6.1 Conclusions of the Technical Chapters

The purpose of this thesis was to contribute to item analysis of single-peaked items, that is, the psychometric evaluation of bipolar scales. We proposed correspondence analysis (CA) as a method to find scale values (i.e., location estimates) for both subjects and items on the same underlying bipolar measurement scale. In the following we will give a chapter-wise summary of the main conclusions presented in Chapters 2, 3 and 4 of this thesis. Section 2 of this chapter summarizes and discusses the results presented in Chapter 5 that concerned the validation study of the Developmental Profile. In section 3 of the current chapter we discuss the results of Chapters 2, 3 and 4, which yields a number of suggestions for future research.

Main conclusions of Chapter 2. In Chapter 2, we compared CA with the unfolding IRT models GGUM and MUDFOLD for the item analysis of single-peaked response items based on real data and simulated benchmark data. Furthermore, we proposed constrained CA (or CCA), an extension of CA in which the measurement scale is constrained to be a linear combination of explanatory variables, as a counterpart of explanatory IRT. This approach is a recent development in monotonic IRT models (cf. De Boeck & Wilson, 2004), but does not yet exist for unfolding IRT models. CCA is explained using results for real data from the field of personality assessment.

The use of CA as a technique for item analysis and psychometric evaluation of a bipolar measurement scale, was illustrated by the analysis of real data from the field of attitude research. 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, and consequently resulted in a more accurate representation of the subject locations. Next, 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 existed for item selection based on the CA solution. We followed Ter Braak and Prentice (2004) for the identification of “rare” and deviant items. As a result, we discarded two items, which 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 (in the study described in Chapter 2, less than 5 %). The argument given in Chapter 2, 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 retention 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 of such an item is an item that operationalizes rare, but clinically relevant behavior.

In Chapter 4 we 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 study described in Chapter 2, 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 described in Chapter 4, is necessary to decide whether or not items must be discarded from the CA solution.

The analyses of simulated benchmark datasets showed 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 showed some important differences. First, overall, CA and GGUM performed better than MUDFOLD with respect to recovering the correct ordering of subjects along the scale. Second, results showed that the true ordering of subjects on the scale is determined more accurately in the conditions with 20 items,
6.1. Conclusions of the Technical Chapters

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. The poorer recovery of the true spread of subject locations in CA may be explained by the edge-effect in CA. The edge-effect 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. This result 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. A first advantage of unfolding analysis in this context is that, in contrast to the Likert approach, unfolding analysis represents the most extreme subjects correctly.

A second advantage of unfolding is, 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”).

Main conclusions of Chapter 3. In Chapter 3, we showed that principal component analysis (PCA) is not appropriate for analyzing bipolar scales. This chapter gives an overview of earlier papers on the results of PCA on bipolar scales, and classifies the type of response models discussed in these earlier papers as either a quadratic function of the person-to-item distances or an exponential function of these distances. It was shown that this distinction is easy to recognize empirically because the inter-item correlation matrix for the two types of data typically shows different patterns. Furthermore, we showed that for both types of unfolding models, CA, which is a rival method for dimensionality reduction, outperforms PCA in terms of representation of both person and item locations, especially for the exponential model.

Finally, we showed that undoubled CA outperforms doubled CA for both types
6. **General Discussion**

of unfolding models. This result is important, since doubling is the standard procedure for analyzing ratings with CA (see for example, Greenacre, 1993, chap. 19; Greenacre, 2007, chap. 23). In Chapter 3 it was shown in which situations this procedure is not suited. Results showed that performing CA on the raw data matrix is an unconventional, but meaningful approach to scaling items and persons on an underlying unfolding scale. A real data example on personality assessment was given, which showed that for this type of data (undoubled) CA outperforms PCA.

**Main conclusions of Chapter 4.** In Chapter 4 we proposed a model-free diagnostic for single-peakedness (unimodality) of item responses. Presuming a unidimensional unfolding scale, and a known item ordering, it approximates item response functions (IRFs) of all items by computing ordered conditional means (OCM) under the assumption that these functions are unimodal. The proposed OCM methodology was based on the criterion of irrelevance, which is a graphical, exploratory method for evaluating the “relevance” of dichotomous attitude items. We generalized this criterion to polytomous items and quantified the relevance by fitting a unimodal smoother (Eilers, 2005). The resulting goodness and badness of fit measures, respectively, $R^2$ and $RMSE$, were used as measures of scale fit. Item fit was determined by the change in scale fit “if item deleted”.

Sampling behavior of both measures of scale fit and item fit was explored in a Monte Carlo simulation varying several conditions, including, subject distribution, scale length, number of deviant items, type of deviation, and location of the deviant item on the scale.

Both measures of scale fit showed, as expected, a diminishing decline as a function of the number of deviant items in the scale. Values of $R^2 = .97$ and $RMSE = .044$ were found as thresholds of good fit. The scale fit did not depend on type of subject distribution, nor did the scale fit differ for both conditions of scale length.

Both measures of item fit strongly discriminated between the deviant and regular items. Values of “change in $R^2$ if item deleted” $\geq .025$ and “change in $RMSE$ if item deleted” $\leq -.005$ were found as thresholds for substantial improvement in scale fit, which could be used for identifying deviant items.

ANOVA showed that the measures of item fit were neither affected by subject distribution, nor by scale length. However, “change in $R^2$ if item deleted” was moderately affected by “type of deviant IRF”, where this measure of fit had more power for identifying non-discriminatory items than items with an irregular IRF.
A second reservation we have to make is that “change in $RMSE$ if item deleted” was affected by “number of deviant items”, where this measure of fit had more power for identifying deviant items in those conditions where the total number of deviant items in the scale was relatively small.

Overall, results of the simulation study indicated, that the power of identifying deviant items on the basis of threshold-values of “change in $R^2$ if item deleted” $\geq .025$ and “change in $RMSE$ if item deleted” $\leq -.005$, was high, while at the same time the type I error rates remained acceptable in most conditions.

It was concluded that the surplus value of the proposed OCM method is, that it provides approximations of the IRFs of all items in a scale, and it is not, like the existing unfolding IRT methods, dependent on model convergence or a parametric model form. As was also shown in Chapter 2, unfolding IRT models such as GGUM and MUDFOLD may require a pre-selection of items, in order to allow the model to converge and provide results. This pre-selection-step may not be straightforward for practical researchers, as it requires other methods than provided by the models themselves.

The proposed methodology could be used in combination with any of the existing unfolding models. In that case, the estimated item locations based on the specific model could be used, instead of the item ranks, to make up the OCM diagrams.

Taken together, the results presented in Chapters 2, 3 and 4 indicate that we succeeded in providing the hitherto lacking methodology in the collection of scaling techniques that were classified in Figure 2 in Chapter 1. CA is a good method for a first selection of items that together form a single bipolar measurement scale. More importantly, CA can be used to find location estimates for items and subjects on the bipolar scale. CA does not suffer the problem of underestimating the locations of extreme items and extreme subjects, like PCA does.

To measure the internal consistency of the set of items, selected on the basis of CA, we developed a methodology (OCM) that approximates item response functions (IRFs) of all items. The OCM methodology incorporates a unimodal smoother that is used to quantify the item fit for each item in the scale. The methodology can be used on the basis of a sound hypothesis concerning the ordering of the items along the scale. Thus, it does not necessarily require the CA location estimates. An attractive property of the OCM methodology is that it can be used in combination with any unfolding method, hence also with unfolding IRT methods. In that case, item location estimates from those methods can be
used instead of the hypothesized rank numbers. We recommend to use the OCM methodology after a CA solution is found to judge the relevance of the items in the solution. For instance, poorly discriminating items, or items with an irregular IRF, can be identified as candidates for deletion.

6.2 Conclusions and Discussion of the Applied Study on the Validity of the Developmental Profile

In Chapter 5 we reported results of the validation study concerning the internal structure of the Developmental Profile (DP; Abraham, 1993, 2005; Abraham et al., 2001). We presented a combination of confirmatory factor analyses (CFA), complemented with Cronbach’s alpha coefficients (Cronbach, 1951), and CA to evaluate the main theoretical assumptions underlying the DP. A large sample was studied (N = 763) with participants from various clinical and nonclinical settings in the Netherlands.

CFA showed an overall good fit, thereby providing a justification for the organization of item scores into level scores. Furthermore, the CFA results justify aggregating the various levels into three clusters, thus supporting constructs of a primitive maladaptive cluster (Lack of Structure, Fragmentation, and Egocentricity), a neurotic maladaptive cluster (Symbiosis, Resistance, and Rivalry) and an adaptive cluster (Individuation, Solidarity, and Generativity).

To provide a further justification for working with the level scores and cluster scores (e.g., to correlate developmental level scores with other psychological variables, such as psychiatric diagnosis, in further research) we also reported Cronbach’s alpha.

We mentioned in the discussion section of Chapter 5, that we used Cronbach’s alpha, because it is the most common statistic for this purpose. Recently, there has been a debate in psychometrics about the tenability of Cronbach’s alpha that was started by Sijtsma (2009a; see also Bentler, 2009; Green & Yang, 2009a, 2009b; Revelle & Zinbarg, 2009; Sijtsma, 2009b). The key points of Sijtsma (2009a, 2009b) are, in short, first, that Cronbach’s alpha underestimates the reliability of the total score and better estimates exist that deserve more attention (special attention is called to the greatest lower bound). Second, that Cronbach’s alpha should neither be interpreted as a measure of internal consistency, which is motivated by the fact that alpha merely reveals the “average degree of interrelatedness”, nor as an indication of unidimensionality. Third, that alpha is not really
meaningful when it comes to diagnosing individuals, because: “statistical results based on a single test administration convey little if any information about individuals measurement accuracy reflected by their propensity distributions” (Sijtsma, 2009a, p. 119).

At this point, we want to call attention to this discussion concerning the use of Cronbach’s alpha. We agree with Sijtsma that psychometricians have an obligation to communicate these reservations and alternatives to psychological researchers who work outside the field of psychometrics.

In the light of this discussion concerning Cronbach’s alpha, we make the following recommendations for future research concerning the psychometric quality of the DP\(^1\). With respect to reliability research, we think an important topic should be to establish the stability of the level scores. A challenge is to unravel the various sources of error (i.e., random disturbances) that are part of the observed level scores. A way to go would be to include test-retest conditions in interrater reliability studies. Thus crossing the factor “rater” with the factor “time”. Other methodological recommendations for further interrater reliability research were made by Polak, Abraham, Van, and Ingenhoven (2010).

As a further recommendation, we agree with Ingenhoven (2009), that to make the assessment less time consuming, but also less susceptible to rater variability, a self-report questionnaire based on the DP would be a possibly attractive addition to the standard DP scoring procedure.

It is important to note, that the items of this instrument are organized primarily in cumulative subscales, but that results support the notion that underlying these subscales is a substitutive (bipolar) scale ranging from strongly maladaptive functioning to strongly adaptive functioning. To conclude this section we want to address the topic of the bipolar continuum underlying the nine levels (i.e., subscales) of the DP. It has been shown, that subjects’ positions on a bipolar scale can not be derived from their general total score (cf. Thurstone, 1928; see also Chapter 1, sections 2 and 5 of the current thesis), that is, the sum over all level scores. Instead, a subject’s position on this bipolar continuum should be derived as the (weighted) mean of the positions of those levels on which the subject scored highly. Throughout this thesis, CA was shown to be explicitly suited for this purpose. In DP study, CA was performed on the level scores (thus the subscale

---

\(^1\)The DP is the focus of an ongoing research project, carried out by a multi-disciplinary team made up of practitioners and researchers with various backgrounds in psychiatry and psychology.
scores), with the objective to find locations of these levels, and of the subjects on the same scale.

Results indicated that the developmental levels were indeed arranged on a bipolar continuum ranging from maladaptive to adaptive psychosocial functioning. Subject locations were considered theoretically meaningful, as a strong distinction was found between several patient groups varying with respect to their psychiatric or psychological complaints, and a group of healthy controls.

Constrained correspondence analysis (CCA) on an a subset of the DP data ($N = 105$), presented in Chapter 2, provided additional support for the underlying bipolar scale. In the CCA diagram resulting from this analysis, the dimensions were constrained to be linear combinations of the following explanatory variables: 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 = patient, 1 = healthy control).

The first dimension in the CCA diagram (cf. Chapter 2, Figure 7) was indeed bipolar, with maladaptive (read: psychiatric) subjects on one pole and adaptive subjects (read: healthy controls) on the other pole. Furthermore, it could 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.

We regard CCA as a useful methodology for further studies concerning the organization of both developmental levels and individual patients on this underlying bipolar continuum. In particular, we aim to study explanatory variables, which might point out the true position of the level of Egocentricity, and correspondingly, the positions of patients with narcissistic personality characteristics on the bipolar continuum.

6.3 General Discussion and Recommendations for Future Research

A limitation of the present thesis 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 throughout this thesis), 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 correspondence analysis (CA) as alternative to principal component analysis (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.

To start with the first, in the context of item selection and subscale evaluation for unipolar scales (see the scheme in Figure 2, Chapter 1, for the differences between bipolar and unipolar scales), PCA with varimax rotation is the most common approach (see for example, Tabachnick & Fidell, 2001, chap. 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 correspondence analysis. 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, which 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. The arch-effect 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 underlying the data there is not only a strong first dimension, 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, may 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 re-
search should reveal which solution suits the purpose of recovering several bipolar subscales best.

Both in Chapter 2 and in Chapter 3, it appeared that in some conditions, 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 arch-effect mentioned in the previous paragraph. In both chapters it was observed that, the edge-effect is related to the distribution of the subjects’ total scores. When it holds that subjects with more extreme locations, also have lower total scores, than the edge-effect does not occur. Additional to the methods to eliminate the edge-effect mentioned in the above, we can conclude with Ter Braak and Prentice (2004), that the edge-effect becomes less strong, as the range of subject scores becomes wider and the spacing of the subject scores and item scores becomes closer relative to the average item discrimination. On the basis of the results presented in Chapter 3 it may be recommended, to choose evenly spaced items when constructing a bipolar measurement scale; and rather a total number of items close to 20, than close to 10.

In Chapter 4 we presented a new methodology to estimate item response functions (IRFs) of single-peaked items with a model-free approach (OCM). A weakness of the methodology is that it depends on a sound hypothesis concerning the true item ordering. Failure in specifying the correct item ordering will also result in misfit.

A first solution to this problem is, to derive an optimal item ordering data-analytically by performing CA on the raw data (see also Chapters 2 and 3). Future research will be performed to evaluate the integration of the CA item location estimates in the OCM methodology.

A second solution to this problem is, that when for a set of items the correct ordering is known only for a subset of the items, the columns of OCM matrix could also be based on these items only. In other words, the OCM matrix does not have to be square. In that case the estimated IRF for each item is based only on ordered conditional means on this subset of items. For instance, for the Thurstone scale discussed in Chapter 4, an optimal subset was selected by Roberts and Laughlin (1996) based on the IRT model GUM. Instead of the 24 points that originally made up the OCM diagrams in the Thurstone scale example, we could have used the 12 points based on the GUM selection.

Another way to prevent misfit due to an incorrect item ordering is to obtain subject matter expert judgments of item location, especially when sample sizes
6.3. General Discussion and Recommendations for Future Research

are too small to estimate location parameters with unfolding IRT models.

Finally, it might be difficult to specify the item ordering when items vary in how strongly they discriminate among respondents. Particularly, when several items appeal to respondents from a broader range of locations on the continuum, the (observed) preferred item ordering will vary across respondents.

A limitation of the simulation study design that was presented in Chapter 4, was that sample size was not varied. In the current simulation procedure sample size was fixed at \( N = 300 \) with the motivation that this sample size is realistic in applied research. However, to provide further support for the cutoff values of the fit indices that were presented in Chapter 4, further simulation studies with varying sample sizes are desired. This will result in a more complete overview of Type I error rate and power of the proposed fit indices and their cutoff values. Furthermore, new simulations may be enhanced by including the family of \( S \)-Chi-square statistics (Roberts, 2008) that are included in the GGUM2004 software as a comparison with the proposed OCM methodology.

Although the evaluation of OCM methodology in Chapter 4 was limited to item fit, the OCM methodology could also be generalized to measure person fit. For that purpose, instead of the OCM matrix, the raw data matrix could be plotted, with a diagram of each subject’s score pattern. Given that the order of the columns reflects to hypothesized item order, the expected score pattern in each row is also single-peaked. Several authors (e.g., Emons, Sijtsma, & Meijer, 2004; Sijtsma & Meijer, 2001) showed that the person response function is an important tool in person-fit research. Analogous to the interpretation of the OCM matrix proposed in Chapter 4, we could interpret the score pattern (standardized with respect to the maximum score) within each row of the data matrix as a rough estimate of the person function. Accordingly, we could assess the person fit by fitting the smoother as proposed in Chapter 4.

Altogether, we think the OCM methodology provides useful diagnostics for both item fit and overall scale fit for single-peaked response items. Further research has to been done to the generalization of this methodology to assess person fit as well.

In conclusion, this thesis called attention to the topic of uncovering bipolar scales. In the current practice of scale and test construction the Likert approach seems to have become far more popular than the methodologies for scale construction originally proposed by Thurstone. It was argued that when one has indicative and contra-indicative items, one should consider whether these items do, in fact,
form two opposite poles of one dimension. Because in that case one should not use the Likert approach, which implies reverse-scoring either one of both sets of items, so that all items point in one direction. Instead, one should use an approach that results in a bipolar scale, which makes it possible to derive separate location estimates for both indicative and contra-indicative items. It would be interesting to apply CA (but also unfolding IRT approaches) to practical data examples where the Likert approach was used, but a bipolar measurement scale is suspected. This re-introduction of Thurstone’ s concept of bipolar measurement scales into the field of Likert scaling, will broaden the range of applications of the unfolding techniques that were evaluated in the current thesis.