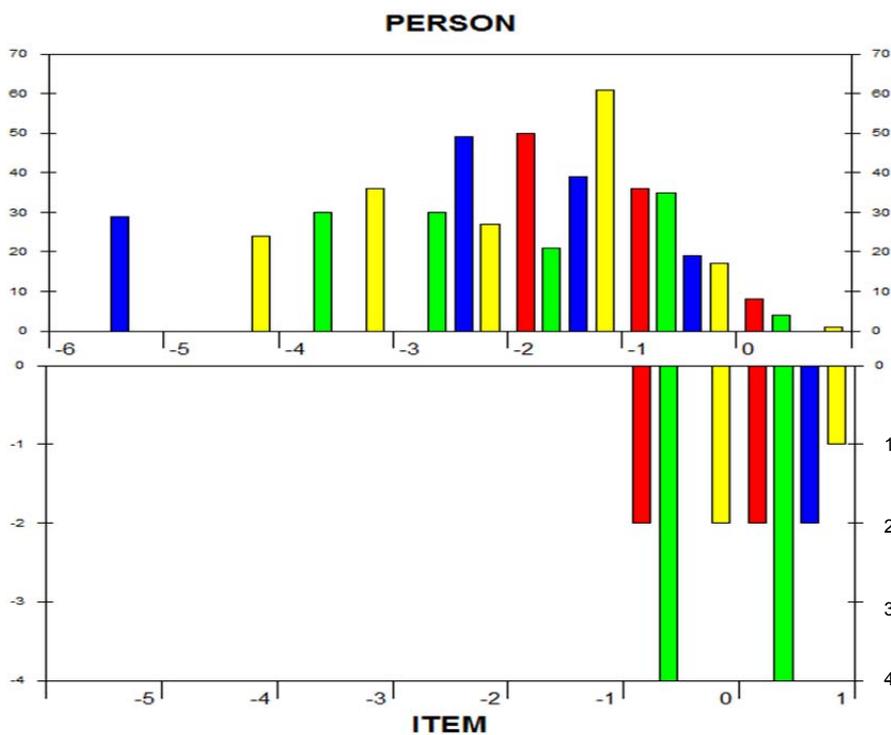
## 5.4 STUDY FOUR: RANDOMIZED CLINICAL TRIALS UNDERESTIMATE THE EFFICACY OF ANTIDEPRESSANTS IN LESS SEVERE DEPRESSION

The fourth study examined how shortcomings of measurement might bias conclusions of clinical trials of antidepressants (RCT-ADs). A recent meta-analysis of six RCT-ADs concluded that the efficacy of antidepressants was “nonexistent to negligible” in mild and moderate depression.<sup>68</sup> The aim of this study was to reanalyze the same data in order to investigate if the meta-analysis could be biased by shortcomings of the rating scale used in the RCT-ADs, the 17-item Hamilton Depression Rating Scale (HDRS). The study was performed on the primary data from five of the six meta-analyzed RCT-ADs. The authors of the sixth study were not willing to share their data. 516 individuals had ratings at endpoint and were included in the analysis. The dataset conformed well to the Rasch model of IRT, and this model was considered appropriate for further analysis.

### 5.4.1 Results

The analysis demonstrated a lack of items corresponding to lower levels of depression in the combined sample (figure 13).

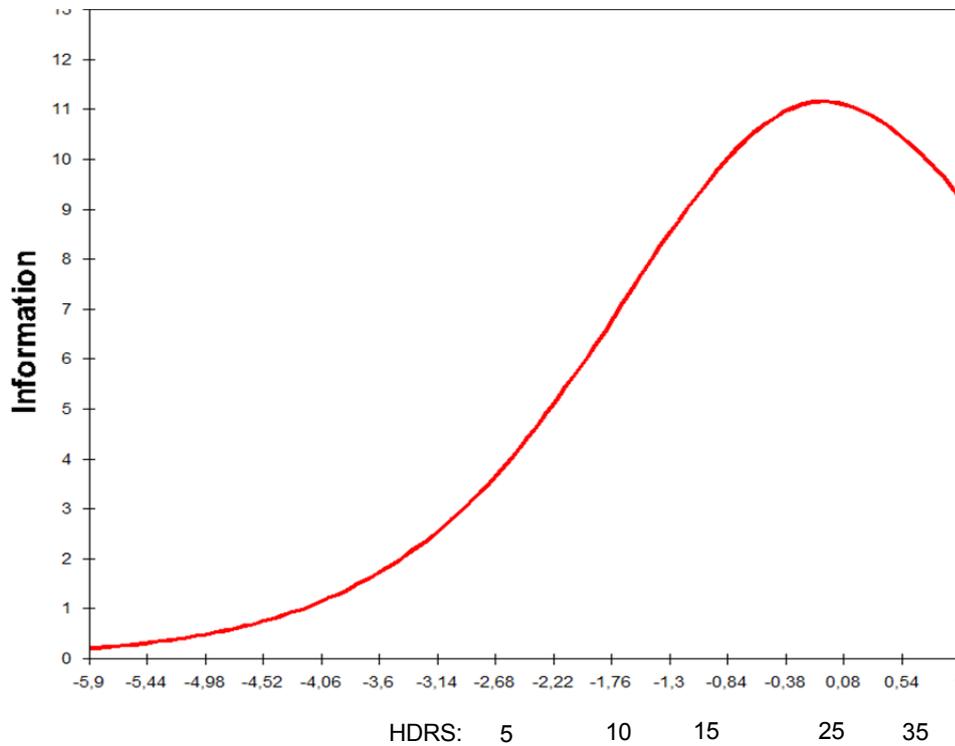
**Figure 13. Person-item match in the combined sample at end-point**



The x-axis represents depression severity with increasing levels from left to right. The figures on the x-axis show the Rasch measure of depression severity (the transformed HDRS score). The height of the bar charts above the x-axis represents the number of persons at different levels of severity. The height of the bar charts below the x-axis represents the number of items with maximum measurement capability at different levels of severity.

This mismatch between depression severity in individuals and measurement capacity of items resulted in a rapidly decreasing TIF at lower degrees of depression severity (figure 14).

**Figure 14. Test Information Function of HDRS at endpoint**



The Test Information Function (TIF) graph describes the relative amount of information that the HDRS extracts at different levels of depression severity in the combined sample at end-point. The x-axis represents depression severity increasing from left to right. The figures on the x-axis are the Rasch modeled measure of severity and below them the corresponding HDRS-score in this sample.

The same pattern emerged in the individual studies analyzed, where large proportions of the samples were measured with low information content (table 10).

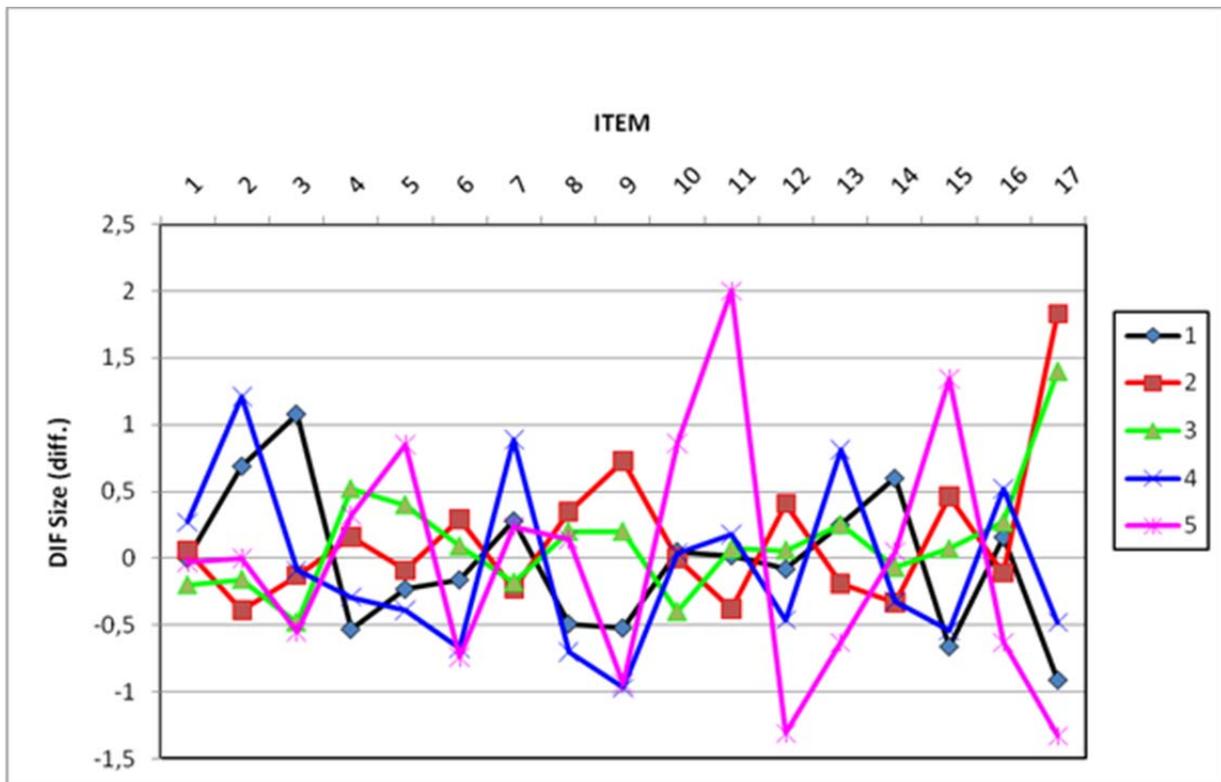
**Table 10. Proportions of samples measured below 50 % information content**

Study	Proportion (%)	N
Combined sample HDRS <sup>68</sup>	38	516
Philipp et al. <sup>69</sup>	23	157
De Rubeis et al. <sup>70</sup>	21	180
Elkin et al. <sup>71</sup>	38	89
Wichers et al. <sup>72</sup>	58	29
Barret et al. <sup>73</sup>	48	61

In all five studies included in the Fournier meta-analysis, as well as in the combined sample, a large proportion of the sample was measured with low precision.

An analysis of Differential Item Functioning (DIF) revealed large and significant differences in DIF-values ( $>0.5$ ,  $z$ -value  $>2$ ) between the studies on all items but one (figure 16). This indicates that there are substantial variations in the understanding of the items between study populations.

**Figure 16: Differential Item Functioning (DIF) reported by study**



Differences in DIF relative to the overall item difficulty (=0) on the 17 items in HDRS. 1: Philipp et al., 2: de Rubeis et al., 3: Elkins et al., 4: Wicher et al., 5: Barret et al.-Differences in DIF-measure  $< 0.5$  is considered negligible.<sup>66</sup>

### 5.4.2 Discussion

The HDRS yields less information as depression severity decreases. This has two important consequences. First, rating precision decreases as the patient improves. At endpoint, a large proportion of the ratings will have low precision, increasing the risk that real differences between study groups treated with antidepressants and placebo go undetected (type II error). Second, HDRS' sensitivity to change declines when depression severity decreases. Comparisons of HDRS score reductions between study persons with different levels of depression severity at base-line will lack validity since improvement from lower levels of depression will be systematically underestimated compared to improvement from higher levels of depression. The substantial DIF-contrasts found in this study indicate important differences in how the items were understood by the study persons in the different RCT-ADs, and signals risk of bias when data from the different studies are pooled and meta-analyzed.

This study indicates that the conclusion of the Fournier et al. meta-analysis is unfounded. The clinical value of antidepressants should not be evaluated from unreliable rating scale data.

## 6 CONCLUSIONS FROM THE DOCTORAL PROJECT

1. The first aim of this doctoral project was to develop and evaluate a self rating scale for simultaneous measurement of severity in depressive, manic and mixed affective states. To meet this aim, the AS-18 was developed.

The reliability of the AS-18 was assessed in two samples of patients with mainly bipolar diagnoses recruited from the same affective disorder outpatient clinic. Measures of reliability, using both CTT- and IRT-methods, have shown values that usually are interpreted as indicating a “good scale” and a “strong scale” respectively.

Establishing validity is a process and the studies in this project have gathered evidence of different aspects of validity:

- The *test content* aspect of validity was addressed by using the current symptom list in the definitions of depression and mania in DSM-IV as starting point for item development. The content of the AS-18 was somewhat restricted compared to DSM-IV definitions, however. In order to get better discriminant validity between depression and mania, symptoms common to the symptom dimensions were excluded.
- The *internal structure* of the AS-18 was investigated by a factor analysis within the CTT-paradigm showing overall support for the expected dimensionality of the rating scale. The IRT analysis gave support to the property of unidimensionality of the subscales by good estimates of scalability H in the Mokken method and low systematic residual variance in the Rasch analysis.
- The *response processes* were tested during the item development by taking into account misunderstandings and opinions expressed by patients and clinicians. This process was not systematically recorded, however. The scalability coefficients and estimates of fit to the model are statistical tests on how well the response patterns conform to expectations. The results of these tests were favorable, indicating that respondents comprehend the items as intended.
- The subscales of the AS-18 showed *association with other variables* (interview rated scales) in the predicted way, implicating good convergent and discriminant validity.
- The test of Differential Item Functioning (DIF) of sex bias is a form of test of the *consequences of testing* aspects of validity. The DIF did not indicate such sex bias.

The studies also revealed shortcomings in the AS-18. The capacity to reliably separate different levels of severity was limited to just a few levels (person reliability in the IRT-analysis). This finding seems surprising, considering the strong overall reliability indicators and the finding that almost all items showed good indicators of performance. The explanation appeared in the analysis of item coverage and Test Information Function in the IRT-analysis, which showed that the optimal measurement range of the items was clustered at medium levels of depression and mania severity. As a result, high and low grades of the dimensions were measured with substantially lower reliability. One item in the AS-18 was found to have a low discrimination capacity and might be replaced.

There are several limitations to the evaluation of AS-18. Both samples were recruited from the same clinic and the scale was tested in (mainly) bipolar samples. An evaluation of the properties of the rating scale in other settings, and for other affective diagnoses than bipolar I, is warranted. The test-retest aspect of reliability and the sensitivity to change has not so far been examined. Several aspects of validity should be further evaluated.

2. The second aim of the project was to explore if IRT-methods are useful for the evaluation and improvement of rating scales for mania and depression. The IRT-based tests for monotonicity, Invariant Item Ordering (IIO), Test Information Functioning (TIF) and Differential Item Functioning (DIF) provided valuable information, not readily accessible within the Classic Test Theory (CTT) framework. The ‘three step approach’ used in paper 3 is a method useful in the early stages of the development of a rating scale or when a rating scale is being utilized in a new setting. Some other potential advantages of IRT over CTT, like the robustness of the overall measures of reliability or the usefulness of interval scaled measures of respondents, were not specifically investigated. The overall conclusion is that IRT-methods can be complementary to CTT-methods, providing important information about rating scale functioning that usually are overlooked, as demonstrated in paper 4. Thus, IRT-methods seem useful in development, evaluation and use of rating scales for mania and depression.

3. The third aim of the doctoral project was to compare the psychometric properties of the depression subscale of AS-18 to the PHQ9 and the MADRS. In paper 3 it was demonstrated that PHQ9 and AS-18-D showed good over-all measurement properties, while the MADRS properties were weaker. Similar problems to AS-18 concerning limited item coverage and Test Information Function were found in the PHQ9 and MADRS. In the MADRS, one item degraded the measurement properties of the scale. In all three scales several items were found to provide relatively small amounts of information. This suggests that the performance of the rating scales

could be improved. Items with poor measurement capacity could be replaced with items with better capacity and coverage of mild and severe depression.

The findings in paper 3 should be considered only as indicative, due to the limited sample size. Considering that MADRS is the most widely used rating scale in treatment studies of bipolar depression it is important to investigate the indications of poor performance further.

4. The fourth aim was to investigate if insufficient measurement properties of the Hamilton Depression Rating Scale (HDRS) might explain the low or absent efficacy of antidepressant medication, in less severe depression, as often found in Randomized Controlled Trials (RCT). Therefore, in paper 4 we analyzed a dataset from a meta-analysis of RCT of antidepressants that concluded low or absent efficacy of antidepressant medication. The analysis showed that the HDRS yields rapidly decreasing precision and sensitivity to change with diminishing depression severity. The consequence is low measurement precision at endpoint in studies. Furthermore, comparisons of score reductions between study groups starting at different levels of depression severity will be biased. Thus, declared low or absent efficacy of antidepressant in less severe depression might be explained by measurement bias. The study also revealed large differences in Differential Item Functioning (DIF) between the studies, indicating that the different study populations had perceived the HDRS items in different ways, which might have increased the risk for bias in the meta-analysis. Study four indicates that unreliable measurement can bias treatment studies. In consequence, scientific conclusions, treatment recommendations and policy decisions may be based on faulty assumptions. The study indicated that a short version of HDRS, with only 6 items, performed better than the full 17 item version.

The Affective Self Rating Scale (AS-18) is a scale measuring severity in depressive, manic, hypomanic and mixed affective states. It takes only a few minutes for respondents to use and has shown good reliability estimates and evidence of validity in samples from an affective disorder outpatient clinic. It should be useful in clinics with similar populations.

The same kind of measurement problems were found in four depression rating scales, indicating a general problem in depression measurement. Established rating scales for depression show dubious properties. Newly developed rating scales seem to perform better, but can be improved. The results indicate that unreliable measurement biases conclusions about antidepressant treatment. It is likely that rating scales used in other areas of psychiatry have similar weaknesses as HDRS and MADRS.

The finding in this project that a shortened version of HDRS performed better than the full version is in line with other studies and exposes a potential for improvement. One study calculated that RCTs of antidepressants using the six item version of HDRS would require one third less patients than

studies using the full version of the scale.<sup>74-75</sup> In a recently published study of rating scales for physical functioning, the best IRT-based instruments required only one-quarter of the sample sizes compared to the conventional rating scales.<sup>76</sup> In order to provide better treatments for depressed persons, there is an urgent need of improved rating scales and the IRT-approach is a promising tool for such an endeavor.

## 7 SAMMANFATTNING PÅ SVENSKA

**Bakgrund:** Inom psykiatrin sker utvärdering av patienter i ökad utsträckning med hjälp av skattningsskalor, i såväl klinisk praxis som forskning och kvalitetskontroll. Samtidigt är omvandlingen av de subjektiva symtomen av psykisk sjukdom till trovärdiga siffror utsatt för många felkällor. Noggrann utvärdering av skattningsskalor är därför nödvändig.

Detta doktorandprojekt uppstod ur ett kliniskt behov av en användbar självskattningsskala för affektiva symtom vid en öppenvårdsmottagning för affektiv sjukdom. Inga befintliga skattningsskalor som uppfyllde de kliniska behoven hittades i litteraturen.

**Syfte:** Syftet med doktorandprojektet var att utveckla och utvärdera en skala för mätning av svårighetsgraden av depressiva, maniska och blandade affektiva tillstånd och att undersöka om Item Response Theory (IRT) är användbar för utvärdering och förbättring av skattningsskalor för mani och depression. Ett ytterligare syfte var att undersöka om randomiserade kontrollerade studier av antidepressiva läkemedel (RCT-AD) kan ge missvisande resultat som följd av bristande mätegenskaper hos den vanligaste skattningsskalan för depression, Hamilton Depression Rating Scale (HDRS).

**Metoder:** En skala med 18 items utvecklades och gavs namnet *Affektiv självskattningsskala (AS-18)* med separata delskalor för depression och mani/hypomani (se appendix). Den utvärderades i två patientmaterial (N = 61 och N = 231) och jämfördes med Patient Health Questionnaire (PHQ9) och Montgomery Åsberg Depression Rating Scale (MADRS). Data från fem RCT-AD som ingår i en nyligen publicerad meta-analys analyserades (N = 516). Statistiska metoder från klassisk testteori (CTT) och IRT användes.

**Resultat:** AS-18 visade god av tillförlitlighet (Cronbachs alpha på 0,89 och 0,91 för depressions- respektive manidelskalorna). AS-18 visade också en stark korrelation till referensskalor. En faktoranalys bekräftade i stort sett den förväntade faktorstrukturen. Items för irritabilitet, risktagande och ökad sömn avvek dock från det förväntade. IRT- analysen visade att AS-18 och PHQ9 hade stark förmåga att rangordna patienterna enligt deras summapoäng, medan MADRS hade svaga sådana egenskaper. Flera items i skattningsskalorna bidrog med lite information till mätningen. Det var få items som täckte lägre nivåer av depression och mani, vilket gjorde mätningen av lindriga nivåer av symtom oprecisa.

I analysen av fem RCT-AD fann vi att informationsmängden som HDRS kunde samla avtog med minskande depressionsgrad. Dessutom konstaterades att de flesta items i HDRS förstods olika av de olika studiepopulationerna. Slutsatsen av meta-analysen, att antidepressiva medel har försumbar effekt i lindrig till måttlig depression, kan ha orsakats av brister i mätmetoder.

**Slutsatser:** AS-18 har visat reliabilitet och validitet i två studier. I öppenvård för patienter med affektiv sjukdom kan den användas som ett tidseffektivt stöd för att identifiera patienter med olika affektiva tillstånd och för att mäta symtomens svårighetsgrad. IRT-metoder visade sig vara användbara på flera sätt, bland annat för att analysera skattningsskalor avseende mängden information som enskilda items bidrar med till mätningen, hur precisionen i mätningen varierar över olika svårighetsgrader och för att undersöka om olika studiepopulationer uppfattar items på lika sätt. Studier av antidepressiva mediciner kan ge missvisande resultat till följd av bristfälliga mätmetoder.

## 8 ACKNOWLEDGEMENTS

First and foremost I want to thank my main supervisor, associate professor Göran Isacson. Göran recruited me to the Affective disorder clinic, encouraged my research and have been continuously supportive and collaborative in all parts of this project. I also want to extend this thank to my co-supervisor professor Jerker Hetta, who have contributed more to this project than what could be expected from a co-supervisor.

A special thanks goes to Ulf Brodin, who initially was my teacher in statistics at LIME, Karolinska Institutet. This evolved to a collaboration in two of the studies. I want to thank Ulf for introducing me to the modern methods of scale evaluation and for his everlasting patience when explaining the mysteries of statistics.

I also want to thank my collaborators in study one, Benny Liberg and Stig Andersson for their dedicated work.

I want to thank Lena Backlund, Urban Ösby and Gunnar Edman for collaborating in the research that made study two possible.

Research nurse Inger Römer and the assistant nurses Anneli Isoniemi and Margareta Norén have been invaluable supportive in the patient recruitment and the retrieval of data.

I am grateful to Dr., Ph D. Caroline Wachtler for her generous sharing of her excellent language skills, helping me preparing the manuscript. Secretary Mia Pettersson have been most helpful in the preparation of the dissertation and thesis.

There are many other people whom I am grateful to for making this doctoral project possible. Many colleagues, friends and patients have contributed to this project. Thanks to all of you.

Last, but not least, I want to thank my family; Hanna, Maja, Julia and Ella. Without your love, tolerance and support this project would not have been possible.

## 9 REFERENCES

1. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med.* Sep 2001;16(9):606-613.
2. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry.* Apr 1979;134:382-389.
3. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry.* Feb 1960;23:56-62.
4. *MINI-D IV Diagnostiska kriterier enligt DSM-IV.* Danderyd: Pilgrim Press; 1995.
5. Goodwin FK, Jamison KR. *Manic-depressive illness: bipolar disorders and recurrent depression.* New York: Oxford University Press, Inc.; 2007.
6. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry.* June 1, 2005 2005;62(6):593-602.
7. Judd LL. The clinical course of unipolar major depressive disorders. *Arch Gen Psychiatry.* Nov 1997;54(11):989-991.
8. SBU. *Behandling av depressionsjukdomar. En systematisk litteraturöversikt.* Stockholm: Statens beredning för medicinsk utvärdering;2004.
9. Kessing LV, Hansen MG, Andersen PK, Angst J. The predictive effect of episodes on the risk of recurrence in depressive and bipolar disorders - a life-long perspective. *Acta Psychiatr Scand.* May 2004;109(5):339-344.
10. Coryell W, Scheftner W, Keller M, Endicott J, Maser J, Klerman GL. The enduring psychosocial consequences of mania and depression. *Am J Psychiatry.* May 1993;150(5):720-727.
11. Tham A, Engelbrektson K, Mathe AA, Johnson L, Olsson E, Aberg-Wistedt A. Impaired neuropsychological performance in euthymic patients with recurring mood disorders. *J Clin Psychiatry.* Jan 1997;58(1):26-29.
12. Barch DM. Neuropsychological abnormalities in schizophrenia and major mood disorders: similarities and differences. *Curr Psychiatry Rep.* Aug 2009;11(4):313-319.
13. Tondo L, Isacson G, Baldessarini R. Suicidal behaviour in bipolar disorder: risk and prevention. *CNS Drugs.* 2003;17(7):491-511.

14. Welton RS. The management of suicidality: assessment and intervention. *Psychiatry (Edgmont)*. May 2007;4(5):24-34.
15. McElroy SL, Keck PE, Jr., Pope HG, Jr., Hudson JI, Faedda GL, Swann AC. Clinical and research implications of the diagnosis of dysphoric or mixed mania or hypomania. *Am J Psychiatry*. Dec 1992;149(12):1633-1644.
16. Suppes T, Mintz J, McElroy SL, et al. Mixed Hypomania in 908 Patients With Bipolar Disorder Evaluated Prospectively in the Stanley Foundation Bipolar Treatment Network: A Sex-Specific Phenomenon. *Arch Gen Psychiatry*. October 1, 2005 2005;62(10):1089-1096.
17. Benazzi F, Akiskal HS. Delineating bipolar II mixed states in the Ravenna-San Diego collaborative study: the relative prevalence and diagnostic significance of hypomanic features during major depressive episodes. *J Affect Disord*. Dec 2001;67(1-3):115-122.
18. Perlis RH, Ostacher MJ, Goldberg JF, et al. Transition to mania during treatment of bipolar depression. *Neuropsychopharmacology*. Dec 2010;35(13):2545-2552.
19. Dilsaver SC, Chen YW, Swann AC, Shoaib AM, Krajewski KJ. Suicidality in patients with pure and depressive mania. *Am J Psychiatry*. Sep 1994;151(9):1312-1315.
20. Perugi G, Akiskal HS, Micheli C, Toni C, Madaro D. Clinical characterization of depressive mixed state in bipolar-I patients: Pisa-San Diego collaboration. *J Affect Disord*. Dec 2001;67(1-3):105-114.
21. McElroy S, Freeman MP, Akiskal HS. The mixed bipolar disorders. In: Marneros A, Angst J, eds. *Bipolar Disorders 100 years after manic depressive insanity*. Dordrecht: Kluwer Academic Publishers; 2000:63-88.
22. Mitchell PB, Goodwin GM, Johnson GF, Hirschfeld RM. Diagnostic guidelines for bipolar depression: a probabilistic approach. *Bipolar Disord*. Feb 2008;10(1 Pt 2):144-152.
23. Forty L, Smith D, Jones L, et al. Clinical differences between bipolar and unipolar depression. *Br J Psychiatry*. May 2008;192(5):388-389.
24. Berk M, Malhi GS, Cahill C, et al. The Bipolar Depression Rating Scale (BDRS): its development, validation and utility. *Bipolar Disord*. Sep 2007;9(6):571-579.
25. Weinstock LM, Strong D, Uebelacker LA, Miller IW. Differential item functioning of DSM-IV depressive symptoms in individuals with a history of

- mania versus those without: an item response theory analysis. *Bipolar Disord.* May 2009;11(3):289-297.
26. Bernstein IH, Rush AJ, Suppes T, et al. A psychometric evaluation of the clinician-rated Quick Inventory of Depressive Symptomatology (QIDS-C16) in patients with bipolar disorder. *Int J Methods Psychiatr Res.* Jun 2009;18(2):138-146.
  27. Benazzi F. Depressive mixed state: testing different definitions. *Psychiatry Clin Neurosci.* Dec 2001;55(6):647-652.
  28. Bech P. The Bech-Rafaelsen Melancholia Scale (MES) in clinical trials of therapies in depressive disorders: a 20-year review of its use as outcome measure. *Acta Psychiatr Scand.* Oct 2002;106(4):252-264.
  29. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch Gen Psychiatry.* Jun 1961;4:561-571.
  30. Spitzer RL, Kroenke K, Williams JBW, and the Patient Health Questionnaire Primary Care Study Group. Validation and Utility of a Self-report Version of PRIME-MD: The PHQ Primary Care Study. *JAMA.* November 10, 1999 1999;282(18):1737-1744.
  31. Kupfer DJ, Reiger DA. DSM-5: The Future of Psychiatric Diagnosis. <http://www.dsm5.org/Pages/Default.aspx>.
  32. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. *Br J Psychiatry.* Nov 1978;133:429-435.
  33. Bech P, Rafaelsen OJ, Kramp P, Bolwig TG. The mania rating scale: scale construction and inter-observer agreement. *Neuropharmacology.* Jun 1978;17(6):430-431.
  34. Williams JB, Terman M, Link M, Amira L, Rosenthal N. Hypomania Interview Guide (including hyperthymia) (HIGH-C), 1994 Edition (rev. 2000). Vol Clinical Assessment Tools Packet. Broadway, Norwood.: Center for Enviromental Therapeutics; 1994.
  35. Altman EG, Hedeker D, Peterson JL, Davis JM. The Altman Self-Rating Mania Scale. *Biol Psychiatry.* Nov 15 1997;42(10):948-955.
  36. Hirschfeld RM, Williams JB, Spitzer RL, et al. Development and validation of a screening instrument for bipolar spectrum disorder: the Mood Disorder Questionnaire. *Am J Psychiatry.* Nov 2000;157(11):1873-1875.

37. Angst J, Adolfsson R, Benazzi F, et al. The HCL-32: towards a self-assessment tool for hypomanic symptoms in outpatients. *J Affect Disord.* Oct 2005;88(2):217-233.
38. Gonzalez JM, Bowden CL, Katz MM, et al. Development of the Bipolar Inventory of Symptoms Scale: concurrent validity, discriminant validity and retest reliability. *Int J Methods Psychiatr Res.* 2008;17(4):198-209.
39. Zheng YP, Lin KM. The reliability and validity of the Chinese Polarity Inventory. *Acta Psychiatr Scand.* Feb 1994;89(2):126-131.
40. Stevens SS. On the Theory of Scales of Measurement. *Science.* Jun 7 1946;103(2684):677-680.
41. de Ayala RJ. *The Theory and Practice of Item Response Theory.* New York: The Guilford Press; 2009.
42. Jamieson S. Likert scales: how to (ab)use them. *Med Educ.* Dec 2004;38(12):1217-1218.
43. Garson GD. Statnotes: Topics in Multivariate Analysis. Retrieved 04/02/2011. 2011; <http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm> .
44. Furr RM, Bacharach VR. *Psychometrics: An Introduction:* SAGE Publications, Inc.; 2008.
45. American Educational Research Association PA, & National Council on Measurement in Education. . *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.; 1999.
46. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess.* Feb 2009;13(12):iii, ix-x, 1-177.
47. Sijtsma K. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika.* Mar 2009;74(1):107-120.
48. Reise SP, Waller NG, Comrey AL. Factor analysis and scale revision. *Psychol Assess.* Sep 2000;12(3):287-297.
49. Embretson SE, Reise SP. *Item Response Theory for Psychologists.* Mahawah, New Jersey: Lawrence Erlbaum Associates, Publishers; 2000.
50. Sijtsma K, Molenaar IW. *Introduction to Nonparametric Item Response Theory.* Vol 5. Thousand Oaks: SAGE Publications; 2002.
51. Molenaar IW, Sijtsma K. *Introduction to non parametric Item Response Theory:* SAGE Publications; 2002.

52. van der Ark LA. Mokken Scale Analysis in R. *Journal of statistical computing*. Februari 2007 2008;20(11).
53. Meijer RR, Baneke JJ. Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychol Methods*. Sep 2004;9(3):354-368.
54. van der Ark AL. Mokken Scale Analysis in R. *Journal of Statistical Software*. May 2007;20(11):1-19.
55. De Koning E, Sijtsma K, Hamers JHM. Comparison of Four IRT Models When Analyzing Two Tests for Inductive Reasoning. *Applied Psychological Measurement*. September 1, 2002 2002;26(3):302-320.
56. Gillespie M, Tenvergert EM, Kingma J. Using Mokken scale analysis to develop unidimensional scales. *Ouality & Quantity*. 1987(21):393-408.
57. Wilson M. *Constructing measures. An Item Response Modeling Approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers; 2005.
58. Reise SP, Haviland MG. Item response theory and the measurement of clinical change. *J Pers Assess*. Jun 2005;84(3):228-238.
59. Doucette A, Wolf AW. Questioning the measurement precision of psychotherapy research. *Psychother Res*. Jul 2009;19(4-5):374-389.
60. Zumbo BD. *A Handbook on the Theory and Methods of Differential Item Functioning (DIF)*: Directorate of Human Resources Research and Evaluation, National Defense Headquarters, Ottawa, Canada;1999.
61. Nguyen HT, Clark M, Ruiz RJ. Effects of acculturation on the reporting of depressive symptoms among Hispanic pregnant women. *Nurs Res*. May-Jun 2007;56(3):217-223.
62. Prieto G, Delgado AR, Perea MV, Ladera V. Differential functioning of minimal test items according to disease. *Neurologia*. Mar 15 2011.
63. Reise SP, Waller NG. Item response theory and clinical measurement. *Annu Rev Clin Psychol*. 2009;5:27-48.
64. Foley BP. *Improving IRT parameter estimates with small sample sizes: Evaluating the efficacy of a new data augmentation technique*. Lincoln: University of Nebraska - Lincoln, University of Nebraska - Lincoln; 2010.
65. Spearing MK, Post RM, Leverich GS, Brandt D, Nolen W. Modification of the Clinical Global Impressions (CGI) Scale for use in bipolar illness (BP): the CGI-BP. *Psychiatry Res*. Dec 5 1997;73(3):159-171.
66. Linacre JM. Winsteps 3.70.0 help section: Winsteps Rasch measurement; 2010.

67. Davison AC, Hinkley DV. *Bootstrap Methods and their Application*: Cambridge University Press; 1997.
68. Fournier JC, DeRubeis RJ, Hollon SD, et al. Antidepressant drug effects and depression severity: a patient-level meta-analysis. *JAMA*. Jan 6 2010;303(1):47-53.
69. Philipp M, Kohlen R, Hiller KO. Hypericum extract versus imipramine or placebo in patients with moderate depression: randomised multicentre study of treatment for eight weeks. *BMJ*. Dec 11 1999;319(7224):1534-1538.
70. DeRubeis RJ, Hollon SD, Amsterdam JD, et al. Cognitive therapy vs medications in the treatment of moderate to severe depression. *Arch Gen Psychiatry*. Apr 2005;62(4):409-416.
71. Elkin I, Shea MT, Watkins JT, et al. National Institute of Mental Health Treatment of Depression Collaborative Research Program. General effectiveness of treatments. *Arch Gen Psychiatry*. Nov 1989;46(11):971-982; discussion 983.
72. Wichers MC, Barge-Schaapveld DQ, Nicolson NA, et al. Reduced stress-sensitivity or increased reward experience: the psychological mechanism of response to antidepressant medication. *Neuropsychopharmacology*. Mar 2009;34(4):923-931.
73. Barrett JE, Williams JW, Jr., Oxman TE, et al. Treatment of dysthymia and minor depression in primary care: a randomized trial in patients aged 18 to 59 years. *J Fam Pract*. May 2001;50(5):405-412.
74. Bech P, Gram LF, Dein E, Jacobsen O, Vitger J, Bolwig TG. Quantitative rating of depressive states. *Acta Psychiatr Scand*. Mar 1975;51(3):161-170.
75. Faries D, Herrera J, Rayamajhi J, DeBrotta D, Demitrack M, Potter WZ. The responsiveness of the Hamilton Depression Rating Scale. *J Psychiatr Res*. Jan-Feb 2000;34(1):3-10.
76. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. *Arthritis Res Ther*. Sep 14 2011;13(5):R147.

## 10 APPENDIX: THE AFFECTIVE SELF RATING SCALE

 **Stockholms läns landsting**  
Affektiva mottagningen M59  
Fax 585 866 30  
Tel 585 866 26, 585 866 34

**AS-18**  
Affektiv självskattningsskala

NAMN: .....

PERS-NR: ..... DATUM: .....

Hur stora problem har du haft under <i>den senaste veckan</i> med: (Ringa in det alternativ som stämmer bäst).	Inga	Små	Måttliga	Stora	Mycket Stora
1 Att du varit så pratsam att andra tyckt det varit svårt att komma till tals.	0	1	2	3	4
2 Att du sovit mer än vanligt.	0	1	2	3	4
3 Att du behövt sova mindre och ändå varit pigg.	0	1	2	3	4
4 Att du känt hopplöshet.	0	1	2	3	4
5 Att du rört dig långsammare än vanligt.	0	1	2	3	4
6 Att du varit uppvarvad eller överaktiv.	0	1	2	3	4
7 Att du varit kroppsligt rastlös så att det har varit svårt att sitta stilla.	0	1	2	3	4
8 Att dina tankar rusat snabbt i huvudet.	0	1	2	3	4
9 Att du varit lättirriterad.	0	1	2	3	4
10 Att du känt dig nedstämd eller deprimerad.	0	1	2	3	4
11 Att du inte kunnat glädja dig eller intressera dig för sådant du annars tycker om.	0	1	2	3	4
12 Att du saknat energi.	0	1	2	3	4
13 Att du har haft skuld känslor och känt dig värdelös.	0	1	2	3	4
14 Att dina tankar har gått trögt och långsamt.	0	1	2	3	4
15 Att du haft alltför hög självkänsla.	0	1	2	3	4
16 Att du har haft alltför stark känsla av glädje och intresse.	0	1	2	3	4
17 Att du haft tankar på att skada dig själv eller ta ditt liv.	0	1	2	3	4
18 Att du tagit risker, t ex med pengar, i trafiken eller i kontakten med andra människor.	0	1	2	3	4

**OBS! Alla frågor gäller problem du haft under *den senaste veckan*.**

*Mats Adler, Affektiva mottagningen M59, Karolinska Universitetssjukhuset i Huddinge, 070612.*

## **Translation of the Swedish original of AS-18**

Items of the Affective Self Rating Scale, translated from Swedish by the authors.

Response categories are “none”, “a little”, “moderate”, “severe” and “very severe”, graded numerically from 0 to 4. Items labelled (M) are included in the mania subscale and items labelled (D) are included in the depression subscale.

### **During the last week, to which extent have you experienced the following problems?**

- 1) (M) Talkativeness. “. . . having been so talkative that it has been hard for others to make themselves heard?”
- 2) (D) Increased sleep. “. . . sleeping more than usual”.
- 3) (M) Less need for sleep. “. . . having less need for sleep but still felt energetic and awake?”
- 4) (D) Hopelessness. “. . . feeling hopeless?”
- 5) (D) Retardation. “. . . your movements have been slower?”
- 6) (M) Overactive. “. . . being wound up or overactive.”
- 7) (M) Agitation. “. . . being so physically restless that you have had trouble keeping still?”
- 8) (M) Racing thoughts. “. . . that your thoughts race.”
- 9) (M) Irritability. “. . . that you have been easily irritated?”
- 10) (D) Depression. “. . . feeling low or depressed?”
- 11) (D) Anhedonia. “. . . inability to take an interest or pleasure in things that you normally enjoy?”
- 12) (D) Low energy. “. . . a lack of energy?”
- 13) (D) Guilt. “. . . feelings of guilt or worthlessness?”
- 14) (D) Slow thinking. “. . . that your thoughts have been sluggish and slow?”
- 15) (M) Increased self-esteem. “. . . that you have been over confident?”
- 16) (M) Euphoria. “. . . that you have had an overly strong sense of happiness and increase in interest?”

17) (D) Suicidal ideation. “. . . that you have had thoughts of harming yourself or taking your own life?”

18) (M) Risk-taking. “. . . that you have been taking risks; for example with money, in traffic or in your social contacts?”