Chapter 2

The Psychometric Evaluation of Bipolar Measurement Scales: Correspondence Analysis as an Alternative to Unfolding IRT Models

Abstract

In psychometrics, an increasingly popular approach for evaluating bipolar scales is unfolding item response theory (unfolding IRT). However, unfolding IRT, like monotonic IRT, still has some drawbacks compared to techniques that are based on the classical test theory (CTT), which explain the persisting popularity of the latter among practical researchers. In the current chapter we propose correspondence analysis (CA), which is available in SPSS, as a CTT-like counterpart of the unfolding IRT models for the analysis of bipolar scales. This technique has been widely used in the field of ecology for scaling species and sites along an environmental gradient. In this study we compare CA with both a parametric and a nonparametric unfolding IRT model: GGUM (Roberts, Fang, Cui, & Wang, 2006) and MUDFOLD (Van Schuur & Post, 1998), respectively. For this purpose we use simulated data, as well as psychological data from the fields of personality assessment and attitude research. Finally, we explore the surplus value of constrained CA that allows incorporating additional explanatory variables in the analysis, which option is not yet available in unfolding IRT.

2.1 Introduction

Item response data are made up of responses of a group of subjects to a set of questions or stimuli (items). Usually the interest lies in whether the variation in these responses can be attributed to systematic differences among subjects, as well as among items on a so-called measurement scale. During the process of scale
construction items are selected on the basis of item analysis. The goal of item analysis is to determine whether there is one underlying scale, or whether there are several scales, and whether items must be discarded or included. Subsequently, the reliability and validity of the resulting measurement scale need to be determined. Item analysis has become a fundamental part of the development of a variety of measurement instruments in psychology also (or perhaps especially) outside the field of cognitive abilities (Dawis, 1987).

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 the upper left cell in the scheme in Figure 2.1, 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 (i.e., 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 the lower left cell in the scheme in Figure 2.1, for an example of a typical single-peaked response function). 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. Hence his position can be computed by taking the mean (or median) of the locations of the preferred items (cf. Thurstone’s scaling approach, e.g., Thurstone, 1928).
2.1. Introduction

Another example of a bipolar scale is a scale that measures the attitude toward immigration, which ranges from a negative extreme (very much against immigration), via a neutral midpoint (neither against nor in favor of immigration), to a positive extreme (very much in favor of immigration). The fact that items reflecting a moderate tolerance toward immigration appeal to subjects in the middle of the scale, but neither appeal to subjects that are very much pro immigration, nor to subjects that are very contra immigration, makes the item response function of these items single-peaked.

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 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 (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, which are estimated from the data. IRT-based item analysis was originally exclusively developed for dominance (monotonic) items, but today, it also provides models for proximity (single-peaked) items; either nonparametric (e.g., MUDFOLD; Van Schuur, 1984), or parametric, (e.g., PARELLA; Hoijtingk, 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 first aim of this chapter is to contribute to item analysis by providing a CTT-based approach to the analysis of single-peaked items, that is, the psychometric evaluation of bipolar scales. Correspondence analysis (CA) (Benzécri, 1992; Greenacre, 1984), which is available in standard software packages such as SPSS and SAS, is proposed as a CTT-like counterpart of the unfolding IRT models. In Figure 2.1, a scheme is given that classifies the various approaches to item analysis. Note that the subject of the present chapter is in the lower right cell of this scheme. The reliability analysis for single-peaked items that is also mentioned in the lower right cell is addressed in Polak, De Rooij, and Heiser (2010a; see Chapter 4 of this thesis). The authors presented diagnostics for single-peakedness
of item responses based on ordered conditional means (OCM).

<table>
<thead>
<tr>
<th><strong>Unipolar scales:</strong></th>
<th><strong>Item response theory</strong></th>
<th><strong>Classical test theory</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dominance items</td>
<td>Monotonic models:</td>
<td>Principal Component Analysis or</td>
</tr>
<tr>
<td>1000</td>
<td>- parametric,</td>
<td>Factor Analysis</td>
</tr>
<tr>
<td>1100</td>
<td>e.g., Rasch, 3PLM, SGRM</td>
<td>Reliability analysis,</td>
</tr>
<tr>
<td>1110</td>
<td>- nonparametric,</td>
<td>e.g., Cronbach’s alpha, test-retest</td>
</tr>
<tr>
<td></td>
<td>e.g., Mokken</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Bipolar scales:</strong></th>
<th><strong>Unfolding IRT models:</strong></th>
<th><strong>Thurstone’s scaling</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-peaked items</td>
<td>- parametric,</td>
<td>Correspondence Analysis</td>
</tr>
<tr>
<td>1100</td>
<td>e.g., GGUM, HCM, MUM</td>
<td></td>
</tr>
<tr>
<td>0110</td>
<td>- nonparametric,</td>
<td></td>
</tr>
<tr>
<td>0011</td>
<td>e.g., MUDFOLD</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.1: Classification of scaling techniques for item response data, with CA as proposed CTT-like approach to the analysis of single-peaked items.

The second aim of the present chapter is to compare CA with IRT-based item analysis of single-peaked items. For this purpose we selected both a parametric and a nonparametric unfolding IRT model: the generalized graded unfolding model (GGUM; Roberts Donoghue, & Laughlin, 2000) and the multiple unidimensional unfolding model (MUDFOLD; Van Schuur, 1984), respectively. We chose GGUM and MUDFOLD, since both models have been well-developed and provide user-friendly, Windows-based software (i.e., resp., GGUM2004; Roberts, Fang, Ciu, & Wang, 2006, and MUDFOLD4.0; Van Schuur & Post, 1998).

CA is known to represent bipolar scales correctly (e.g., Heiser, 1981, 1987a, 1987b), unlike FA/PCA, which is only suited for the analysis of unipolar or cumulative scales (e.g., Van Schuur & Kiers, 1994; Polak, Heiser, & De Rooij, 2009). Heiser (1985) showed that for single-peaked response functions, even nonlinear principal component analysis does not lead to a correct representation of person and item locations. In Section 2.2.1 we will explain CA as a CTT-like approach to item analysis. Note that the topic of the current chapter is the analysis of one-dimensional proximity data (or single-peaked items), thus we focus on the bottom cells in the scheme in Figure 2.1. A comparison of CA with PCA is described in Polak et al. (2009; see Chapter 3 of this thesis).

Why should we be interested in CA as an extension of the CTT approach to item analysis of single-peaked items, 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 that 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 option is not yet available 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. We will explain CCA in Section 2.2.2. Since CCA has some very attractive features, especially for practical researchers, we will illustrate the use of CCA for real psychological data in a separate section (see Section 2.5).

The outline of this chapter is as follows, first, in Section 2.2, we present the theoretical background of the compared techniques, that is, CA as an approach to CTT-like item analysis (Section 2.2.1), CCA for item analysis using explanatory variables (Section 2.2.2), parametric unfolding IRT with GGUM (Section 2.2.3), and nonparametric unfolding IRT with MUDFOLD (Section 2.2.4). In the method section (Section 2.3) we explain the data characteristics of the real data sets, and present a rationale for the choice of characteristics of the simulated benchmark datasets. In the results section (Section 2.4) we firstly compare the real data
results of the three techniques, that is, we illustrate the application of CA as a CTT-like approach to item analysis for a real dataset, and subsequently we present the results of both GGUM and MUDFOLD for the same data (Section 2.4.1). Secondly, in Section 2.4.2 we compare results for the simulated benchmark data. Finally, we illustrate the use of CCA for data from the field of personality assessment in Section 2.5. In Section 2.6 we discuss the results.

2.2 Theory

In the subsections below we present the theoretical background of CA, CCA, and the unfolding IRT models GGUM and MUDFOLD.

2.2.1 CA as an Approach to CTT-like Item Analysis

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).

In PCA, item loadings are found by performing an eigenvector/eigenvalue decomposition on the inter-item correlation matrix; if necessary, person (subject) scores are then found in a second step by a regression procedure. In contrast, in CA, item locations and person (subject) locations are found by computing a singular value decomposition of a matrix $D$ with standardized deviations from independence. Computational details and the rationale of using CA for analyzing single-peaked item response data are presented in Appendix A.

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 environmental gradient (e.g., vegetation structure or pH-value)(cf. Ter Braak & Prentice, 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 the following we will translate the findings concerning CA from the field of ecology to the field of
2.2. Theory

psychology. We will use the term subjects instead of sites and items instead of species. 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 underlying measurement scale.

Note that, instead of some directly observable environmental variable, psychological research is often concerned with a latent variable (such as an attitude or a personality characteristic) for which a measurement scale is constructed. CA is well suited to estimate the subject and item location parameters of Gaussian models for the subject-to-item distance on such latent scale. Ter Braak (1985) showed that CA approximates the ML solution of the Gaussian ordination model for binary single-peaked data. Polak et al. (2009) have used results from Ter Braak (1985) and Ihm and Van Groenewoud (1984) to show the relation between the first CA component and the Gaussian ordination model, where the latter can be written as

\[
\pi_{ij} = \beta_j^* \exp \left( -\frac{1}{2} \left( \theta_i - \delta_j \right)^2 / \alpha_j^2 \right),
\]

(2.1)

where

- \( \pi_{ij} \) is the probability that subject \( i \) agrees with item \( j \),
- \( \theta_i \) is the location of subject \( i \) on the underlying scale,
- \( \delta_j \) is the location of item \( j \) on the underlying scale,
- \( \beta_j^* \) is the maximum of the curve for item \( j \), and
- \( \alpha_j \) is the discrimination parameter for item \( j \) with large values indicating poor item discrimination (i.e. relatively flat IRFs).

An example of an IRF conforming to (2.1) is given in Figure 2.2a. Note that the probability that a subject agrees with item \( j \) is maximal for subjects with the same location as item \( j \), and decreases for subjects that lie at a greater distance from item \( j \) in both directions of the scale.

Ter Braak (1985) discusses the conditions under which the CA approximation of (2.1) works best, namely equally spaced item locations, \( \delta_j \), which extend past all \( \theta_i \) on both ends of the scale. Furthermore, items should have equal or independent maxima \( \beta_j^* \), and equal discrimination parameters \( \alpha_j \); subject locations \( \theta_i \) should be evenly distributed over the whole range of \( \delta_j \), and closely spaced relative to \( \alpha_j \).

Simulation studies have shown that CA is reasonably robust when these conditions (Ter Braak, 1985) are not completely met. However, for unequal discrimination parameters the accuracy of the approximation diminishes. Furthermore,
2. Psychometric Evaluation of Bipolar Measurement Scales

Figure 2.2: IRFs for an item located on the midpoint of the scale defined by two different response models, the Gaussian ordination model (a) and the generalized graded unfolding model (b).

(a) Gaussian ordination model with parameters $\delta_j = 0, \beta_j^* = 1, \alpha_j = 1$.

(b) Generalized graded unfolding model with parameters $\delta_j = 0, \alpha_j = 1, \tau_{jm} = -0.5, 0.0, 0.5$.

Ter Braak and Prentice (2004) recommend to particularly check for subjects with an extreme response style (that is, subjects that agree or disagree with all items) and for items that are either deviant (i.e., items that are unlikely to have a regular single-peaked IRF) or “rare” (i.e., items that only very few respondents agree with).

To start with the latter, rare items are easily recognized in the CA solution, as they often appear as outlying points. That is, a point that has a very large contribution to one of the major dimensions (see Appendix A for a more detailed explanation of how this contribution is defined) and a high scale value on a major dimension, can be considered an outlier. Usually, these outlier points are discarded, since they offer little information about the individual differences among subjects.

Deviant items with an irregular IRF, and hence items that are not directly related to the underlying dimension will not appear as outliers in the solution, but, in contrast, will be located near the center (origin) of the solution. To distinguish these items from items with a “true” central location, one needs to check the assumption of single-peaked IRFs.

Ter Braak (1985; see also, Ter Braak & Prentice, 2004) recommends to check the requirements concerning the IRFs, by regressing the observed ratings on each item on the CA estimates of the subject locations (using non-linear regression). This is also a way to estimate the goodness of fit of the model. Polak et al. (2010a;
see Chapter 4 of this thesis) propose a model-free methodology to approximate the
IRFs of the items, which is less restrictive concerning the specific shape of the IRF
than the Gaussian regression function. It provides a nonparametric procedure to
determine item and model fit.

In practical applications of CA, the aim is to find an optimal graphical rep-
resentation of both subjects and items in as few dimensions as possible, which
usually results in a two-dimensional plot. Since the CA dimensions are nested,
like in PCA, the scores on each dimension do not change when a lower or higher
number of dimensions is selected.

In a two-dimensional CA solution each item and subject has two scores that
can be depicted as a point in a plane. Item points will be located relatively
close to each other, when the corresponding item responses are relatively similar
across subjects, and will be far apart when this is not the case. In the same way,
subjects with relatively similar score patterns (also referred to as profiles) will be
represented by subject points that are located relatively close to each other.

When data are strongly one-dimensional, a two-dimensional representation
will show what is often referred to as the arch-effect. In such case, the items and
subjects are ordered along an arch, but also along the first dimension, according
to their position on the scale (see for example, Hill & Gauch, 1980, or Greenacre,
1984 p. 227). In such case only the scores on the first dimension are used as
scale scores. The scores resulting from CA can be standardized in different ways
depending on the focus of the analysis. In the current chapter we performed so-
called row principal analyses (i.e. analysis of row profiles), meaning that the item
scores are standardized to have (weighted) mean 0 and (weighted) variance 1 and
each subject score is computed as the (weighted) average of his corresponding
item scores, with the ratings used as weights.

In short, CA is a dimension reduction technique that always results in a unique
solution for all items and subjects. Furthermore, CA allows the user to make a
two-dimensional graphical display of the solution, where distances between points
representing subjects and items are inversely related to their similarity. When
item responses conform to the Gaussian ordination model, the first CA dimension
approximates the ML solution of the parameters in the Gaussian ordination model
(Ter Braak, 1985).
2.2.2 Constrained Correspondence Analysis (CCA): Item Analysis using Explanatory Variables

In this section we give the theoretical background of CCA. It translates the theory presented in Ter Braak (1986) and Ter Braak and Verdonschot (1995) into the context of item analysis of (psychological or sociological) item response data.

For a set of items located on a standardized measurement scale, with location parameters $\delta_j$, for which

$$
\sum_{j=1}^{k} z_{ij} \delta_j = 0 \land \sum_{j=1}^{k} z_{ij}^2 \delta_j^2 = 1,
$$

where $z_{ij}$ is the observed response of subject $i$ ($i = 1, ..., n$) to item $j$ ($j = 1, ..., k$), and the symbol “+” indicates the sum over the omitted index.

The location of subject $i$ is defined as the weighted average ($\theta_i$) of the locations $\delta_j$ of the items that are endorsed by subject $i$, i.e.

$$
\theta_i = \sum_{j=1}^{k} \frac{z_{ij}}{\sum_i z_{ij}} \delta_j.
$$

When we choose a row principal normalization, the weighted variance of the subject locations $\theta_i$ ($i=1,...,n$) is defined by

$$
\lambda = \sum_{i=1}^{n} \frac{z_{i+}}{z_{++}} \theta_i^2.
$$

In CCA $\theta_i$ is constrained to be a linear combination of explanatory variables,

$$
\theta_i = \sum_{h=1}^{p} c_h x_{ih},
$$

with $x_{ih}$ the score of subject $i$ on explanatory variable $h$ ($h=1,...,p$), and $c_h$ as the optimal weight of explanatory variable $h$.

CCA chooses optimal weights $c_h$, i.e. weights that result in a measurement scale with item locations $\delta_j$ and subject locations $\theta_i$, for which the weighted variance defined by equation (2.4) is at its maximum.
2.2. Theory

To make the explanatory power of the explanatory variables comparable, each variable must first be standardized to mean 0 and variance 1. The relative importance of each variable for predicting the measurement scale can be inferred from the signs and the magnitudes of the canonical coefficients and the so-called intraset correlations, where the latter are defined as the correlations between each explanatory variable and the measurement scale. Note that when the explanatory variables are uncorrelated, the canonical coefficients and the intraset correlations are identical.

The subject locations and item locations are often represented graphically in a CCA diagram (usually a two-dimensional display). Because each subject lies at the weighted mean of the item locations (with the subject’s ratings used as weights), we can infer from the diagram the probability that a subject chooses each item: the smaller the distance between a subject’s location $\theta_i$ and an item location $\delta_j$, the higher the probability of endorsement. In the CCA diagram the explanatory variables $x_h$ are displayed as vectors from the origin of the plot, where the angle of the vector with each dimension $s$ is determined by the intraset correlation of the explanatory variable with dimension $s$. The orthogonal projection, of each item point $\delta_j$ on a vector representing an explanatory variable $x_h$ gives the weighted average $\bar{x}_{jh}$ of $x_h$, with the item scores of item $j$, $z_{ij}$, as weights:

$$\bar{x}_{jh} = \frac{\sum_{i=1}^{n} z_{ij} x_{ih}}{z_{+j}}. \tag{2.6}$$

It follows from equation (2.6) that the more association there is between high scores on item $j$ with high scores on explanatory variable $x_h$, the closer the projection of item point $\delta_j$ on the vector of $x_h$ is to the arrow head of this vector. The differences between the values of $\bar{x}_{jh}$ indicate differences between the item distributions with respect to the explanatory variables. An example of the interpretation of a CCA diagram is given in Section 2.5.

2.2.3 Parametric Unfolding IRT: the Generalized Graded Unfolding Model (GGUM)

The GGUM (Roberts & Laughlin, 1996; Roberts et al., 2000) is a parametric IRT model that incorporates features such as variable item discrimination and variable threshold parameters for the response categories. The model assumes the existence of a latent trait (i.e., unidimensionality), local independence, and symmetric (around the item location), bell-shaped IRFs.
2. Psychometric Evaluation of Bipolar Measurement Scales

The GGUM allows for binary or graded responses. One premise of the GGUM is that for each subject there are two subjective responses associated with each observed response. These subjective responses can be seen as two distinct reasons for a subject’s response. For instance, when a subject strongly disagrees with a certain item this could be for either of two reasons. If on the underlying (bipolar) continuum the item is located more to the right extreme than the subject, the subject disagrees from below the item. However, if the item is located more to the left extreme than the subject, the subject disagrees from above the item. The probability that a subject will respond using a particular answer category is defined as the sum of the probabilities associated with the two corresponding subjective responses.

Specifically, the model has the form:

\[
P(Z_{ij} = z | \theta_i) =
\frac{\exp\{\alpha_j[z(\theta_i - \delta_j) - \sum_{m=0}^{z} \tau_{jm}]\} + \exp\{\alpha_j[(S - z)(\theta_i - \delta_j) - \sum_{m=0}^{z} \tau_{jm}]\}}{\sum_{\omega=0}^{M} \left( \exp\{\alpha_j[\omega(\theta_i - \delta_j) - \sum_{m=0}^{\omega} \tau_{jm}]\} + \exp\{\alpha_j[(S - \omega)(\theta_i - \delta_j) - \sum_{m=0}^{\omega} \tau_{jm}]\} \right)},
\]

(2.7)

where

- \(Z_{ij}\) is a random variable indicating the observed response of subject \(i \ (i = 1, ..., n)\) to item \(j \ (j = 1, ..., k)\) with \(Z_{ij} = z\), and
- \(z = 0, 1, ..., M\), with \(z = 0\) indicating the strongest level of disagreement,
- \(z = M\) indicating the strongest level of agreement,
- \(S = 2M + 1\),
- \(\theta_i\) is the location of subject \(i\),
- \(\delta_j\) is the location of item \(j\) (on the same metric as \(\theta\)),
- \(\alpha_j\) is the discrimination of item \(j\), and
- \(\tau_{jm}\) is the relative location of response category \(m\) within item \(j\).

An example of an IRF conforming to equation (2.7), that is \(P(Z_{ij} = 1 | \theta_i)\) for a dichotomous item, is given in Figure 2.2b. The model parameters defined in equation (2.7) can be estimated with GGUM2004 (Roberts et al., 2006). Free copies of the program are readily available to readers at:

http://www.psychology.gatech.edu/unfolding/FreeSoftware.html

GGUM2004 parameter estimation works as follows. Item parameters are estimated using a marginal maximum likelihood (MML) approach (Bock & Lieber-
The algorithm is based on an expectation maximization (EM) strategy, which is used to solve the likelihood equations for the item parameters, $\delta_j$, $\alpha_j$ and $\tau_{jm}$. Subject parameter estimates are obtained by using an expected a posteriori procedure (EAP).

Several simulation studies have been published that evaluate the performance of the GGUM procedure in various conditions (Roberts et al., 2000, Roberts, Donoghue & Laughlin, 2002). Results indicate, that MML estimates are relatively insensitive to the choice of the prior distribution. Similarly, EAP estimates are also robust to the choice of the prior distribution, except in cases involving extreme $\theta_i$ values with correspondingly extreme response patterns. Furthermore, it was shown that accurate item parameter estimates could be obtained with a sample size of at least 750; accurate subject estimates were obtained with 15-20 items with 6 response categories per item.

A possible limitation of the simulation studies described above, is that items ($\delta_j$) were always located at equally distant positions on the latent continuum that always ranged from -2.0 to 2.0, regardless of the number of items studied. This equal-spacing strategy was based on a general principle in Thurstone’s (1928) attitude scale construction procedure in which items are explicitly chosen to represent the latent continuum in an approximately uniform fashion. In practice, it might also be relevant to test the recovery of parameters, when we have two clusters of items on both sides of the continuum, and thus with a gap in the middle. For instance, Roberts, Laughlin and Wedell (1999) explain the difference between the Thurstone and Likert approach, and show that, in practice, the Likert approach (Likert, 1932) results in selection of two clusters of items (contra-indicative and indicative) on both ends of the latent continuum, and thus leaves out more nuanced items in the center of the scale (see also Andrich, 1996).

### 2.2.4 Nonparametric Unfolding IRT: the Multiple Unidimensional Unfolding Model (MUDFOLD)

The second IRT model, which we discuss in more detail, is MUDFOLD (Van Schuur, 1984, Van Schuur & Post, 1998). MUDFOLD is a nonparametric IRT model that results in ordinal scale values for both items and subjects. This model assumes unidimensionality, local independence, and single-peaked IRFs, but in contrast to GGUM and several of its predecessors (e.g., Andrich and Luo, 1993; Hoijtink, 1991), the model does not assume a specific shape of the IRFs.

From a set of $k$ items MUDFOLD forms an optimal unfolding scale consisting
of a subset of \( k^* \leq k \) items. The procedure is based on the following rationale. If a set of items conforms perfectly to the unidimensional unfolding model, then for each subset of three items that are ordered according to their position along the unfolding scale, it must hold that if a subject agrees with the two outer items, he also agrees with the intermediate item. Hence, for binary items (where 1 indicates agreement), each score pattern 1, 0, 1 on three subsequent items is regarded a violation of the unfolding model.

In searching for an optimal unfolding scale, MUDFOLD calculates the frequency of the so-called observed error patterns \( O \) (i.e. 1, 0, 1), for each subset of three items, in each of its three possible different orders. These frequencies are compared to the frequencies of expected error patterns \( E \), under statistical independence.

A final unfolding scale is formed by maximizing the \( H \)-coefficient (adapted for unfolding items by Van Schuur (1984)), where \( H = 1 - O/E \). Note that \( H = 1 \) for a perfect unfolding scale, and \( H = 0 \) for a set of items which are statistically independent. \( H \) can be calculated for the entire scale as well as for each individual item. In the first step of an iterative procedure the \( H \)-coefficient is used to find the best three item scale that is selected as the elementary scale. Iteratively the item that leads to the highest \( H \)-value for the scale as a whole is selected and added to the scale. Items are added to the scale under the condition that the \( H \)-coefficient of both the scale as a whole, and of the individual item has a higher value than a user-specified value between 0 and 1, where 0.30 is used as the default value for acceptable fit.

Note that MUDFOLD results in a set of ordered items, and not item locations, that have an acceptable fit under the MUDFOLD model assumptions. Subject scale values are computed after the optimal unfolding scale, consisting of a subset of \( k^* \leq k \) items, has been selected. The procedure for determining subject scale values is based on the procedure used in Mokken’s nonparametric cumulative scaling (see Mokken, 1971). A detailed explanation of the MUDFOLD subject parameter estimation can be found in Van Schuur and Post (1998, pp. 31-34).

### 2.3 Method

The aims of the present research are, first, to explain and evaluate the use of CA as a an approach to CTT-like item analysis of single-peaked items, that is, as a method for the psychometric evaluation of bipolar scales. For this purpose we use a real dataset. The second objective is to compare the performance of CA to the
2.3. Method

performance of unfolding IRT methods (GGUM and MUDFOLD) in terms of the recovery of the “true” scale (location) values for both subjects and items. For this purpose we compare the results of analysis of the real dataset obtained with the various methods. Additionally, we generated and analyzed benchmark datasets for a more systematic comparison. In the subsections below we will explain the respective methods for realizing both objectives.

The attractive features of CCA are discussed separately in Section 2.5. For this purpose we analyze a real dataset from the field of clinical psychology. In Section 2.5 this dataset is presented, as well as the results of the CCA.

2.3.1 Real Data: Thurstone’s Capital Punishment Scale

The data were graded responses to Thurstone’s (1932) attitude toward capital punishment scale, which Roberts and Laughlin (1996) obtained from 245 American undergraduates on the 24 items of the scale. The scale consists of statements that are listed in Table 2.1, which is presented in Section 2.4.1. The response format is a six-point rating scale (with response categories 0 = strongly disagree to 5 = strongly agree), varying from strongly against the death penalty to strongly in favor of it.

We analyzed this dataset with CA, GGUM, and MUDFOLD. In particular, we were interested in the subject and item locations, as well as the model fit.

2.3.2 Simulated Benchmark Data

The main focus of this part of the study was to investigate whether CA would also be applicable for evaluating scales with items grouped into two clusters, one on each end of the scale, and thus with a gap in the middle. This is a departure from the ideal conditions for CA as were derived by Ter Braak (1985, as explained in Section 2.2.1 of the current chapter). This situation is of particular interest, since it provides a test whether CA and unfolding IRT methods can be used to analyze scales that are bipolar in nature, but that are constructed according to the Likert-approach (which is very common nowadays).

As explained in Section 2.2.3, the Likert approach results in selection of two clusters of items (contra-indicative and indicative) on both ends of the latent continuum, and requires the reverse-scoring of either of both subsets of items. A typical advantage of CA and unfolding IRT methods is that they do not require this reverse scoring procedure, since they uncover bipolar scales, with the two clusters of items located on both ends of the scale. It is of particular interest whether
the techniques are capable of recovering the true subject locations, especially for the subjects with positions in the middle of the scale, in the absence of items at that part of the scale. Since it is known that both CA and unfolding IRT work well for scales with equidistant items, we will use datasets with equidistant items as reference conditions.

We generated eight benchmark datasets to compare the performance of CA to the performance of the unfolding IRT methods GGUM and MUDFOLD. We first discuss the design of the data generation procedure, and subsequently the procedure itself.

We crossed the factor number of items (10; 20) with the factor item spacing (equidistant; grouped into two clusters, one on each end of the scale, and thus with a gap in the middle). For both the 10-item scales, and the 20-item scales we generated four datasets; one with evenly spaced items as a reference condition, and three datasets with a gap in the middle of the item locations, where we varied the amount of error in the responses.

In each condition of our design, we generated 100 random datasets by sampling from a multinomial distribution with 5 response categories (0-4) with probabilities computed with the GGUM (see equation (2.7)). GGUM parameter values were set as follows. We generated responses to either 10 or 20 items that were located either on both ends of the scale, and not in the center (the left clusters equidistant from -2 to -1, the right cluster equidistant from 1 to 2), or evenly spaced over the scale (from -2 to 2). We set $\alpha_j = 1$ for all items and the interthreshold distance to .5 for all items, to produce typical GGUM IRFs (cf. Figure 2.2b). The subject locations were sampled from a $N(0,1)$ distribution, with $N = 300$ in all conditions.

Note that as a consequence of the random data generation procedure the simulated datasets vary in the way they resemble the expected response pattern following from the GGUM model. That is, no dataset will show a perfect deterministic structure, but some datasets will resemble this “ideal” structure more than others. For both the 10-item scales, and the 20-item scales with the clustered items, our aim was to select three data sets from the 100 we generated that represented, respectively a “weak”, a “moderate”, or a “strong” resemblance to a perfect deterministic structure. For both the 10-item scales, and the 20-item scales with evenly spaced items, our aim was to select an “moderate” data set from the 100 we generated, which we could use as a reference.

In order to select the intended datasets, we first performed CA on all 100 datasets in each condition. Subsequently, in each condition, we computed the correlations between the true subject locations and the CA estimated subject
locations for all datasets. Our assumption is, that the higher this correlation, the stronger resemblance the data will show to a perfect deterministic structure (Heiser, 1981, p. 118-129, showed that CA can perfectly recover a dataset with a deterministic unfolding structure).

For both the 10-item scale, and the 20-item scale with the clustered items, we selected from the distribution of the 100 correlations, three datasets corresponding to, respectively, the first, the second (median) and the third quartile. We consider these three datasets to represent, respectively, a relatively weak, a moderate, and a strong deterministic structure.

For both the 10-item scale, and the 20-item scale with the equidistant items, we selected from the distribution of the 100 correlations, the datasets corresponding to the second quartile (median). We consider such a dataset to represent a moderate deterministic structure.

Of each CA solution, the first dimension estimates were taken to compare the performance of CA and both unfolding IRT models. Since the model-generated data are one-dimensional, the major target is that the first dimension of each solution shows the true order of items and subjects. The quality of recovery of the subject locations and item locations is expressed in terms of correlations (Pearson $r$ and Spearman rank correlation $r_s$).

### 2.4 Results

In the sections below we discuss the analysis results for Thurstone’s attitude scale data (Section 2.4.1), and the benchmark datasets that were simulated based on the GGUM model (Section 2.4.2).

#### 2.4.1 Real Data: Thurstone’s Capital Punishment Scale

In this section we analyze responses obtained by Roberts and Laughlin (1996) to Thurstone’s (1932) attitude toward capital punishment scale, which was introduced above (see Section 2.3.1). In the following sections we will first present the results of each analysis, that is, CA, GGUM, and MUDFOLD separately, and subsequently, we will compare the results of the three analyses.

**CA on Thurstone’s Capital Punishment Scale**

CA was performed on the total set of 24 items. The estimated item locations, $\hat{\delta}_j^1$, resulting from CA, are given in Table 2.1. The two-dimensional display of both
item and person location estimates, resulting from the CA, is presented in Figure 2.3.

Three items were discarded, after they were identified as deviant items based on the CA solution. The data-analytic strategy we used was a step-by-step-procedure, in which we first checked the two-dimensional CA solution for outliers, which are remote points, with location values on either of both dimensions exceeding ±3. This criterion was used since the item scores are standard scores and standard scores exceeding ±3 are commonly regarded as extreme. In the second step we inspected the frequency distribution of the most remote outlier. When the item showed to be a “rare” item, meaning that only few people expressed strong agreement with it, it was deleted. In the third step we performed the CA for the remaining items and we repeated this stepwise process until no further outliers were present.

This procedure led to the removal of items 24 (“Every criminal should be executed”) with coordinate (0.95, 8.18) and 12 (“I think the return of the whipping post would be more effective than capital punishment”) with coordinate (0.29, 5.62), which showed, respectively, 2.4% and 3.3% strongly agree responses. Compared to the average percentage of strongly agree responses over all items of 11.9% these two items can clearly be considered “rare”. The relative unpopularity of these two items can be explained by their relatively extreme wording.

One point in the resulting solution, corresponding to item 13 (“It doesn’t matter to me whether we have capital punishment or not”), was located near the origin of the solution (coordinates, -0.15, 0.62). This item was also relatively unpopular with 2.4% strongly agree responses in the current sample. We decided to consider item 13 as deviant as well, and to remove it from the solution, because it provided almost no information about the differences among subjects (see also, Section 2.2.1).

Note that the diagram (which uses only two dimensions) gives an approximation of the original data. The quality of this approximation can be derived from the eigenvalues associated with each dimension. The total inertia (or variance in the data) is 0.547. The amount of variation in the data that can be represented by each dimension is expressed by the eigenvalue of each dimension. The eigenvalues sum to the total inertia. Eigenvalues of the first two axes are 0.228 (42%) and 0.036 (6%), respectively. The diagram thus represents 48% of the total variance in the observed scores, with a clearly dominant first dimension.

An important result of the CA is the arch pattern that is visible in the item points (indicated with dots) in Figure 2.3. In Section 2.2.1 it was explained, that
Table 2.1: Statements of the Capital Punishment Scale (Thurstone, 1932) with their original Thurstone scale values, $T_j$, CA location estimates, $\hat{\delta}_1^j$, GGUM location estimates, $\hat{\delta}_2^j$, and the MUDFOLD estimated rank numbers $\hat{o}_j$.

<table>
<thead>
<tr>
<th>Statement</th>
<th>THUR</th>
<th>CA</th>
<th>GGUM</th>
<th>MUDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Capital punishment is absolutely never justified.</td>
<td>0.0</td>
<td>-1.39</td>
<td>-2.09</td>
<td>2</td>
</tr>
<tr>
<td>2. I do not believe in capital punishment under any circumstances.</td>
<td>0.1</td>
<td>-2.28</td>
<td>-2.45</td>
<td>1</td>
</tr>
<tr>
<td>3. Capital punishment is the most hideous practice of our time.</td>
<td>0.6</td>
<td>-1.22</td>
<td>-1.90</td>
<td>7</td>
</tr>
<tr>
<td>4. Execution of criminals is a disgrace to civilized society.</td>
<td>0.9</td>
<td>-1.33</td>
<td>-1.91</td>
<td>5</td>
</tr>
<tr>
<td>5. We can’t call ourselves civilized as long as we have capital punishment.</td>
<td>1.5</td>
<td>-1.59</td>
<td>-2.00</td>
<td>3</td>
</tr>
<tr>
<td>6. The state cannot teach the sacredness of human life by destroying it.</td>
<td>2.0</td>
<td>-0.86</td>
<td>-1.44</td>
<td>11</td>
</tr>
<tr>
<td>7. Capital punishment cannot be regarded as a sane method of dealing with crime.</td>
<td>2.4</td>
<td>-1.17</td>
<td>-1.77</td>
<td>8</td>
</tr>
<tr>
<td>8. Capital punishment has never been effective in preventing crime.</td>
<td>2.7</td>
<td>-0.64</td>
<td>-1.63</td>
<td>9</td>
</tr>
<tr>
<td>9. Capital punishment is not necessary in modern civilization.</td>
<td>3.0</td>
<td>-1.48</td>
<td>-2.03</td>
<td>4</td>
</tr>
<tr>
<td>10. Life imprisonment is more effective than capital punishment.</td>
<td>3.4</td>
<td>-1.16</td>
<td>-1.79</td>
<td>6</td>
</tr>
<tr>
<td>11. I don’t believe in capital punishment but I’m not sure it isn’t necessary.</td>
<td>3.4</td>
<td>-0.65</td>
<td>-1.63</td>
<td>10</td>
</tr>
<tr>
<td>12. I think the return of the whipping post would be more effective than capital punishment.</td>
<td>3.9</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>13. I don’t matter to me whether we have capital punishment or not.</td>
<td>5.5</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>14. I do not believe in capital punishment but it is not practically advisable to abolish it.</td>
<td>5.8</td>
<td>-0.29</td>
<td>-1.43</td>
<td>12</td>
</tr>
<tr>
<td>15. I think capital punishment is necessary but I wish it were not.</td>
<td>6.2</td>
<td>0.59</td>
<td>.59</td>
<td>15</td>
</tr>
<tr>
<td>16. Capital punishment is wrong but is necessary in our imperfect civilization.</td>
<td>6.2</td>
<td>0.53</td>
<td>.53</td>
<td>14</td>
</tr>
<tr>
<td>17. Capital punishment may be wrong but it is the best preventative to crime</td>
<td>7.2</td>
<td>0.79</td>
<td>1.28</td>
<td>.</td>
</tr>
<tr>
<td>18. Capital punishment is justified only for premeditated murder.</td>
<td>7.9</td>
<td>0.53</td>
<td>1.29</td>
<td>13</td>
</tr>
<tr>
<td>19. We must have capital punishment for some crimes.</td>
<td>8.5</td>
<td>0.80</td>
<td>.57</td>
<td>16</td>
</tr>
<tr>
<td>20. Capital punishment should be used more often than it is.</td>
<td>9.1</td>
<td>1.15</td>
<td>1.11</td>
<td>19</td>
</tr>
<tr>
<td>21. Capital punishment gives the criminal what he deserves.</td>
<td>9.4</td>
<td>0.90</td>
<td>.96</td>
<td>17</td>
</tr>
<tr>
<td>22. Capital punishment is just and necessary.</td>
<td>9.6</td>
<td>0.98</td>
<td>.90</td>
<td>18</td>
</tr>
<tr>
<td>23. Any person who commits murder should pay with his own life.</td>
<td>10.4</td>
<td>1.00</td>
<td>1.44</td>
<td>20</td>
</tr>
<tr>
<td>24. Every criminal should be executed.</td>
<td>11.0</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

the arch is indicative of a dominant first dimension in terms of explained variation. These data are meant to be one-dimensional, whereas the corresponding latent scale is meant to be bipolar, ranging from strongly against the death penalty to strongly in favor of it. The obtained solution may be interpreted as a validation of this presumption. The fact that the arch-effect is only mildly apparent for the subject points (indicated with the red stars), is a consequence of the current (row principal) normalization, where the subject points are found as the weighted average of the item points. As explained in Section 2.2.1, when data are considered to be one-dimensional, the item scores on the first CA dimension are used as
Figure 2.3: CA diagram for the capital punishment scale, where item locations are depicted with a dot and the item number. The item labels are given in Table 2.1. Subject locations are depicted with a star.

location estimates on the latent scale.

In Figure 2.3 we see item 2 (I“ do not believe in capital punishment under any circumstances”) as the most extreme, negative statement, and item 20 (“Capital punishment should be used more often than it is”) as the most extreme, positive statement. Near the midpoint of the first dimension we find statements 11 and 14 that both reflect more nuanced opinions concerning capital punishment (see Table 2.1). The statements do not seem evenly distributed across the dimension, but rather seem divided in three clusters.
First, a cluster with negative statements on the left hand side of the first dimension (i.e., statements 1 to 10 that all express a disapproval of capital punishment); second, a cluster with two statements that express both disapproval and approval of capital punishment (i.e., statements 11 and 14); and third a cluster with positive statements on the right hand side of the first dimension (i.e., statements 15 to 23 that both express an approval of capital punishment).

Furthermore, Figure 2.3 shows that items 15 and 16, which both express reservations concerning capital punishment, are found more toward the midpoint cluster. Item 16 is closest to the midpoint, which can be understood from its content, since it seems to express the doubts concerning capital punishment more strongly.

Finally, Figure 2.3 shows that the cluster with negative statements has a larger variation in item scale values, than the cluster with positive statements. An explanation for this difference could be that in the current sample of American undergraduates more subjects endorse the positive items than the negative items (59 % of the subjects lies on the right hand side of the origin). In CA solutions, items which are both extreme and less popular tend to lie more in the periphery of the solution.

In Figure 2.3 it can further be seen that subject points are spread out along the first dimension, reflecting individual differences in attitude concerning capital punishment. There were no outlying subjects.

**Unfolding IRT on Thurstone’s Capital Punishment Scale**

In this section we will present the results of both parametric and nonparametric unfolding IRT analysis on the capital punishment data. We will start with the former by discussing results from GGUM analysis of the same data. Second we will present the results of a MUDFOLD analysis on the capital punishment data.

**GGUM analysis of the capital punishment data.** Roberts and Laughlin (1996) analyzed this dataset with GUM (a predecessor of GGUM2004), where they first eliminated items that were unlikely to conform to the unidimensionality assumption, based on PCA, leaving 17 items for the analysis. In short, the authors described the following procedure for item selection and model fit evaluation.

After a first analysis with GUM another 5 items were eliminated based on Wright and Masters’ (1982) item-infit $t$-statistic and an item content evaluation. This approach resulted in a scale consisting of 12 items, where the scale fit was judged as reasonably well, based on a visual inspection of the model fit plot and
the squared correlation \((r^2 = .987)\) between the average observed and expected values in 70 \((\theta - \delta)\) fitgroups. These fitgroups were formed by the program based on the signed difference between each subject’s and each item’s location. In each fitgroup, the average observed and average expected item response were calculated for all subject/item pairs. The average \(\theta - \delta\) value was also calculated. In the model fit plot (cf. Figure 8 in Roberts & Laughlin (1996, p. 251)), the average observed and expected item responses are plotted against the average \(\theta - \delta\) value.

In the current chapter we re-analyzed the capital punishment dataset with GGUM2004, but for reasons of comparison, with the same selection of items as was found in the final CA solution. Hence, items 24, 12, and 13, which were deviant items in the CA solution, were removed from the scale. The 21 estimated item locations, \(\hat{\delta}_j\), resulting from GGUM2004, are given in Table 2.1. The analysis converged without problems and the global model fit seems satisfactory (see Figure 2.4 for the GGUM model fit plot).

Although in the current version of the GGUM2004 software, no quantification of model fit is given in terms of the squared correlation, it is clear in Figure 2.4 that the observed scores are strongly in agreement with the expected scores based on the GGUM model in the 70 fitgroups. We compared Figure 2.4, with Figure 8 in Roberts and Laughlin (1996; p. 251), where the first indicates the GGUM model fit based on the current (CA-based) selection of 21 items, and the latter on a selection of 12 items (based on PCA and GGUM item fit statistics). Both figures indicate a satisfactory model fit, although Figure 2.4 shows a relatively better fit around the peak of the function. This means that the \textit{strongly agree} responses are more accurately described by the model estimates based on the current selection of 21 items. Roberts and Laughlin reckoned that, for their 12-item-scale, the estimated item locations showed a gap in the middle of the scale (between the locations -.60 and 2.05), where unfortunately the majority of the subjects was located. In the current analysis, for the 21-item-scale, the gap is also apparent (between the locations -1.43 and .59), but it is less wide (cf. the item location estimates in Table 2.1).

We have to make one reservation concerning the fit of GGUM on the 21-item scale. Although all localized item \(t\)-statistics were below the recommended cut-off value of 2.576, the item fit plot of item 14 showed a relative large amount of misfit with an irregular estimated IRF. In a previous study by Polak et al. (2010a) this item was also identified as deviant. In the CA solution (see Figure 2.3), this item has a high score on the second dimension, but was not identified as deviant, given the expected arch pattern.
2.4. Results

Figure 2.4: GGUM2004 global model fit plot for the 21-item capital punishment scale: average observed item responses (squares) and average expected item responses (solid line) as a function of $\theta_i-\delta_j$.

*MUDFOLD analysis of the capital punishment data.* For the comparison in the current chapter we analyzed the capital punishment dataset with MUDFOLD for the same selection of items as was used above. The estimated item rank numbers, $\hat{o}_j$, resulting from MUDFOLD, are given in Table 2.1.

The MUDFOLD analysis of the 21-item scale resulted in a satisfactory scale fit, with $H = .50$, which is above the recommended value of .30. Item 17 was discarded by the program, based on the criterion that for all combinations of three subsequent items the $H$-coefficient (see Section 2.2.4) must be positive. The MUDFOLD manual warns that this criterion might be too strict (Van Schuur & Post, 1998, p. 58).

Neither CA, nor GGUM showed substantial misfit for item 17. The resulting item order can be found in Table 2.1. Since MUDFOLD results in ordinal scale
values it is not possible to comment on the spacing of the items. We see that the MUDFOLD scale selects the same items as extremes (i.e., resp. items 2 and 23) as GGUM, whereas CA picked items 2 and 20 as the left and right extreme (see Table 2.1). However, Table 2.1 shows only a minor difference in the CA estimates for item 20 and 23.

Comparison of Results for Thurstone's Capital Punishment Scale

In the following we will compare the three data analysis techniques with respect to the ordering of the estimated item locations. Spearman rank correlations among the various estimates, including the original Thurstone scale values are reported in Table 2.2. To complete the comparison, we included not only the parameter estimates for the selection of 21 items reported in Table 2.1, but also for the selection of the 12 items, which was reported in Roberts and Laughlin (1996). CA on the subset of 12 items (see $\hat{\delta}_{1-12}$ in Table 2.2) resulted in a 50 % variance accounted for by the first dimension. MUDFOLD on the subset of 12 items (see $\hat{o}_{1-12}$ in Table 2.2) resulted in $H = .55$ for the scale. Correlations that were not of direct interest were omitted from the table.

Specifically, we will look at four different aspects. First, we will discuss how well each analysis corresponds with the original Thurstone order (see the correlations with subscript $a$ in Table 2.2). Second, we compare the three analysis techniques (CA, GGUM, and MUDFOLD) with respect to the estimated item ordering for the 21-item scale (see the correlations with subscript $b$ in Table 2.2). Third, we compare the three analysis techniques with respect to the estimated item ordering for the 12-item scale (see the correlations with subscript $c$ in Table 2.2). Fourth, we will evaluate the stability of each analysis technique by comparing the item ordering in the subset of 12 items with the item ordering of the same set items, but than within the subset of 21 items (see the correlations with subscript $d$ in Table 2.2). Similarity of both orderings indicates stability of the estimates of a particular technique.

The first row in Table 2.2 (correlations with subscript $a$) shows that, for both the 21-item, and the 12-item scale, the CA item order most strongly resembles the original item order based on the Thurstone scale values. In particular for the cluster of items with a negative location in the CA and GGUM solutions, the CA estimates have a wider range and a stronger resemblance to the original Thurstone order (cf. Table 2.1).

The correlations with subscripts $b$ and $c$ in Table 2.2 show that for both the
Table 2.2: Spearman rank correlations between original Thurstone scale values, $T_j$, CA item location estimates, $\hat{\delta}^1$, GGUM item location estimates, $\hat{\delta}^2$, and the MUDFOLD estimated rank numbers $\hat{o}$, for both the 21-item and the 12-item selection of the capital punishment scale.

<table>
<thead>
<tr>
<th></th>
<th>CA $\hat{\delta}^{1-21}$</th>
<th>GGUM $\hat{\delta}^{2-21}$</th>
<th>MUDF $\hat{\delta}^{2-21}$</th>
<th>CA $\hat{\delta}^{1-12}$</th>
<th>GGUM $\hat{\delta}^{2-12}$</th>
<th>MUDF $\hat{o}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>THUR</td>
<td>$T_j$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>$\hat{\delta}^{1-21}$</td>
<td>$.94_a$</td>
<td>$.87_a$</td>
<td>$.92_a$</td>
<td>$.95_a$</td>
<td>$.90_a$</td>
</tr>
<tr>
<td>GGUM</td>
<td>$\hat{\delta}^{2-21}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$.98_d$</td>
</tr>
<tr>
<td>MUDF</td>
<td>$\hat{o}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$.97_d$</td>
</tr>
<tr>
<td>CA</td>
<td>$\hat{\delta}^{1-12}$</td>
<td></td>
<td></td>
<td>$.94_b$</td>
<td>$.99_b$</td>
<td>$.100_d$</td>
</tr>
<tr>
<td>GGUM</td>
<td>$\hat{\delta}^{2-12}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$.97_c$</td>
</tr>
</tbody>
</table>

For the eight benchmark datasets results are presented in Table 2.3. Table 2.3 shows the correlations (Pearson $r$ and Spearman rank correlation $r_s$) between the true and estimated location parameters of, respectively, CA, GGUM, and MUDFOLD. In general, higher correlations between the true and estimated location parameters, indicate a better recovery. In particular, Spearman rank correlation $r_s$ can be considered as an index for the quality of recovery of the ordering of the true locations, and Pearson $r$ can be considered as an index for the quality of recovery of the variance (or spacing) of the true locations as well as of the ordering of the true locations. First we will explain the results or the subject locations, and second for the item locations.
Table 2.3: Parameter recovery for CA, GGUM and MUDFOLD; Pearson ($r$) and Spearman ($r_s$) correlations between true and estimated parameter values for items and subjects, for either a 10- or 20-item scale.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Item spacing</th>
<th>CA</th>
<th>GGUM</th>
<th>MUDFOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>items</td>
<td>subjects</td>
<td>items</td>
</tr>
<tr>
<td>10</td>
<td>Gap (1)</td>
<td>$r$ .998</td>
<td>.908</td>
<td>.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gap (2)</td>
<td>$r$ .998</td>
<td>.925</td>
<td>.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gap (3)</td>
<td>$r$ .999</td>
<td>.926</td>
<td>.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evenly (2)</td>
<td>$r$ .997</td>
<td>.934</td>
<td>.999</td>
</tr>
<tr>
<td>20</td>
<td>Gap (1)</td>
<td>$r$ .998</td>
<td>.946</td>
<td>.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gap (2)</td>
<td>$r$ .982</td>
<td>.970</td>
<td>.997</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gap (3)</td>
<td>$r$ .999</td>
<td>.948</td>
<td>.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evenly (2)</td>
<td>$r$ .998</td>
<td>.962</td>
<td>.999</td>
</tr>
</tbody>
</table>

Note. Items were either distributed in two equal sized clusters at both extreme ends of the scale (Gap), or distributed evenly along the scale (Evenly); and the data showed either (1) weak, (2) moderate, or (3) strong resemblance to a prefect deterministic structure.

Recovery of the Subject Locations. We will first interpret the results for the subject location estimates in terms of Spearman rank correlation $r_s$. This correlation can be considered as an index for the quality of recovery of the ordering of true subject locations. Table 2.3 makes clear that overall, CA and GGUM perform better than MUDFOLD with respect to recovering the correct ordering of the subjects along the scale. CA and GGUM show highly similar values for $r_s$ in each condition, although GGUM has slightly higher values in all cases (but most differences only occur with respect to the third decimal).

Furthermore, it can be seen that, as expected, the values of $r_s$ for all techniques in all conditions are higher for the longer scales. This means that the true ordering of subjects on a scale can be determined more accurately on the basis of 20 items, than on the basis of 10 items. The strength of this improvement is approximately the same for all techniques.

Additionally, we see as expected, that within each condition of scale length, the
values of $r_s$ increase with the strength of the deterministic structure. The pattern of improvement of $r_s$ is very similar for CA and GGUM. For MUDFOLD the relation between scale length and quality of recovery is less consistent. Apparently in the current study, CA and GGUM result in relatively more stable estimates of the subject ordering than MUDFOLD.

Next, we compare the values of $r_s$ for the datasets with unevenly spaced items (Gap) with the values of $r_s$ for the dataset with evenly spaced items (Evenly), within each condition of scale length. For the 10-item scale the values of $r_s$ are the highest for the equidistant dataset for all techniques. In contrast, for the 20-item scale the values of $r_s$ are at its highest for the unevenly spaced (Gap) dataset with a strong deterministic structure for both CA and GGUM. Also, for the 20-item scale, the values of $r_s$ for both the unevenly spaced, and the evenly spaced dataset with a moderate deterministic structure are almost equal for CA and GGUM. For MUDFOLD, the recovery of the ordering of true subject locations is most accurate when the items are evenly spaced regardless of scale length.

These results indicate that, in general, the quality of recovery of the ordering of true subject locations improves when the items are evenly spaced, but a gap in the item locations can be overcome by both CA and GGUM as long as there are enough items in the scale. MUDFOLD is more sensitive to departures from equidistance of the items, that is, a gap in the item locations.

When we inspect the results for the subject location estimates in terms of Pearson $r$, that is, only comparing CA with GGUM, we see a similar pattern of results as described above. However, some differences are apparent. First, the difference between the values of $r$ for CA and for GGUM within each condition are larger than for the values of $r_s$ (the differences are now apparent in the second decimal); in all cases $r$ values for GGUM are slightly higher than for CA. Second, these differences are at its minimum for the 20-item scales with equidistant spacing of the item locations. To understand these differences we depicted the relation between the true subject locations and the respective estimates. Figure 2.5 shows the relation between the true and estimates subject locations for the 20-item scales with evenly spaced items.

In Figure 2.5, it can be seen that for this benchmark dataset all techniques result in a good recovery of the subject locations, with a strong, linear relation between the true location values and the respective estimates. Comparing Figures 2.5a, 2.5b and 2.5c it can be seen that this relation is slightly stronger for GGUM than for the other techniques, with, in particular, less variability in the estimates of the extreme subject locations. Note that even MUDFOLD shows a relatively
strong linear relation between the true and estimated subject locations, despite of
the ordinal nature of its solution.

Next, in Figure 2.6, we depicted the relation between the true subject locations
and the respective estimates for the condition with 20 unevenly spaced items (gap,
weak structure).

First, Figure 2.6 shows that for this benchmark dataset, GGUM is the only
technique that results in subject locations estimates that are linearly related to
the true subject locations. Both CA and MUDFOLD show an S-shaped relation
between both sets of subject location values. However, judging from the values
of $r_s$ for CA the subject ordering is approximately equal to that of GGUM. For
MUDFOLD, however, the ordering shows more discordance. Judging from Figure
2.4. Results

Figure 2.6: Recovery of subject location values by CA, GGUM and MUDFOLD for the dataset (weak structure) with 20 unevenly spaced items (gap).

2.6c, most errors are found at the right extreme of the scale. The S-shaped relation between the estimated and true subject locations is a known phenomenon in CA. It is often referred to as the edge-effect (Gauch, 1982), which can be recognized by a considerably smaller variance of subject locations at the beginning and the end of the first dimension, than in the middle. Apparently, the edge-effect does not occur in the equidistant condition, but rather in the conditions with a gap in the item locations. We will discuss this phenomenon in more detail in Section 2.6.

Recovery of the Item Locations. In Table 2.3 it can be seen that the recovery of the item locations is almost perfect in all conditions for all techniques. Apparently, the sample size of 300, used in the current study is sufficient to find accurate item location estimates under the various conditions studied. One thing should be
noted, however: to find a proper solution in all conditions with unevenly spaced items, MUDFOLD required an appropriate starting set of items. That is, without a minimal set of three items in their correct order, MUDFOLD did not converge. But provided with any such starting set, MUDFOLD always resulted in the correct item order.

To sum up, the analyses of the benchmark datasets show that, in the studied conditions, all techniques recover the true item locations (or their ordering in the case of MUDFOLD) very well. With respect to the subject locations, the techniques show some important differences. First, overall, CA and GGUM perform better than MUDFOLD with respect to recovering the correct ordering of subjects along the scale. Second, results show that the true ordering of subjects on the scale is determined more accurately in the conditions with 20 items, than in those with 10 items. Third, although CA performed approximately equally well as GGUM in retrieving the correct subject ordering, the Pearson correlations between the true and estimated subject locations are higher for GGUM than for CA. This may be explained by the edge-effect in CA, which was apparent in the conditions with unevenly spaced items, but not in the conditions with equidistant items.

2.5 CCA for Item Analysis using Explanatory Variables: the Developmental Profile

Here we analyze a subset \((N = 105)\) of the Developmental Profile (DP; Abraham, 1993, 2005; Abraham et al., 2001) data, including three (standardized) explanatory variables, namely, (1) the Symptom Check List 90-item version (SCL-90; Arrindell & Ettema, 1975) that measures psychological distress, (2) Age and (3) the dummy variable “Healthy” \((0 = \text{patient}, 1 = \text{healthy control})\). See Polak, Van, Overeem-Seldenrijk, Heiser, & Abraham (2010), or Chapter 5 of this thesis, for the theoretical framework of this measurement instrument, and the analysis of the complete dataset \((N = 736)\).

The DP is an instrument for personality assessment consisting of nine subscales, referred to as developmental levels, each consisting of nine items, referred to as developmental lines. The items reflect personality characteristics and are scored by a trained mental health professional following a semi-structured interview with a client. A four-point scale is used to indicate the degree to which each personality characteristic is present \((0 = \text{not present} \text{ to } 3 = \text{very clearly present})\). The developmental profile of an individual is defined as the set of total scores on
the nine developmental levels. The total scores were recoded as follows: 0 (0), 1(1-3), 2(4-6), 3(7-9), 4(9). From a developmental perspective, it is presumed, that an individual’s profile (i.e., the pattern of total scores) shows a peak at the level that characterizes the person’s current level of functioning best. As an individual develops, for instance in the course of therapy, his peak will shift up the hierarchy of the DP. This shift indicates that the person learns to replace maladaptive behavior with more adaptive behavior.

The current sample consisted of 105 individuals, of whom 82 individuals (78%) were classified as psychiatric inpatients and 23 individuals (22%) as healthy controls.

Previous studies have shown that the nine developmental levels (i.e., the subscales of the DP) are ordered on one underlying bipolar dimension ranging from maladaptive to adaptive psychosocial functioning (Polak, Overeem-Seldenrijk, & Abraham, 2005; Polak et al., 2010). In the current chapter the aim is to validate the interpretation of this dimension as ordering subjects from maladaptive to adaptive. For this purpose the aforementioned explanatory variables are incorporated in the analysis in such a way that the dimensions are constrained to be linear combinations of the explanatory variables. The main focus here is to interpret the first dimension as the intended underlying bipolar scale.

The vegan package in R (Oksanen et. al, 2010) was used to perform the analysis and create the two-dimensional CCA solution for the DP data. The resulting CCA diagram is presented in Figure 2.7.

For the current DP data, the total inertia is 0.5595. Eigenvalues of the first two dimensions are 0.124 (22%) and 0.010 (2%), respectively. The diagram thus represents almost 24% of the variance in the observed scores, with a clearly dominant first dimension. Eigenvalues of the first two dimensions in the unconstrained CA solution for this data are 0.225 (40%) and 0.103 (18%), respectively. The two-dimensional CCA solution thus explains (0.124 + 0.010)/(0.225 + 0.103) = 41% of the unconstrained CA solution.

According to Ter Braak (1986) the aim must not be to explain 100% of the variation, because part of the total variance is error variance. The obtained solution can be considered as satisfactory, given we have only three explanatory variables in the current example.

The first CA dimension can be expressed as the following linear combination of the explanatory variables: \( \theta_i = -0.23 \ SCL-90_i + 0.20 \ Age_i + 0.78 \ Healthy_i \), in which the coefficients are interpreted as standardized regressions weights.

This means that, correcting for age and health-status, higher scores on the
SCL-90 (i.e., suffering from more psychological distress) are associated with subject locations toward the left extreme of the scale. Furthermore, a higher age is associated with subject locations toward the right extreme of the scale, after correcting for the other explanatory variables. And even more strongly, normal controls (i.e., Healthy = 1) are associated with subject locations toward the right extreme of the scale as well, or conversely, psychiatric patients (i.e., Healthy = 0) are associated with subject locations toward the left extreme of the scale, after correcting for SCL-90 scores and age. This finding is a confirmation of the earlier conclusion that underlying the scores on the DP, is a bipolar scale that reflects the extent to which an individual’s psychosocial functioning is adaptive or maladaptive.

The explanatory variables are displayed as vectors from the origin, of equal length (scaled to fit in the diagram), where a vector’s angle with dimension $s$ represents the correlation between the variable and the dimension (if the angle is sharp, the correlation is positive, if obtuse, negative). These angles thus indi-
2.5. CCA for Item Analysis: the Developmental Profile

cate how well the explanatory variables explain the patients’ scores via the CCA dimensions. In Figure 2.7, it can be seen that the dummy variable “Healthy” (0 = psychiatric inpatient, 1 = healthy control) is most strongly related to the first CCA dimension \( r = .97 \), which means that this first dimension strongly discriminates psychiatric inpatients from healthy controls. The correlations of the first dimension with, respectively, SCL-90 and Age, are -.72 and .40.

We learn from the CCA diagram that the levels of the DP (with the exception of Egocentricity) are indeed ordered on the first dimension according to the theory. The canonical coefficients (see Section 2.2.2) provide support for interpreting this dimension as bipolar with maladaptive (read: psychiatric) subjects on one pole and adaptive subjects (read: healthy controls) on the other pole. Furthermore, it can be concluded that functioning on maladaptive levels is associated with more psychiatric complaints (as measured by the SCL-90), and (to a lesser extent) that developmental differences in adults, as measured by the DP, are partly a natural result of aging.

Note that the vectors point in the direction of the maximum change in value of the associated variable. In the perpendicular direction the variable does not change in value. Note that, in principle, one may extend the vector to infinity in both directions. The perpendicular projection, of each item point \( \delta_j \) on a vector representing an explanatory variable \( x_h \) gives the weighted average of the subject distribution of \( x_h \). Thus the item with the highest projected value on, for instance, the explanatory variable SCL-90, is the item that is most strongly associated with psychological distress. In Figure 2.7 it can be seen that levels 2 and 1 (i.e., respectively Fragmentation and Lack of Structure) are, as expected, most strongly associated with psychological distress.

Furthermore, the (weighted) centering of the solution enables the following interpretation: if the projected item point is on the same side of the origin as the arrowhead of the associated vector representing the explanatory variable, then the inferred weighted average is higher than average, and lower otherwise. For instance, in Figure 2.7 it can be seen that the subjects with high scores on the level Generativity, are (much) older than average.

Egocentricity is an outlier in the CCA diagram. The orthogonal projections of Egocentricity on the vectors of Healthy and SCL-90 show, that on the one hand Egocentricity is indeed a characteristic that is present in the psychiatric subgroup, rather than in the subgroup of healthy controls; but, on the other hand, that functioning on the level of Egocentricity is associated with a (very) low level of psychological distress (as measured by the SCL-90). With respect to the scores on
the SCL-90 people with high scores on Egocentricity are comparable with healthy controls, rather than with the subgroup of psychiatric patients. The orthogonal projection of Egocentricity on the vector of age shows that the subgroup of people with high scores on Egocentricity is relatively young. In other studies, this level was also found to represent an atypical subset of patients (cf. Polak et al., 2010). Future research may provide explanatory variables, which cast a light on the unique position of this level.

2.6 Discussion

This chapter explained CA as a CTT-like approach to item analysis for bipolar measurement scales. Furthermore, the use of CA as an alternative to the unfolding IRT models, in particular GGUM and MUDFOLD, was investigated. For this purpose both real data from the field of attitude research, and simulated benchmark datasets were used. Finally, constrained CA (or CCA), an extension of CA in which the measurement scale is constrained to be a linear combination of explanatory variables, was proposed as a counterpart of explanatory IRT. This approach is a recent development in monotonic IRT models (cf. De Boeck & Wilson, 2004), but has not yet been fully incorporated in unfolding IRT methods. CCA was explained by means of a real data set from the field of personality assessment.

The analysis of real data from the field of attitude research illustrated the use of CA as a technique for item analysis and psychometric evaluation of a bipolar measurement scale. From a comparison with two unfolding IRT models (i.e., GGUM and MUDFOLD), it became clear that there are several advantages of performing CA on single-peaked item response data. First, the analysis is relatively simple, that is, CA always results in a solution for all items and subjects. In contrast to GGUM and MUDFOLD, CA does not require, respectively, either a pre-selection of items (which is based on PCA in GGUM) or a starting set of items, in order to converge. Furthermore, the CA approach resulted in retaining more items in the scale, which also seemed to result in a more accurate representation of the subject locations. The CA solution showed to be stable, that is, the item location estimates were the least affected by discarding items from the scale. Finally, the CA solution could be validated, yielding high correlations with the original Thurstone scale values that were available for this dataset.

Since CA is regarded as an exploratory technique, no fixed criteria exist for item selection based on the CA solution. We followed Ter Braak and Prentice (2004) for the identification of “rare” and deviant items, where we discarded two
items that both had a remote coordinate in the two-dimensional CA solution and one item that was a central point in the solution. All three items were “rare” given the low percentage of strongly agree responses (i.e., 2.4 % to 3.2 %). The argument for discarding these items from the scale is that they hardly discriminate between individual subjects (as they appeal to almost none of the subjects).

However, it is difficult to give a lower limit for the percentage of strongly agree responses that justifies retainment of an item in the scale. Ter Braak and Prentice (2004) point out that additional analyses need to be done to check the assumption of single-peaked, and equally discriminating IRFs. When additional analyses show that an item discriminates between subjects, a relatively rare item may still be considered important by a practical researcher and could therefore be retained in the scale. An example would be an item that operationalizes rare, but clinically relevant behavior. A previous study by Polak et al. (2010a) presented a method based on ordered conditional means (OCM) to roughly estimate the IRFs, which can be used to check the assumption of single-peaked, and equally discriminating IRFs. In that study the capital punishment data were also analyzed and the three items that were discarded in the current chapter, were indeed identified as deviant, since they all showed a flat (i.e., non-discriminating) IRF. We argue that an additional analysis, such as the OCM method, is necessary to decide whether or not items must be discarded from the CA solution.

The analyses of the benchmark datasets show that, in the studied conditions, all techniques recover the true item locations (or their ordering in the case of MUDFOLD) very well. With respect to the subject locations, the techniques show some important differences. First, overall, CA and GGUM perform better than MUDFOLD with respect to recovering the correct ordering of subjects along the scale. Second, results show that the true ordering of subjects on the scale is determined more accurately in the conditions with 20 items, than in those with 10 items. Third, although CA performed approximately equally well as GGUM in retrieving the correct subject ordering, the Pearson correlations between the true and estimated subject locations are higher for GGUM than for CA. This may be explained by the edge-effect in CA, which was apparent in the conditions with unevenly spaced items, but not in the conditions with equidistant items.

A remarkable outcome was that the relative estimates of the subject locations around the center of the scale were accurate for all techniques, even in those conditions where the items around the center of the scale were lacking (see Figure 2.6). This finding is important, since it implies that even for scales consisting of items that were selected with the criteria for Likert scales (thus selecting two sets of,
respectively, strongly negative and strongly positive wording), unfolding analysis (either performed with CA or unfolding IRT) is still appropriate for recovering item and subject locations. An advantage of unfolding analysis in this context is that, in contrast to the Likert approach, it is not required to reverse-score one of the sets of items before the data can be analyzed. Unfolding analysis results in a bipolar scale, which makes it possible to derive separate location estimates for both indicative and contra-indicative items. Thus unfolding does not require the strong assumption of the Likert procedure, that disagreement with a certain indicative item (e.g., “Capital punishment is just and necessary”) is equal to (or implies) agreement with a certain contra-indicative item (e.g., “I do not believe in capital punishment under any circumstances”).

Furthermore, Figure 2.6 showed that, in contrast to GGUM, CA underestimated the spread of subject locations at both extreme ends of the scale. In the CA literature, the compression of the subject locations near both ends of the first dimension is known as the edge-effect, which sometimes occurs as a side-effect of the better known arch-effect (see Ter Braak & Prentice, 2004). It is interesting to see that the edge-effect did not occur in the conditions with evenly spaced item locations, although the arch-effect did occur. Thus, a two-dimensional CA display of one-dimensional data results in an arch-shaped pattern of both subject and item locations. We found the subjects to be evenly spaced along the arch (and along the first dimension) for all benchmark datasets with equidistant items, but unevenly spaced along the arch (and along the first dimension) for all benchmark datasets with items that were clustered on both ends of the scale. In the latter conditions the subject locations near the ends of the dimensions were relatively more compressed than those in the middle.

Ter Braak and Prentice (2004) state that the edge-effect becomes less strong, as the range of site (subject) scores becomes wider and the spacing of the site (subject) scores and species (item) scores becomes closer relative to the average species tolerance (item discrimination).

An explanation for the fact that the edge-effect did occur in the datasets with unevenly spaced items (gap), but did not occur in the datasets with evenly spaced items, can be found in the differences between the distributions of the (row-)marginal proportions (i.e., the subjects’ total scores divided by the grand total) of the datasets resulting from the different conditions (see appendix A for the computational details of CA). Note that with attitude data using graded response scales (e.g., with response categories 0 = strongly disagree to 5 = strongly agree) a relatively high total score indicates that a subject agrees with relatively many
2.6. Discussion

Items.

Subjects with a relatively extreme position on the first CA dimension, are expected to have high scores exclusively on the most extreme item(s) on the corresponding end of this dimension. For these subjects the scores on the other items are less high as these items lie further away from the extreme end of the scale. When the items are evenly spaced along the dimension, in general, subjects with an extreme position have the lowest total score (or marginal proportion), since less items appeal to these subjects than to subjects with a more moderate position. In the conditions with evenly spaced items, we observed a bell-shaped relation between the subject locations and the respective total scores: the more extreme the subject location, the lower the total score. Since in CA the inverse marginal proportions serve as weights in determining the inter-point distances, the subjects with an extreme position lie relatively far away from the origin. This explains the accurate representation of the subject locations of CA in the conditions with evenly spaced items.

However, when we have unevenly spaced items, like in the present chapter, with a gap in the center of the scale, not only subjects on the extreme ends of the scale have relatively low total scores, but also subjects that lie near the center of the scale (because for the latter there are no items located close to their position, hence no items to strongly agree with). In the conditions with unevenly spaced items, we did not observe a bell-shaped relation between the subject locations and the respective total scores, instead there was no clear relation at all. This means that in these conditions the subjects with an extreme position lie relatively closer to the origin, since their weighting is less extreme than in the evenly spaced conditions. This explains the underestimation of the subject locations of CA in the conditions with the unevenly spaced items.

A limitation of the present chapter is, that it is exclusively concerned with the analysis of one-dimensional data. Although in many instances the interest lies in recovering one latent scale (see the examples given in the current chapter), it is also often the case that a measurement instrument consists of several (sub-)scales. An example of a measurement instrument that consists of several subscales is an attitude questionnaire that measures attitudes toward several issues, such as immigration, death penalty, and aid to the developing countries, using subsets of items. The use of CA as alternative to PCA for evaluating more than one (bipolar) subscale needs still to be addressed. Two topics of future research into the use of CA for evaluating several bipolar (sub-)scales are of particular interest, that is, the topic of rotation and that of the arch-effect in CA.
2. Psychometric Evaluation of Bipolar Measurement Scales

To start with the first, in the context of item selection and subscale evaluation for unipolar scales (see the scheme in Figure 2.1 for the differences between bipolar and unipolar scales), PCA with varimax rotation is the most common approach (see for example, Tabachnick & Fidell, 2001, Chapter 13). As this approach results in an orthogonal solution, with usually one cluster of highly loading items for each component, the components point out potentially relevant subscales (subsets) of items. Van de Velden and Kiers (2005) introduced orthogonal varimax rotation in CA. However, it is not straightforward to generalize the PCA approach for recovering several unipolar scales to a CA approach for recovering several bipolar scales. For one, in contrast to unipolar scales, bipolar scales also include items on the midpoint of the scale that necessarily have low correlations with the underlying dimension. Therefore, maximizing the separation of subjects or items (depending on the type of normalization in CA, see appendix A) along the dimensions, might not necessarily result in uncovering the relevant (bipolar) subscales. Furthermore, it seems relevant to investigate oblique rotation as well, since often we deal with (cor-)related subscales.

A second hurdle to overcome is the so-called arch-effect. This is an artifactual second dimension (namely a quadratic function of the first dimension), which shows up as an arch pattern of both the item points and the subject points that has no further meaningful interpretation. When data are truly one-dimensional, the arch-effect is not a serious problem, because it can be regarded as evidence for a strong first dimension, and the ordering of the subjects and items on the first dimension is not disturbed by the arch-effect. However, in cases where, there is not only a strong first dimension underlying the data, but also a second dimension, with a less strong, yet relevant interpretation, this second dimension may be obscured by the arch-effect. In those cases, the true second dimension, might appear as the third dimension in the CA solution. Several methods to eliminate the arch-effect have been proposed, including detrended correspondence analysis (Hill & Gauch, 1980; Ter Braak & Prentice, 2004), detrended canonical correspondence analysis (Ter Braak, 1986), and principal curve fitting (De’ath, 1999). Future research should find out which solution suits the purpose of recovering several bipolar subscales best.