Item analysis of single-peaked response data : the psychometric evaluation of bipolar measurement scales

Polak, M.G.

Chapter 1

General Introduction

1.1 The Application of Measurement Scales in Psychological Research

Since over a century (i.e., since Cattell, 1890), psychological scaling methods have become more and more common in everyday life. The goal of psychological (psychometric) scaling is to compare individuals in terms of certain psychological qualities; to name just a few: intelligence, personality traits, job aptitude, leadership qualities, preference of food products, attitude toward abortion, quality of life, depression, and anxiety.

The field of psychometrics is concerned with the development of valid and reliable (psychological) measurement scales. The challenge lies in finding stimuli (or items) that reflect different locations on the scale, so that, subsequently, individual responses to these stimuli can be used to measure the individual differences on the same scale. For instance, an individual's "answers" to the items of an intelligence test are used to determine his location on the intelligence scale. Individuals who answer more items correctly have higher intelligence scores, and vice versa, those who answer less items correctly have lower intelligence scores.

Conversely to locating individuals on a scale, psychological scaling is also used to scale the items themselves. For instance, we can determine the relative positions of various statements concerning abortion in the view of a group of persons. In fact, in psychometric (data-) analysis these two procedures are alternated: on the one hand we analyze item responses to select those items that together constitute a valid and reliable measurement scale; and, on the other hand, we analyze the responses to these items to pinpoint individuals on the scale. Obviously, the process of constructing and evaluating psychological measurement scales requires several samples of individuals from the population of interest (e.g., Dutch pre-schoolers, people suffering from phobic fear, job applicants, or consumers of cosmetic products).

The American psychologist Louis Thurstone was among the fist psychometricians who offered a number of theoretical and statistical rationales for constructing psychometric scales (e.g., Thurstone, 1927, 1928). Ever since his founding work, there has been a large-scale development of quantitative or statistical methods to construct and evaluate psychological measurement scales. However, psychological measurement scales and their evaluation are, and always have been, an issue of ongoing controversy and debate (for a recent example, see Borsboom, 2006a, 2006b; Clark, 2006; Heiser, 2006; Kane, 2006; Sijtsma, 2006). We refer the reader to Borsboom (2005) for a recent treatment of the issue of measurement in psychology.

In the preface of his book on measuring attitude toward the church, Thurstone commented the following about the challenges of psychological measurement:

> The true allocation of an individual to a position on an attitude scale is an abstraction, just as the true length of a chalk line, or the true temperature of a room, or the true spelling ability of a child is an abstraction. We estimate the true length of a line, the true temperature of a room, or the true spelling ability of a child by means of various indices, and it is a commonplace in measurement that all indices do not agree exactly. In allocating an individual to a point on the attitude continuum we may use various indices, such as the opinions that he endorses, his overt acts, and his past history, and it is to be expected that discrepancies will appear as the "true" attitude of the individual is estimated by different indices. The present study is concerned with the allocation of individuals along an attitude continuum based on the opinions that they accept or reject. (Thurstone & Chave, 1929, p. xii)

The present thesis is a contribution to the vast body of work containing many practical and (statistical-)theoretical approaches to the psychometric evaluation of measurement scales. In particular, the aim is to contribute to the psychometric evaluation of so-called *bipolar* measurement scales, specifically, a scale for personality development that ranges from maladaptive functioning to adaptive functioning. A key characteristic of such scale is that there are two theoretically opposite poles. It will be argued that this type of scales needs to be distinguished from the more common unipolar scales (e.g., a scale that measures the severity of maladaptive functioning, hence a scale with only one theoretical pole), because it requires a whole other type of statistical and theoretical approach to psychometric evaluation.

The distinction between unipolar and bipolar scales, and the consequences this distinction yields for psychometric evaluation, will be explained in more detail in the next section. The remaining sections of this introduction present, first, a classification of scaling techniques for item response data; second, an introduction to using correspondence analysis as an approach to scaling items and persons on a (psychological) bipolar scale, which is the core topic of this thesis; third, an historical perspective on the development of scale construction methods, and fourth an overview of the remaining chapters of this thesis.

In this thesis, items intended to reflect positions on bipolar scales, are referred to as single-peaked response items. Furthermore, models that deal with single-peaked item responses are generally referred to as unfolding models (cf. Coombs, 1964); and likewise the analysis of single-peaked item responses, is generally referred to as unfolding analysis. Furthermore, we will refer to the item response function (IRF) to indicate the function that describes the relationship between the probability of a positive response and the position of the responding individual on the underlying measurement scale.

1.2 Unipolar versus Bipolar Measurement Scales

Response items can be classified as either dominance items or proximity items. Dominance items are organized on unipolar (or cumulative) measurement scales, for which the item responses are monotonically related to a subject's position on this scale (see Figure 1.1, bottom, for an example of a typical monotonic response function). Unipolar scales are typically found in ability research, where items (or tasks) can be ordered from very simple to very difficult, and subjects can be ordered from poorly skillful to highly skillful. Subjects' locations are typically based on their total score, that is, the sum score indicating the total number of items a subject answers correctly.

In contrast, proximity items lie on bipolar (or substitutive) scales, for which the item responses are a single-peaked function of the distance between the position of the item, and the position of the subject on the scale; the closer an item is located near the subject's position on the scale, the higher the value of the expected response (see Figure 1.1, top, for examples of typical single-peaked response functions). Single-peaked items typically arise in the fields of measurement of psychological development (cf. Noel, 1999), personality measurement (cf. Chernyshenko, Stark, Drasgow, & Roberts, 2007; Weekers & Meijer, 2008), and the measurement of preferences (cf. Ashby & Ennis, 2002) and attitudes (cf. Andrich & Styles, 1998).

Subjects' positions on bipolar scales cannot be based on their total scores. Rather, their positions are determined by computing the mean position associated with endorsed items. For instance, suppose we have a number of drinks varying from sweet to sour, and we ask subjects to taste the drinks and subsequently to indicate which of the drinks they liked. Now we expect people who prefer sweet, to pick only the sweet drinks, and people who prefer sour, to pick only the sour drinks. Thus, the total number of drinks a subject picks tells us nothing about his position on the sweet-sour scale. Instead, the subject's position can be found amongst the drinks he picked, for instance, by taking the mean (or median) of the scale values of the preferred drinks (cf. Thurstone's scaling approach, e.g., Thurstone, 1928).

In Figure 1.1, reprinted from Thurstone and Chave (1929, p. 94), examples are given of item response functions *avant la lettre*, typical for either bipolar scales (top), or unipolar scales (bottom).

Thurstone and Chave did not use the typology of bipolar and unipolar scales (nor did they use the term item response function), but rather referred to these scales as, respectively, *maximum probability type of scale* and *increasing probability type of scale*. To start with the first, the *maximum probability type of scale* refers to the principle that items are most likely to be endorsed by those subjects who are scaled at a minimum distance of a particular item. In Figure 1 (top) we see two "item response functions" that depict the probability of endorsement for two single-peaked items. The item's scale value, that is, the location of each item on the attitude scale, is indicated by the peak of the response function. Hence, people whose scale value coincides with the item's scale value are most likely to endorse this item. Conversely, people whose true attitude is more distinct from the attitude that the item reflects, are less likely to endorse it. Hence, the probability function decreases in both directions from the item's location on the attitude scale. The height of the function reflects the item's popularity: the higher the peak of the function, the more popular an item is.

For a bipolar scale one could say that each item on such a scale reflects a specific balance between both poles of the scale. As an example, suppose we have a scale that measures the attitude toward capital punishment, ranging from very much against it to very much in favor of it. Items near the midpoint of the scale reflect both pros and cons (e.g., "I do not believe in capital punishment, but it is not practically advisable to abolish it"), while items located between the left ("contra") pole and the center of the scale are more against capital punishment than in favor of it (e.g., "Life imprisonment is more effective than capital punish-

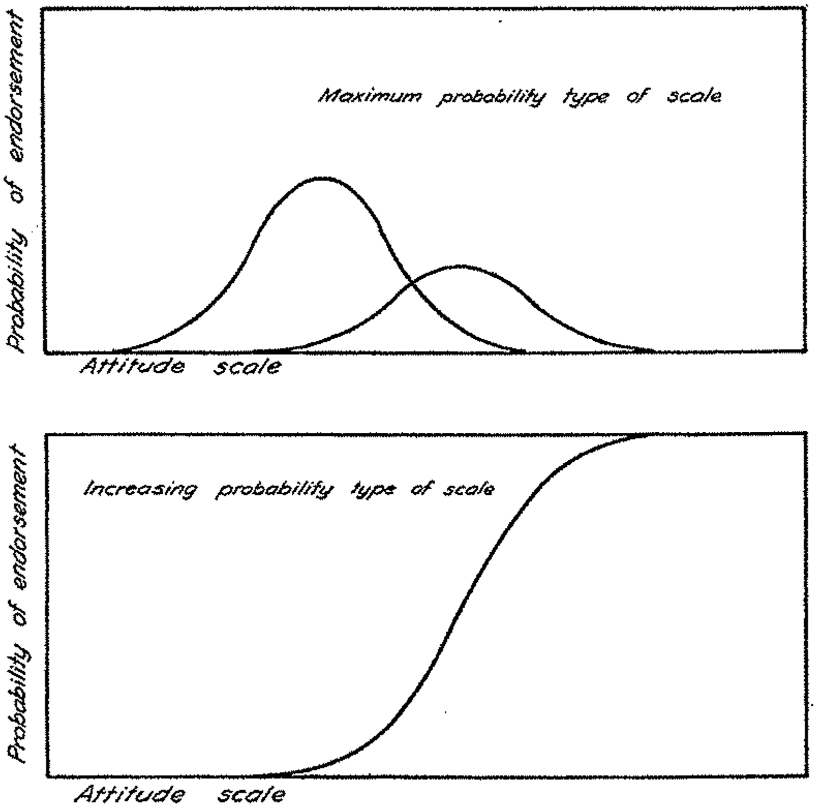


Figure 1.1: Probabilities of endorsement of items on a bipolar scale (top) and a unipolar scale (bottom) from Thurstone and Chave (1929, p. 94).

ment"). Items that are located on the left pole of the scale are unambiguously against capital punishment ("I do not believe in capital punishment under any circumstances").

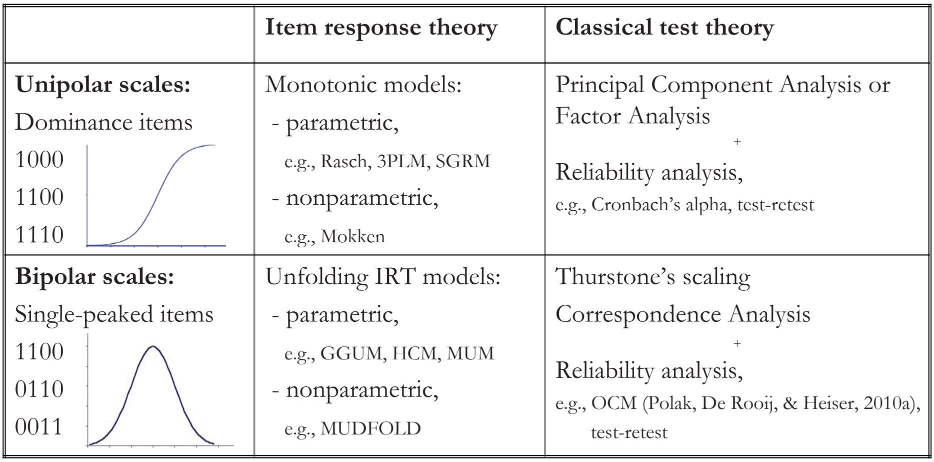
Second, the *increasing probability type of scale* refers to the principle that the probability of endorsing a certain item, increases with the subject's scale value. In Figure 1 (bottom) we see an example of an "item response function" depicting the probability of endorsement for a monotonic item. Thurstone and Chave (1929) give as example of a unipolar scale a scale that indicates the seriousness of crimes. Suppose we present subjects a list of crimes, and subjects are asked to check those crimes in the list which they consider serious enough to deserve capital punishment. Then it seems likely that the probability of checking a crime increases with the seriousness of the crime.

For the measurement of attitude, Thurstone and Chave (1929, p.96) conclude that: "it is probable that most issues will be better described if the scale is intentionally constructed so that a person is more likely to endorse the opinions at some one part of the scale than at any other part. Such as scale is the maximum probability type."

In Section 1.3 below we give an overview of the various approaches that have been developed for analyzing unipolar and bipolar scales, respectively.

1.3 Scaling Techniques for Item Response Data

In Figure 1.2, a scheme is given that classifies the various approaches for analyzing unipolar scales (i.e., dominance items) and bipolar scales (i.e., single-peaked items), respectively. The lower right cell of the scheme in Figure 1.2 shows the methodology in the collection of scaling techniques that this thesis aims to provide as a successor of Thurstone's earlier methodology (cf., Thurstone, 1927, 1928; Thurstone & Chave, 1929).



	Item response theory	**Classical test theory**
Unipolar scales: Dominance items 1000 1100 1110	Monotonic models: - parametric, e.g., Rasch, 3PLM, SGRM - nonparametric, e.g., Mokken	Principal Component Analysis or Factor Analysis + Reliability analysis, e.g., Cronbach's alpha, test-retest
Bipolar scales: Single-peaked items 1100 0110 0011	Unfolding IRT models: - parametric, e.g., GGUM, HCM, MUM - nonparametric, e.g., MUDFOLD	Thurstone's scaling Correspondence Analysis + Reliability analysis, e.g., OCM (Polak, De Rooij, & Heiser, 2010a), test-retest

Figure 1.2: Classification of scaling techniques for item response data, with correspondence analysis as proposed CTT-like approach to the analysis of single-peaked items.

Item analysis is usually based either on classical test theory (CTT), or on item-response theory (IRT). CTT-based item analysis has been developed exclusively for unipolar scales (i.e., dominance items), and consists of two steps, that is, factor analysis (FA) or principal component analysis (PCA) with Varimax rotation followed by reliability analysis (e.g., computation of test-retest reliability or Cronbach's alpha; Cronbach, 1951). The goal of PCA is to determine the homogeneity, or dimensionality, of the items based on their inter-correlations. For a (sub-)set of unidimensional dominance items, Cronbach's alpha gives an estimate of the (lower bound of the) reliability of the total score.

For IRT-based item analysis, a probabilistic model is used to describe the relationship between a response and the underlying measurement scale. Item and subject characteristics are now parameters of a model that are estimated from the data. IRT-based item analysis was originally exclusively developed for unipolar scales (or dominance items), but today, it also provides models for bipolar scales (or single-peaked items); either nonparametric (e.g., MUDFOLD; Van Schuur, 1984), or parametric, (e.g., PARELLA; Hoijtink, 1991; HCM; Andrich & Luo, 1993; GGUM; Roberts, Donoghue, & Laughlin, 2000; MUM; Javaras & Ripley, 2007). These models are often referred to as unfolding IRT models (see Andrich, 1988, for an introduction to this type of models).

The analysis of bipolar scales (or single-peaked items), as described in the bottom cells in the scheme in Figure 1.2, is the topic of Chapter 2. In this chapter, CA is compared with unfolding IRT models for the psychometric evaluation of bipolar measurement scales. A comparison of CA with PCA is described in Chapter 3 (see also Polak, Heiser, & De Rooij, 2009). The reliability analysis for single-peaked items that is also mentioned in the lower right cell in Figure 1.2 is addressed in Chapter 4 of this thesis. Chapter 4 presents diagnostics for single-peakedness of item responses based on ordered conditional means (OCM; see also Polak, De Rooij, & Heiser, 2010a). In Section 1.4 below, we explain the relevance of CA as CTT-like approach to the analysis of single-peaked items, that is, the psychometric evaluation of bipolar measurement scales.

1.4 Correspondence Analysis as a Tool in the Psychometric Evaluation of Bipolar Measurement Scales

CA is a multivariate technique primarily developed for the analysis of contingency table data (for a practical introduction, see Greenacre, 2007). However, the technique can be applied also directly on a subject by item data table, as long as the entries of the table can be considered measures of association strength between row entries and column entries. The association measure is assumed to be some non-negative quantity, where lack of association is indicated by a zero entry (Heiser, 2001). Computational details and the rationale of using CA for analyzing single-peaked item response data are presented in Appendix A of this thesis.

CA is also known as reciprocal averaging (Horst, 1935), homogeneity analysis (Gifi, 1990), and dual scaling (Nishisato, 1996). These methods differ with respect to the underlying rationale (for a recent bibliographic review, see Beh, 2004).

In the field of ecology, CA has been popular for several decades (since Hill, 1973) as a method to estimate the optima of species (e.g., birds, spiders, or plants) on some bipolar environmental gradient (e.g., vegetation structure or pH-value)(cf. Ter Braak & Prentice, 1988, 2004). The data table in ecology is often a species by sites table, where the entries of the table indicate the incidence (1/0 indicating presence/absence) or abundance (e.g., number of individuals of each species present) of a species in each site. Sites are chosen so that they represent the environmental variable evenly (e.g., Ph-values ranging from acidic to basic). In addition to continuing a rich research tradition in psychometrics, this thesis builds on the extensive body of research by ecologists on the application of CA to single-peaked data. For this purpose, we translated the findings from the field of ecology to the field of psychology.

Why should we be interested in CA as an extension of the CTT approach to item analysis of single-peaked items (see Section 1.3), when besides CTT, IRT-based methods have already been developed, providing models for both dominance and single-peaked items? One reason is that, despite all IRT development for evaluation of unipolar scales (i.e., dominance items), the CTT approach is still extremely popular among practical researchers.

One explanation for the persisting popularity of FA/PCA and Cronbach's alpha is that these techniques are available in SPSS and are (partly for this reason) still the basic tools for scale evaluation that are taught to psychology students all over the world. Since CA is also available in SPSS (categories module, Meulman & Heiser, 2004) and SAS/STAT (CORRESP procedure, SAS institute, 2008) it is, for the analysis of single-peaked items, potentially as attractive to practitioners as FA/PCA is for the analysis of dominance items. Even more so, because its output is comparable to PCA (e.g., variance accounted for per dimension, nested dimensions, a perfect solution can be found as long as sufficient dimensions are chosen).

Other advantages of CA are that, first, it is computationally straightforward and has a unique solution. Second, it handles any sort of data as long as the entries of the data matrix contain measures of association strength, which are positive and where 0 indicates lack of association.

Finally, another favorable property of CA in the context of item analysis is that it allows for incorporating explanatory variables in the analysis (cf. explanatory monotonic IRT; De Boeck & Wilson, 2004), which approach has not yet been incorporated in unfolding IRT methods. The technique of incorporating explanatory variables in CA is called constrained (or canonical) correspondence analysis (CCA) (Ter Braak, 1986, 1987; Takane, Yanai & Mayekawa, 1991; Takane & Hwang, 2002). In CCA the dimensions in the solution are constrained to be linear combinations of the explanatory variables, along which the subjects are maximally separated.

1.5 Why Bipolar Scales Are Not Commonly Used in Psychological Measurement: Thurstone versus Likert Scaling

The current thesis is concerned with the psychometric evaluation of bipolar scales, and hence the item analysis of single-peaked response items that are typical for this type of scales. The goal of this section is first, to explain why bipolar scales are not common in psychological research, even when the psychological construct that is being measured is bipolar in nature (e.g., attitude scales, preference scales, or developmental scales ranging from immature to mature). Second, we argue that researchers who are investigating dimensions with opposite extremes, should become more aware of the various scaling techniques that are intentionally developed for the psychometric evaluation and scoring of such scales.

Thurstone's method of constructing and scoring attitude scales. One of the earliest examples of the bipolar scale in psychological research is the Thurstone scale (cf. Thurstone, 1928). Thurstone's scaling procedure for constructing and scoring scales with two opposite extremes, such as attitude scales ranging from very much against a certain issue to very much in favor of it, are very attractive because of their strong theoretical basis (see also Section 2 of the current chapter).

Thurstone scales were introduced in the 1920s to measure attitudes toward controversial issues, such as birth control, communism, war, and even “the Negro”. To measure people's attitude, a set of statements was formulated intended to cover the whole possible range of opinions from extremely pro the issue to extremely contra the issue. Subsequently, according to Thurstone's procedure, a small group of expert-judges was asked to sort these statements onto (always) 11 piles, starting with the item that the judge thought to reflect the most unfavorable opinion toward the issue and on the eleventh pile, the item reflecting the most favorable opinion toward the issue. Scale values for the items were derived as the median "pile number" based on the judges' sorting. Note, that in those days this was all done by hand.

Subsequently, the subjects in the sample from the population of interest were asked to indicate for each of these statements whether or not they agreed with the statement. For the sake of comparison, suppose the response format was a 5-point scale (with 1= strongly disagree, 2 = disagree, 3 = undecided, 4 = agree, 5 = strongly agree). According to the Thurstone procedure, a subject's position that was on the same attitude scale that also included the positions of all statements, can be found as the mean scale value of all statements with which the subject expressed strong agreement.

Clearly, the procedure for development and scoring of a Thurstone scale was rather time-consuming, since at the time when it was fist introduced, there were no computers to facilitate processing of the research data.

Likert's method of constructing and scoring attitude scales. Not surprisingly (but unfortunately), Thurstone's method fell into disuse some decades after Rensis Likert introduced a "simple and reliable method of scoring the Thurstone attitude scales" (cf. Likert, 1932; Likert, Roslow & Murphy, 1934). The most important changes were that, first, the step of using judges to determine item locations on the scale, was omitted from the procedure. Instead, no differentiation between the items in terms of their location on the scale was made, only the following distinction: items were regarded as either favorable toward the issue (nowadays also referred to as indicative items), or unfavorable toward the issue (nowadays also referred to as contra-indicative items).

The second change with respect to the original Thurstone procedure -which was an implication of the first- was that all midpoint items that reflected a more nuanced attitude toward the issue, were discarded.

The third change was, that subjects' positions on the attitude scale were now derived from their total score, thus adding up the responses on the 5-point scales over all items. For this purpose, the set of items reflecting an unfavorable opinion was reverse-scored. That is, the responses to the contra-indicative items were scored as follows: 1 = 5, 2 = 4, 4 = 2 and 5 = 1. Subsequently, the total scores were interpreted as the degree to which a person was favorable toward the issue.

Likert, Rowlow and Murphy (1934, p. 230) explain the reverse-scoring as follows:

> Thus, for example, statement number five, form A, of the Attitude toward the Negro Scale is, "I place the Negro on the same social basis as I would a mule." Obviously to "strongly agree" with this statement reflects an attitude which is "more unfavorable to the Negro" than "to strongly disagree"; consequently, the "strongly agree" alternative is given a score of 1 and "strongly disagree" is scored 5. This assumes, of course, that we have previously designated the "most favorable to the Negro" extreme of the attitude continuum as the numerically high position. In a similar manner, numerical values from 1 to 5 are assigned to each alternative for each statement in the attitude scale.

Note that in doing so, the Likert procedure transformed the *bipolar* Thurstone scale into a *unipolar* scale. Hence, a Likert scale measures the degree to which a subject is in favor of a certain issue (cf. the maximum probability type of scale described in Section 2).

The relative simplicity of the Likert procedure and later, the introduction of factor analysis (ironically Thurstone introduced (multiple-)factor analysis in the field of mental testing in psychology, cf. Thurstone, 1931, 1934) as a generally accepted method for psychometric evaluation of, in particular, unipolar scales, explain the dominance of Likert scales in psychological research up until today.

Why psychologists need to re-adopt Thurstone's rationale for scale construction and scoring. This thesis calls for reappraisal of the theoretical basis of the Thurstone procedure, which seems to be in some aspects a more valid approach to the measurement of those psychological constructs that are bipolar in nature (cf. Andrich, 1996; Roberts, Laughlin, & Wedell, 1999). We will explain this on the basis of two main points of criticism with respect to the use of Likert scales for measuring bipolar constructs.

The first point of criticism concerns the analysis of subjects, that is, determining subject locations on the measurement scale. Roberts, Laughlin, and Wedell (1999) showed that the Likert procedure for scaling subjects (i.e., computing the total score after reverse-scoring the contra-indicative items), results in underestimation of the locations of the most extreme subjects (see also Chapter 3 of this thesis). This drawback of the Likert procedure calls attention to an important distinction between unipolar and bipolar scales that will be elaborated on in the following.

Subjects with an extremely positive opinion, for example, toward capital punishment, tend to agree exclusively with the most extremely formulated favorable (or indicative) items. Items that are only mildly favorable toward the issue (e.g., "Capital punishment is justified only for premeditated murder") are less agreeable to these extreme subjects (in the example given, because of the word "only"). Empirically, this phenomenon can be recognized, by inspecting the item response functions for these moderately positive (and moderately negative) items. Namely, these response functions increase toward the point where the item is located on the scale, but then decrease again from that point onward, to the extreme end of the scale (see for example, Figure 6 in Chapter 4).

Thus, even items selected on the basis of the Likert procedure, show some degree of "bending" (i.e, single-peakedness) of the item response functions. This explains why deriving subjects' locations from their total scores, leads to a misrepresentation of the locations of subjects with relatively extreme opinions.

In short, Roberts, Laughlin, and Wedell (1999) showed that when the psychological dimension that is being measured is in essence bipolar, the procedure of deleting the midpoint items and reverse-scoring the contra-indicative items, does not provide indisputably valid subject locations.

A second point of criticism with respect to the use of Likert scales for measuring bipolar constructs, concerns the item analysis, that is, determining item locations and the psychometric quality of the items. The Likert procedure instructs to remove midpoint items from the scale. This is a necessary condition for deriving subjects' locations from their total scores. This omission of midpoint items in attitude research is often justified by pointing at the ambivalence in the item wording. This ambivalence is the result of the fact that in neutral opinions pros and cons of the issue are assumed to be more or less balanced. The argument for removing ambivalent items is, that if an attitude statement reflects both a favorable and an unfavorable property, it is possible that one subject responds more to one part of the statement, while another subject responds more to the other part. Obviously, this is an undesired item-characteristic. Likert, Roslow, and Murphy(1934, p. 230) write:

> An illustration of such a statement is number 5 in form A of the Dobra War Scale: "Compulsory military training in all countries should be reduced but not eliminated". If a subject "agrees" while following

> the present directions it is impossible to say whether he is agreeing with the "reduction" aspect of this statement or the "not eliminated" aspect. A person who strongly opposes compulsory military training would disagree or strongly disagree with the "not eliminated" aspect, whereas a person who favors compulsory military training would probably disagree or strongly disagree with the "reduction" aspect of the statement. Obviously for the present 1-to-5 method of scoring the statement is double-barreled and of little value because it does not differentiate persons in terms of their attitudes. Persons at opposite ends of the attitude continuum may at times check the same alternative

Today, software and computer capacity make it easy for researchers to evaluate the psychometric quality of their measurement instrument on the basis of empirical test data. Thus there seems no need to discard the presumed ambiguous items beforehand. Instead, it seems more plausible to explicitly test the presumptions that midpoint items "do not differentiate persons in terms of their attitude", or that for midpoint items "persons at opposite ends of the continuum may at times check the same alternative". After all, this is the main purpose of item analysis in general!

To sum up, we argue that researchers who are concerned with measuring constructs that are bipolar in nature, should strive for formulating items that cover the entire range of the scale, thus including midpoint items. This not only enhances the content validity of the scale, but may also provide more information to discriminate among subjects who are located around the midpoint of the scale. Item analysis should point out which items are suited to measure the bipolar construct consistently. This thesis investigates the use of correspondence analysis to select items that together form a bipolar scale, and to estimate both item locations and subject locations on this scale. Furthermore, we propose a method to investigate the psychometric quality of the individual items, as well as the internal consistency of the scale as a whole.

1.6 Outline of this Thesis

The chapters of this thesis are in fact four individual papers. Apart from Chapter 4, all chapters are concerned with the use of correspondence analysis (CA) for the psychometric evaluation of bipolar measurement scales. Chapter 4 is concerned with diagnostics for single-peakedness of item responses that can be used in combination with any unfolding method.

In this thesis, items intended to reflect positions on bipolar scales are referred to as single-peaked response items. Furthermore, models that deal with single-peaked item responses are generally referred to as unfolding models. Likewise the analysis of single-peaked item responses is generally referred to as unfolding analysis. The content of the chapters is as follows.

In Chapter 2, we elaborate on various approaches to the psychometric evaluation of bipolar scales. Specifically, we present a theoretical basis for using CA as unfolding technique. In this chapter, the aim is to study the merits of CA compared to unfolding IRT methods, which are also suited for analyzing bipolar scales. We compare these different approaches using results from both simulated and real data. Furthermore, we propose constrained correspondence analysis (CCA) as an approach to explanatory (unfolding) IRT.

In Chapter 3, we deal with the topic of evaluating bipolar scales again, but here we stress that principal component analysis (PCA) is unsuited for this particular purpose. We compare CA with PCA in analyzing simulated and real data. We explain the two main approaches to data coding in CA, and demonstrate which type of data coding to choose for correctly representing single-peaked response items. Furthermore, we show how to recognize single-peakedness of item responses based on the inter-item correlations. We distinguish between different types of unfolding models, and show that these models yield different patterns of inter-item correlations.

In Chapter 4, we present two diagnostics for single-peakedness of item responses. The proposed methodology approximates item response functions (IRFs) of all individual items by computing ordered conditional means (OCM), given a hypothesized ordering of items on a bipolar scale. By fitting a unimodal smoother to each approximated IRF, we attempt to identify deviant items, in particular items with an irregular, or flat IRF. We aim to use the methodology for estimating the internal consistency of the item responses for the scale as a whole, as well as for computing a diagnostic for individual item fit. We evaluate the OCM diagnostics using results from both simulated and real data.

Chapter 5 comprises the applied part of this thesis, that is, the psychometric evaluation of the Developmental Profile (DP; Abraham, 1993, 2005; Abraham et al., 2001). The DP is an instrument for personality assessment. The DP's subscales cover either varying degrees of maladaptive, or varying degrees of adaptive personality characteristics. Since these subscales reflect different positions on a bipolar scale, ranging from strongly maladaptive functioning to strongly adaptive functioning, the psychometric evaluation of the DP was not straightforward. We present a combination of confirmatory factor analyses (complemented with Cronbach's alpha coefficients) and CA to evaluate the main theoretical assumptions underlying the DP. A large sample is studied ($N = 763$) with participants from various clinical and nonclinical settings in the Netherlands.

Chapter 6 consists of three parts. The first part presents the main conclusions of Chapters 2, 3 and 4. The second part summarizes and discusses the results presented in Chapter 5. In the third part we discuss the results of Chapters 2, 3 and 4, and we end with potential directions for future research.