T E C H N I C A L

R E P O R T

0691

IS DEPRESSION DIMENSION- OR CATEGORY-LIKE ? A PSYHOMETRIC ANALYSIS USING DIMCAT AND TAXOMETRICS

HIDEGKUTI, I., DE BOECK, P. and J. NEZLEK

I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Running head: Is depression dimension- or category-like?

Is depression dimension- or category-like? A psychometric analysis using DIMCAT and taxometrics

István Hidegkuti and Paul De Boeck

K.U. Leuven, Belgium

John Nezlek

College of William and Mary, Williamsburg, VA

Abstract

In psychopathology research it is long debated whether psychological disorders are best conceptualized as categorical or dimensional constructs. Taxometrics is a well established and widely used approach to investigate this question, whereas DIMCAT is a more recent approach that provides an elaborate framework to address the dimensional vs. categorical debate. In this paper both approaches were applied to three data sets to study the underlying nature of depression as measured by the CES-D Scale. In line with the previous findings with other depression measures, the taxometric analyses found the structure underlying depression to be dimensional. Using DIMCAT, however, it was found that depression in fact shows category-like features. These contradictory results can be explained through a different conceptualization of being category-like, but they are also informative for how the two approaches work and for the nature of depression as measured with the CES-D.

Introduction

In the psychopathology research there is a long-lasting and yet unresolved debate about the underlying nature of depression. The focus of the debate is whether depression is best conceptualized as a genuine category that is separated from normality, or, alternatively, as part of a dimension ranging from normality to severe depression, including lots of intermediate cases between the extremes of the dimension. The Diagnostic and Statistical Manual of Mental Disorders 4th Edition (DSM-IV; American Psychiatric Association, 1994) represents the former conceptualization describing depression (and other mental disorders) as a distinct category. The way the list of symptoms is defined in order to make a diagnostic decision suggests some graduality, but the DSM-IV is basically still categorical. The DSM-IV reflects the way medical doctors view illnesses and disorders, considering a certain illness to be either present or absent. Depression is not seen as a continuum, but as a dichotomy instead. This way of thinking of depression and of other mental disorders appears to be primarily intuitive, whereas no sound empirical evidence is presented that supports the categorical nature of depression and most other disorders.

The importance of the question whether psychological disorders are categorical or dimensional is manifold. First, this problem is an interesting one from a theoretical point of view. One of the aims of science is to better understand and describe various facets of the world, to construct theories about it and test those theories. Second, the categorical vs. dimensional issue is very important from a practical point of view. The proven categorical nature of a certain psychological disorder may have several practical implications. On the one hand, it may affect the construction of assessment tools, and the selection of research samples (Meehl, 1992). When the underlying nature of a certain disorder is categorical it may be very important to have well-founded, nonarbitrary cutpoints when studying those phenomena. When, for example, a questionnaire is used for the selection of a research sample with a psychological disorder, the cutpoint has to be such that only participants having the disorder are selected (e.g., J. Ruscio & Ruscio, 2004; Santor & Coyne, 2001). Of course, even if the disorder turns out to be of a dimensional type, a (to some extent) arbitrary cutpoint may be used but one has to be aware of the arbitrariness of that cutpoint (J. Ruscio & Ruscio, 2004). On the other hand, the categorical or dimensional nature of the phenomenon is also relevant for the nature of the cause, the etiology, and also for the treatment and the evaluation of its effects. For example, the categorical nature of a disorder could either suggest a categorical, for example a delineated genetic cause (e.g., Meehl, 1972), or a model such as in catastrophe theory where a specific constellation of continuous processes leads to a sudden, dramatic change, that is, to a qualitative change in the behavior, for instance (e.g., van der Maas & Molenaar, 1992; van Geert, 1991).

The categorical versus dimensional nature of depression can be approached from several different ways. Flett, Vredenburg, and Krames (1997) describe four possible views on the continuity of depression. Based on the review of the literature, they distinguish phenomenological, typological, etiological and psychometric continuity. *Phenomenological* continuity is based on the assumption that persons with different levels of depression experience the same symptoms independent of the severity of depression, and that there are only quantitative differences in the intensity and persistence of those symptoms. The *typological* approach focuses on the question whether well-separated subtypes of depression can be identified, while the *etiological* view investigates if persons having subclinical levels of depression symptomatology are at a risk of developing a more severe form of depression. The fourth approach described by Flett et al. is to investigate the *psychometric* continuity of depression, that is, if a measure of depression is continuous. The focus of the present paper is on the last, that is, on the psychometric continuity. A psychometric approach implies that one concentrates on indicators of a disorder, such as symptoms or items in an inventory. The psychometric approach is not unrelated to the other approaches, because the indicators are of course relevant to the phenomenology, the typology, and sometimes also to the etiology of depression.

There have been numerous methods used to investigate the underlying nature of psychological phenomena. Not only were different clustering methods applied to address this problem, but also a specific methodology called taxometrics has been developed by Paul Meehl and his colleagues (e.g., Meehl, 1973; Meehl & Yonce, 1994, 1996; Waller & Meehl, 1998). This method, or more precisely, group of methods has repeatedly been used to reveal the nature of different psychological phenomena (e.g., borderline personality disorder, Trull, Widiger, & Guthrie, 1990; hypnotic susceptibility, Oakman, & Woody, 1996; schizotypy, Lenzenweger & Korfine, 1992; type A personality, Strube, 1989), of which depression is one of the most intensely investigated. The authors of most studies using taxometrics (e.g., Hankin, Fraley, Lahey, & Waldman, 2005; A. M. Ruscio & Ruscio, 2002, J. Ruscio & Ruscio, 2000) conclude that there is evidence for the continuity of depression. Beach and Amir (2003) report ambiguous results, suggesting that the conclusions concerning the categorical or dimensional nature of depression are also dependent on the indicators that are used. Santor and Coyne (2001) used a nonparametric IRT approach to investigate the issue, comparing the item characteristic curves across two subsamples one of which consisting of depressed participants, and the other consisting of nondepressed participants. They found that several symptoms behave differently depending on the subsample; consequently, they argued that depressed and nondepressed people cannot be described by a single continuum, that is, the differences between the two classes are not dimensional but categorical.

A recent development in the continuity vs. discontinuity field is the Dimension/Category framework (DIMCAT; De Boeck, Wilson, & Acton, 2005). DIMCAT is based on item response modeling (IRT) and provides a broad and elaborated framework for the distinction between categories and dimensions. The DIMCAT framework basically makes distinctions along two (primary) axes. The first axis differentiates between *within-category*[1] *heterogeneity* (there is systematic variance within the categories)

[1]The terms category and class are used interchangeably throughout the manuscript.

and *within-category homogeneity* (there is no systematic variance within the categories). The second axis differentiates *between-category qualitative differences* (i.e., the categories cannot be differentiated exhaustively by referring to a degree of something) from *between-category quantitative differences* (i.e., the categories can be exhaustively differentiated indeed by the degree of something; e.g., category B having more of the same category A has). Qualitative differences show in a lack of parallelism of indicator profiles. A clear example of lack of parallelism is that the ordering of the indicator prevalence differs from one category to another. For example, mood indicators may be relatively more prevalent in depressed people than in normals, when compared with other types of indicators. When the categories are heterogeneous in terms of degree within the category, one can differentiate between two aspects of qualitative differences: level differences and relevance differences of the indicator values between the two categories. *Indicator level* differences refer to the relative levels of the indicators being different (e.g., feelings being relatively more prevalent). *Indicator relevance* differences refer to the relevance to discriminate between the persons within a category, or in other words, how much indicators are correlated to the degree of depression (or some other syndrome). In terms of IRT, indicator level corresponds to difficulty (threshold, location) and indicator relevance corresponds to discrimination. Indicator relevance refers also to factor indicator loadings in factor analysis.

When the differences are quantitative, a nested (secondary) axis differentiates between abrupt and smooth differences. Abrupt differences refer to a clear multimodality of the underlying joint distribution, whereas for smooth differences there is a large overlap and the underlying joint distribution is still unimodal.

Homogeneity is considered to be more category-like than heterogeneity, qualitative differences are considered to be more category-like than quantitative differences, and abrupt differences are considered to be more category-like than smooth differences.

The distinctions along the two (primary) axes define four basic types of category-likeness. The only combination that can be considered purely *categorical* is the combination of within-class homogeneity with between-class qualitative differences. The opposite of this combination, within-class heterogeneity with between-class quantitative differences, is considered to be a *dimensional* type of category-likeness, called here *heterogeneous categories on a dimension*. The remaining two combinations are even more hybrid combinations. The first hybrid combination is within-class heterogeneity with qualitative between-class differences, called *dimensional categories*. The other hybrid combination is within-class homogeneity with between-class quantitative differences, called *categories on a dimension* (*homogeneous categories on a dimension*). The graphical representation of category-likeness is depicted in Figure 1. The bottom left quarter of Figure 1 (categorical) is the most categorical, followed by the top left quarter (dimensional categories) and the bottom right quarter (categories on a dimension) and the top right quarter (dimensional) is the least categorical. The different models are to be explained when the DIMCAT model estimation is explicated in more details in the Method section. As explained, all these models are categorical, even when labeled "dimensional". A non-categorical or purely dimensional model is presented in Figure 2. In order to obtain categories from this model, a cutpoint is needed because there is no inherent basis to differentiate between categories.

The aim of the present study is to apply both approaches, taxometrics and DIMCAT, to different data sets, in order to test the categorical versus dimensional nature of the underlying structure for data from a depression questionnaire. Applying both approaches for the same data sets has the potentiality to yield insights in both the methodology and the categorical versus dimensional nature of depression.

Method

In the present paper the categorical versus dimensional nature of depression as measured by the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977) is investigated in three separate analyses using both the taxometric and the DIMCAT approach. Common features of the analyses are the use of the CES-D Scale and the large sample size (at least 2340 participants in each study).

Measure

The CES-D is a twenty-item self-report questionnaire that is one of the most widely used measures of depressive symptomatology in the general population (Santor, Gregus, & Welch, 2006). The questionnaire consists of indicators that represent symptoms that are considered to be characteristic for depression, with an emphasis on depressed feelings, but also somatic, retarded activity and interpersonal indicators are included. The twenty items relate to the past seven days, and the respondent is to evaluate on a four-point rating scale how often he/she felt or behaved in a certain way during the past week. The list of items is shown in Appendix A.

In the original publication of the CES-D, Radloff (1977) assigned the 20 depression indicators to four groups based on a factor analysis. Based on the content of the items, he named the four groups of items: *depressed affect, positive affect, somatic and retarded activity* and *interpersonal* items. In the present application, a maximum likelihood factor analysis was performed and two factors were extracted with eigenvalues larger than one. The first factor can be considered a general depression factor, on which all indicators had high loadings. On the second factor, the indicators with positive statements ("I felt just as good as other people", "I felt hopeful", "I was happy", "I enjoyed life") had high loadings, and hence this second factor can be considered a positive affect factor. From these results, the CES-D appears to be rather unidimensional, except for the four positive affect items that seem to form a secondary dimension, although they have high loadings also on the first factor.

Although there are 20 indicators in the CES-D scale, none of the analyses makes use of all the indicators for reasons to be described later. The indicator selection procedures follow from each of the two procedures, and will be described together with the analytic steps for both procedures. Because of the selection it is important to realize the following. When the structure seems dimensional for a subset of indicators, it might in fact be categorical for the total set of indicators, but when the structure seems categorical for a subset of indicators, it follows that it is also categorical for the total set of indicators.

Participants

Data Set 1. 3405 undergraduate students from a U. S. university completed the CES-D in partial fulfillment of course requirements. At the beginning of the semester, participants logged on to a secure website where they completed the CES-D as part of a battery of psychological tests. In response to the general question "Please indicate how often you felt this way during the past week," participants answered the 20 questions that comprise the CES-D using the standard 0-3 response format, where 0 = "Rarely or none of the time (less than one day)" and 3 = "Most or all of the time (5-7 days)". The average score of this sample on the CES-D was 13.75 ($SD = 9.28$).

Data Set 2.[2] The data stem from the participants of an Internet-based smoking cessation study (Muñoz et al., 2006). The participants of this study answered the 20 CES-D items online as part of a series of measurement tools. The study was conducted either in Spanish or in English. In the current study the data of the 3084 English-speaking participants who answered all the twenty CES-D items were used. People in this data set

[2]Data Set 2 was provided by Ricardo F. Muñoz, Ph.D., Professor of Psychology at the University of California, San Francisco and San Francisco General Hospital. Support for the collection of these data was provided by grants 7RT-0057 and 10RT-0326 from the California Tobacco-Related Disease Research Program (Muñoz, P.I.), and by a grant from the University of California Committee on Latino Research to the UCSF/SFGH Latino Mental Health Research Program, which Dr. Muñoz directs.

scored 17.88 on average ($SD = 12.25$).

Data Set 3. The data are from 2340 noninstitutionalized elderly residents ($age > 65$ years) from New Haven as part of the ICPSR EPESE study (Taylor, Wallace, Ostfeld, & Blazer, 1998). A stratified random sample of clusters of households was used to sample the participants. The participants were asked to fill out a test battery including the CES-D Scale. The data from the baseline questionnaire (the questionnaire administered at the beginning of the study) were used in the analyses. The mean CES-D score of this third group was 8.00 ($SD = 8.28$).

The three data sets represent different subpopulations. Data Set 1 consists of young adults; Data Set 2 covers the age span of the whole adulthood, whereas Data Set 3 represents exclusively elderly people. That is, the mean age is increasing throughout these data sets. The mean score increases from Data Set 1 to Data Set 2, but is much lower in Data Set 3 than in the other two.

DIMCAT analyses

The DIMCAT framework (De Boeck et al., 2005) was originally formulated for manifest categories and binary responses. That a category is manifest implies that the category membership is known beforehand, for example, a diagnosis based on expert opinion. However, in the DIMCAT analysis procedure that will be used here, a series of mixture (latent class) models, in particular a version of the two-parameter logistic mixture model (a mixture version of the 2PL; Birnbaum, 1968) and its restricted variants are fitted, and the fit of those models is compared in order to conjecture the structure underlying the data. All these models share the assumption that for heterogeneous classes the latent distribution is a normal distribution (mixture of normals).

Often, there is no advance knowledge of the category membership and it is even unclear whether more than one class is needed. Although individuals can always be assigned to categories, in many cases these categories are arbitrary, and it is not proven whether there is a real categorical entity in the background. Hence, it is a natural extension of the original DIMCAT formulation to incorporate models that can reveal unknown or latent classes present in the data. Another important extension is that models are formulated for ordinal or rating-scale data, because such data are quite common for questionnaires.

The basic model on which this study relies for the extension to latent classes and rating-scale data is the Modified Graded Response Model (MGRM; Muraki, 1990), which is a generalization of the 2PL model for rating-scale data that models the cumulative probability of responding in response category m or higher. Given the cumulative probabilities, the probability of responding in a certain response category is defined as the difference between the upper and lower cumulative logits (e.g., Tuerlinckx & Wang, 2004), or in other words, a slice of the normal distribution defined by the difference between two category cuts. The MGRM is derived from the more general Graded Response Model (GRM; Samejima, 1969), by restricting the distances between response category thresholds to be equal across the items (but see the analysis of Data Set 3 for a relaxation). This restriction is a reasonable one when the indicators have the same response alternatives (Embretson & Reise, 2000). Hence, the locations of the items along an underlying dimension are unambiguous; independent of the category threshold one would look at, the differences between the items (their relational position) remain the same. Here the second (the middle) threshold was used to describe the position of each item, that is, the second threshold is considered to be the location (difficulty) parameter.

Translated into GRM terms, the DIMCAT approach implies the following. As a first step, a latent class model is specified where in each latent class an MGRM model holds. The item discriminations and also the item locations are allowed to vary across classes in this model. The locations refer to the levels and thus to the means, in fact to a function thereof, and the discriminations refer to relevance for an underlying dimension.

In the first model, called the QUAL1&2-HET model, qualitative differences of both kinds (item location inequality and item discrimination inequality across classes) are allowed. QUAL refers to the qualitative differences, 1 refers to the item discrimination inequality and 2 to the item location inequality across categories, respectively, whereas HET means that within-category heterogeneity is assumed in the model.

In the second model, the QUAL1-HET model, the item locations are restricted to be equal across classes, allowing qualitative differences only of an item discrimination nonequivalence type.

In the third model, the QUAL2-HET model, the item locations are restricted to be equal across classes, allowing qualitative differences only of an item location nonequivalence type.

In the fourth model, called the QUAN-HET model, both the item discriminations and the item locations are restricted to be equal across classes, but the location of the classes on the common underlying dimension may differ (if not, there would be only one class). QUAN refers to quantitative differences. As a result of the restrictions in the QUAN-HET model, there is only a quantitative difference between the categories, so that the categories differ only in their overall level. If the sum of the mixture components is not unimodal, then the differences can be considered to be abrupt, as explained earlier.

In all these four models, two heterogeneous latent categories are assumed. QUAL1&2-HET, QUAL1-HET and QUAL2-HET are models with a different dimension within each class, called *dimensional categories* earlier (see Figure 1). QUAN-HET is a model with the same dimension within both classes, defining a *dimensional* type of category-likeness, called *heterogeneous categories on a dimension* earlier (see Figure 1).

In order to disentangle level differences and variance differences from qualitative differences between classes (between mixture components), the following modeling strategy was pursued. The models were fitted with the mean of the person parameter fixed at 0 and the variance of the person parameter fixed at 1. In order to investigate *discrimination differences* between classes a first model was used with the discrimination parameters being freely estimated in one of the classes, whereas in the other class the discrimination parameters were defined as the discrimination parameters in the first class plus an additive constant (the same for all items). In this the discrimination parameters are restricted to be the same across classes except for the additive constant in one class. This additive constant parameter expresses the difference in variance between the two classes in an indirect way. The variance of the person parameter is fixed so that the model is identified. The second model is one with free discriminations in both classes. If this second model fits better than the first, then there exist differences in discrimination beyond a difference in variance.

In order to investigate *location differences* between classes, similarly a first model was used with the location parameters being freely estimated in the first class, whereas the locations in the second class were defined as the location parameters of the first class plus an additive constant (the same for all items) to account for an overall location difference between the conjectured classes. The mean of the person parameter is fixed for identification reasons. The second model is one with free locations (difficulties) in both classes. If this second model fits better than the first, then there exist differences in locations beyond an overall level difference between the classes.

As a second part of the analyses, the homogeneous counterparts of the above described four models, the QUAL1&2-HOM model the QUAL1-HOM model the QUAL1-HOM model and the QUAN-HOM model are considered, where HOM refers to homogeneity within the class (a variance of zero). However, if there is no variance within a class, then there is also no discrimination, so that the QUAL1&2-HOM and the QUAL1-HOM models do not make sense. As a result, there are only two homogeneous models, the

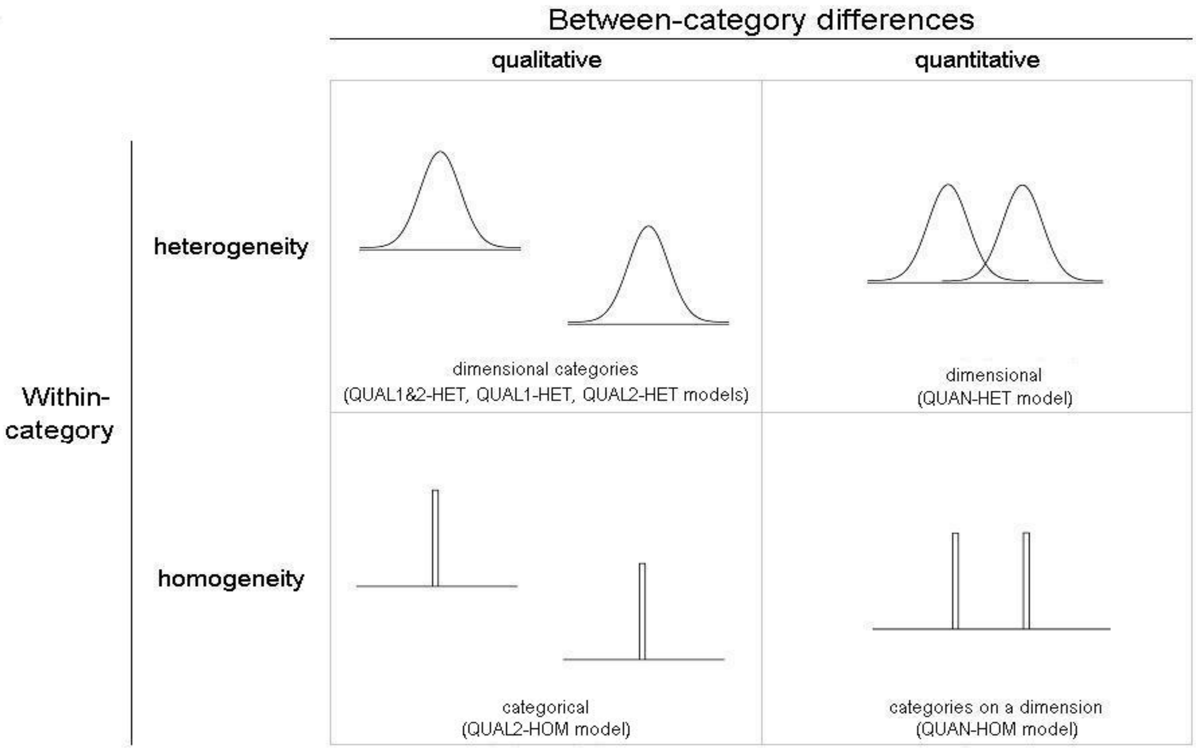

Figure 1: The graphical representation of the various categorical models (assuming two categories).

QUAL2-HOM model which defines *pure categories* and the QUAN-HOM model which defines homogeneous *categories on a dimension* (see Figure 1).

The QUAL1&2-HET, QUAL1-HET, QUAL2-HET, QUAN-HET, QUAL2-HOM and QUAN-HOM models serve as the set of potential category-like structures. The graphical representation of these six models appears in Figure 1.

They are contrasted with a *purely dimensional* model with one single (normal) distribution, so that the categories can be differentiated only by a cutpoint. This seventh model is a model with only one class, called the 2PLGR model. This is the graded response version of the regular 2PL, the same model for the whole data set as used within the classes of the HET-models. The graphical representation of this seventh model appears in Figure 2.

A summary of the basic features of each of the sensible models can be found in

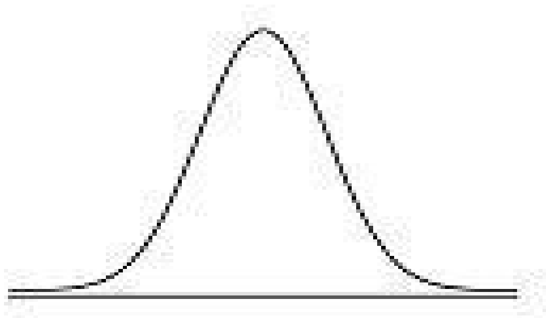

Figure 2: The graphical representation of a purely dimensional model.

Table 1.

In principle, the number of classes can be larger than two, but in the current application of DIMCAT two classes were assumed because of the purpose to study depression in comparison with normality. That is, for this aspect, the approach taken here is rather confirmatory and not exploratory.

Table 1: Summary of the basic features for the DIMCAT models.

Model	Number of classes	Within-class homogeneity	Location equivalence	Discrimination equivalence
QUAL1&2-HET	2	No	No	No
QUAL1-HET	2	No	Yes	No
QUAL2-HET	2	No	No	Yes
QUAN-HET	2	No	Yes	Yes
QUAL2-HOM	2	Yes	No	Yes
QUAN-HOM	2	Yes	Yes	Yes
2PLGR	1	No	Does not apply	Does not apply

Several fit indexes can be used to compare the relative fit of the models described above, of which the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978) were employed. Both of these indexes are based on the deviance (-2*Log Likelihood; -2LL) of the model in question but they differ in how much they penalize for the number of free parameters[3]. The lower the value of these fit indexes is, the better the model fits the data.

The DIMCAT analyses were performed with M*plus*, a flexible general latent variable analysis software (Muthén & Muthén, 2006).

Results

Data Set 1

Of the 20 CES-D indicators, 16 were used in the DIMCAT analyses. The four positive affect indicators were excluded because these were found to form a separate factor. It is rather common to find a separate "method specific" factor whenever positively <u>and</u> negatively worded items are involved (e.g., Rodebaugh et al., 2004; Schmitt & Stuits, 1985; Herche & Engelland, 1996). Positive and negative do not seem to be opposites but rather independent dimensions of affect. Because the focus for depression is on the negative aspects, and because the majority of the items (16 out of 20) loads exclusively on the first factor, we chose to continue with these 16 items.

The seven analysis models described in the DIMCAT analyses section were estimated and their goodness of fit was compared. The goodness of fit indexes for all models appear in Table 2. The eighth model (Saltus) is to be explained later.

The fit of the 2PLGR model was worse than that of the QUAN-HET model following the AIC (100442.15 and 99549.40, respectively) and the BIC (100650.68 and 99776.32, respectively). This means that two categories describe the underlying structure of the data better than a single category. It also means that, at least to some extent, depression does show category-like characteristics when the CES-D indicators are used. As the next

[3] $AIC = -2LL + 2 * npar$, $BIC = -2LL + log(N) * npar$, where $npar$ is the number of estimated parameters, and N is the sample size.

Table 2: Goodness of fit indexes of the DIMCAT models for Data Set 1.

Model	AIC	BIC
QUAL1&2-HET	98228.26	98639.17
QUAL1-HET	99139.96	99458.88
QUAL2-HET	98444.42	98763.34
QUAN-HET	99549.4	99776.32
QUAL2-HOM	102866.68	103179.46
QUAN-HOM	103403.88	103630.8
2PLGR	100442.15	100650.68
Saltus	98302.5	98627.55

step, the QUAL1-HET and the QUAL2-HET models were estimated to test if discrimination equivalence and location equivalence hold across the categories. The QUAL1-HET model fits the data better (AIC = 99139.96, BIC = 99458.88) than the QUAN-HET model, meaning that the item discriminations are different across the classes. It was investigated whether this could be explained by a difference in variance, using the method explained earlier, and the answer was no. Likewise, the QUAL2-HET model yielded a better model fit than the QUAN-HET model, that is, the locations differ across categories. It was investigated whether this could be explained by a difference in the overall level of the categories, using the method explained earlier, but the answer was again no. From these fit indexes one can see that the discriminations have a smaller influence on the model fit: when the locations are allowed to vary across classes, the improvement of the model fit is more substantial than in case the discriminations are allowed to vary across classes. When estimating the QUAL1&2-HET model it turns out to be the best fitting one (AIC = 98228.26, BIC = 98639.17). From the analyses with the homogeneous models, it was established that the models *with* systematic within-category variance do fit the data clearly better than the models assuming homogeneity within the categories. Therefore, the conclusion is that different categories exist with a different dimension in each of both categories and that the dimensions differ in the location and the discrimination of the indicators.

A crucial aspect of a latent class analysis is the interpretation of the classes. It is expected that the classes represent different levels of depression, or different variants with or without a difference in intensity.

In order to check possible level differences, apart from the earlier model comparison, a cutpoint of 20 was used to assign all cases to one of two categories based on their total scores of the CES-D items. Radloff (1977) suggested 16 as a cutpoint to distinguish depressed and nondepressed cases, but the cutpoint of 16 was found to be too low, and cutpoints of 24 and even 27 were suggested instead based on receiver operating curves to find a balance between sensitivity and specificity (e.g., Gotlib, Lewinsohn, & Seeley, 1991). In the current paper, the cutpoint of 20 was chosen as a compromise. According to this assignment 2734 participants (80.3%) had to be categorized as nondepressed and 671 participants (19.7%) had to be categorized as depressed. The class size estimates in the latent class DIMCAT analysis for the best-fitting model are 2708 (79.5%) and 697 (20.5%), which is close to the category sizes of the manifest categories based on the cutpoint of 20 for the total scores. Following a classification on the basis of the model estimation, of the 2734 persons with a total score of 20 or lower, 92.68% was assigned to the majority latent class, and of the 671 persons with a total score higher than 20, 75.07% was assigned to the minority latent class. The corresponding κ coefficient is .66. It may be concluded that the two latent classes correspond to a minority class of depressed and a majority class of nondepressed, respectively.

As for the *indicator locations* within the two classes, a subset of four indicators (nrs. 3, 6, 14, 18) shows clearly larger differences across the categories as compared to the remaining of the indicators. The item locations of the two classes are depicted in Figure

3.

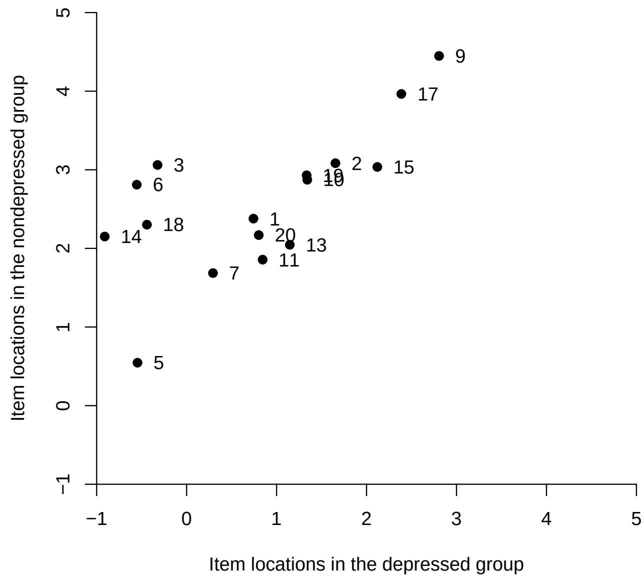

Figure 3: Scatter plot of the item locations for Data Set 1.

The four indicators in question are: "I could not shake off the blues" (3), "I felt depressed" (6), "I felt lonely" (14), "I felt sad" (18). These are four of the five depressed affect indicators. All these four indicators have to do with the affective component of depression. This pattern with the four indicators being more prevalent in the depression class in comparison with other indicators suggests that a single parameter can account for the relative prevalence pattern difference between the categories. Therefore, a new model was estimated with one set of location parameters for both classes, and with two level difference parameters, a first one for the whole set of items and a second one (called the *Saltus* parameter) for the four items in question instead of a whole new set of indicator location parameters in the second class. The corresponding model is called Saltus (Wilson, 1989) referring to the jump of a subset of items when going from one class to another. The AIC and the BIC values of the Saltus model were 98302.5 and 98627.55, respectively (see Table 2). Compared to the thus far best fitting QUAL1&2-HET model, it is not clear what the best fitting model is. When using the AIC, one must conclude that the

QUAL1&2-HET model fits better, whereas based on the BIC, the Saltus model should be preferred. This ambiguous result is due to the fact that the BIC penalizes more for the number of parameters, or in other words stated, the BIC prefers the more parsimonious models. The Saltus parameter for the four indicators is significantly different from zero when estimated as a separate parameter (1.82, $t = 17.7$, $p < .01$). The Saltus parameter is of a kind that the four indicators show a 'jump' towards a relatively higher prevalence (compared to other items) when moving from the nondepressed class to the depressed class. The participants belonging to the depressed class score higher on the indicators in general, but for the four depressed affect indicators they score even higher. The depression category is characterized not just by more of the same or by a higher position on an underlying dimension, but also by a change in pattern, with a relatively higher prevalence and hence, a stronger pronunciation of depressed affect.

When further exploring the indicator discriminations in the two categories, it is clear that three items (three of the four items that show a sudden jump when moving from the nondepressed category to the depressed category) have a substantially higher item discrimination than the remaining items. The scatter plot of the item discriminations appears in Figure 4. This means that these items differentiate the individuals better than the other items in both classes; these items are more relevant for the underlying construct (or constructs, because of the difference between the two classes).

It is not clear, though, whether and how the item discriminations differ between the two classes, except that the discriminations in the nondepressed category are more heterogeneous than the discriminations in the depressed category. From the inspection of the scatter plot in Figure 4 it appears that the relationship between the discriminations in the two classes may differ by a multiplicative constant. Such a difference is also more in line with the multiplicative nature of the discrimination parameter. Hence, an adapted Saltus model was set up which differs from the original one in that a multiplicative

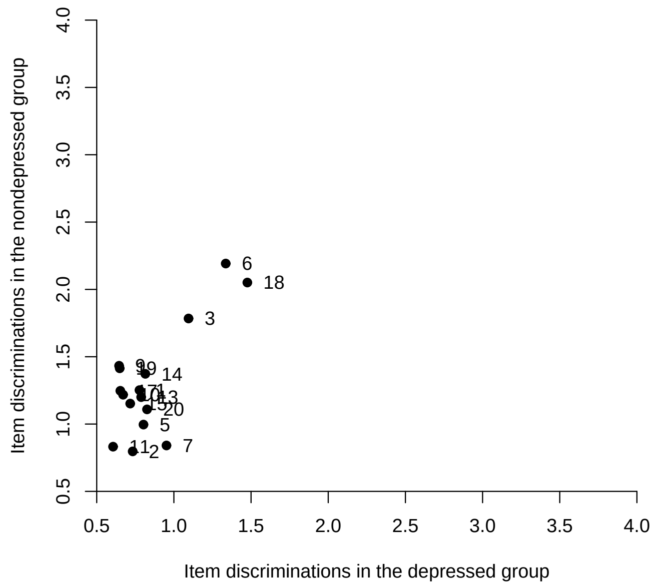

Figure 4: Scatter plot of the item discriminations for Data Set 1.

constant was used for the discriminations in the second class instead of a whole new set of discriminations. The AIC and the BIC values are 98541.14 and 98774.2, respectively, which is inferior to the corresponding values of the QUAL1&2-HET model. Because item 7 and to some degree also item 2 seem to deviate, an extended model was estimated next, one with two multiplicative constants, one for all items except 7 and 2, and one for the latter two. The goodness of fit indexes of this model with the two multiplicative discrimination difference parameters are 98411.765 and 98650.95 for the AIC and the BIC, respectively. Although the BIC value of this model is very close to that of the original Saltus model, the latter still fits the data better. The values of the multiplicative constants are 0.59 ($t = 64.99$, $p < .01$) for the majority of the items and 1 ($t = 23.484$, $p < .01$) for the items 2 and 7. This means that except for the items 2 and 7 the variance in the depressed class is smaller than the variance in the nondepressed class. Depressed people seem more alike than the nondepressed, which makes sense if depression is a category. Although the original Saltus model with a nonequivalence of the discriminations has a better goodness of fit, also the multiplicative model is considered a reasonable model for the data, for two reasons. First, it is the better fitting model for Data Sets 2 and 3, and second, it is a simpler model in that the differences in discriminations can be explained away (except for items 2 and 7) as a difference in variance (as expressed in the multiplicative constant).

Data Set 2

In the second study, the CES-D indicators, again except for the four reverse-coded positive affect indicators, were submitted to a DIMCAT analysis. To start with, the goodness of fit indexes of the 2PLGR and the QUAN-HET model were compared (see Table 3). Like in case of the first data set, it was found that the QUAN-HET model yielded a significantly better model fit than the 2PLGR model. The respective goodness of fit statistics were: 96037.1 and 96697.59 for the AIC, and 96260.36 and 96902.74 for the BIC.

Table 3: Goodness of fit indexes of the DIMCAT models for Data Set 2.

Model	AIC	BIC
QUAL1&2-HET	95318.39	95722.67
QUAL1-HET	95818.2	96131.97
QUAL2-HET	95470.16	95783.92
QUAN-HET	96037.1	96260.36
QUAL2-HOM	101741.29	102049.02
QUAN-HOM	102086.49	102309.74
2PLGR	96697.59	96902.74
Saltus	95527.07	95846.87

These fit values are indicative of a categorical underlying phenomenon, at least to some extent because the QUAN-HET model is a categorical model only in a very limited sense. When testing for qualitative between-class differences it was established that the QUAL1-HET model yields a better model fit ($AIC = 95818.2$, $BIC = 96131.97$) than the

QUAN-HET model, meaning that there is discrimination nonequivalence. It was found that the nonequivalence could not be attributed to differences in variance when modeled with an additive constant. It was also found that indicator location equivalence did not hold across classes, as the QUAL2-HET model also fits better ($AIC = 95470.16$, $BIC = 95783.92$) than the QUAN-HET model and this could not be attributed to an overall difference in level. When both types of qualitative differences were allowed (QUAL1&2-HET model), an even better fit ($AIC = 95318.39$, $BIC = 95722.67$) was obtained. The estimation of the homogeneous models resulted in worse model fit values. As for Data Set 1, the conclusion is that there are two categories, with a different dimension in each category and that the dimensions differ in the location and the discrimination of the indicators.

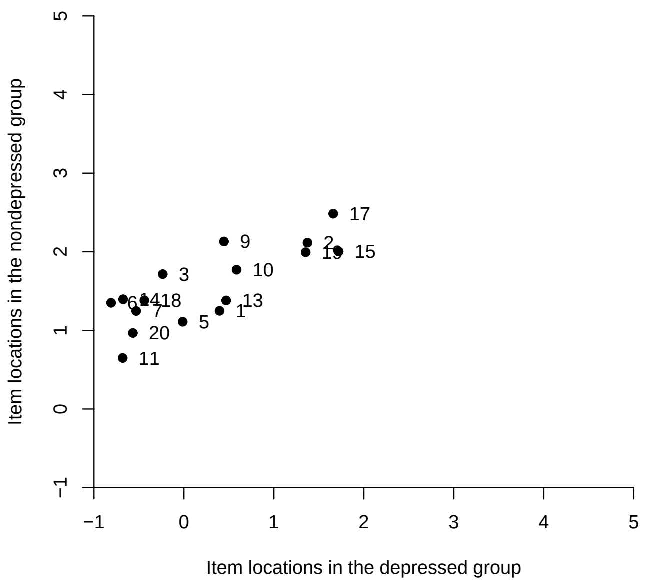

Figure 5: Scatter plot of the item locations for Data Set 2.

In order to interpret the latent classes, the same method was used as for Data Set 1. The participants were assigned to a nondepressed class (with a total score of less or equal to 20) and a depressed class (with a total score of more than 20). According to this assignment 1934 of the 3084 participants (62.71%) had to be categorized as nondepressed and 1150 participants (37.29%) had to be categorized as depressed. The class size estimates in the latent class DIMCAT analysis for the best-fitting model are 2251 (72.99%) and 833 (27.01%), which is different but not largely different from the category sizes of the manifest categories based on the cutpoint of 20 for the total scores. Following a classification on the basis of the model estimation, of the 1934 persons with a total score of 20 or lower, 91.69% was assigned to the majority latent class, and of the 1150 persons with a total score higher than 20, 52.17% was assigned to the minority latent class. The corresponding κ value is .49, so that again, the two classes can be interpreted in terms of nondepressed and depressed, respectively.

Next, the same Saltus model was fitted as in Data Set 1, with a Saltus parameter for the items 3, 6, 14, and 18 as for Data Set 1, although the plot in Figure 5 does not suggest a difference in location for these items. The fit of the Saltus model was better than the fit of any other analysis model except for the QUAL1&2-HET model (AIC = 95527.07, BIC = 95846.87). However, the Saltus parameter was still significant (0.92, t = 10.37, $p < .01$) suggesting that in case of these four indicators there may be a difference in prevalence between the depressed and nondepressed class, although this difference is smaller than the difference found in Data Set 1. A further finding of the inspection of the item locations in the two classes, is that the indicator locations are much more spread out in the depressed class, that is, depressed individuals differentiate the item locations more than nondepressed people do. Among the nondepressed, the prevalence of the symptoms (expressed by the item location) is more homogeneous. The item locations are shown in Figure 5.

As far as the item discriminations are concerned, the same pattern can be seen as in case of the first data set. The same three items show a substantially higher discrimination, and the range of the discriminations is higher in the nondepressed class (Figure 6). Again, the difference of the discriminations in the two classes appears to be rather of a multiplicative type than of an additive type. For this reason, the Saltus model was repeated but with a multiplicative constant for the discriminations in the second class instead of a new set of discrimination parameters. The fit of this new model (AIC = 95604.44, BIC = 95833.73) was better than the fit of the Saltus model when the BIC is used as a criterion, suggesting, that the differences in the item discriminations can be explained away by differences in variance, accounted for by the multiplicative discrimination constant. The value of the multiplicative constant was 0.57 (t = 70.24, $p < .01$) which is very similar to the value that was found in case of the first data set. As expected, the variance is substantially lower in the depressed class than in the nondepressed class. For Data Set 2, there seem to be no exceptions for some of the items, contrary to the results for Data Set 1.

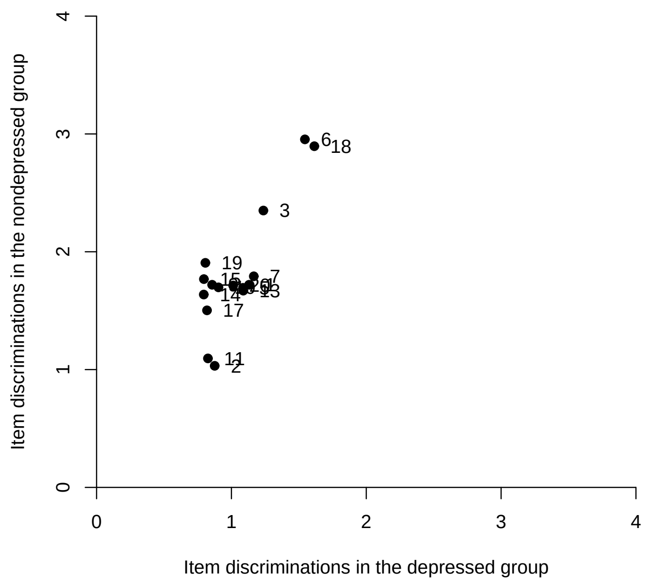

Figure 6: Scatter plot of the item discriminations for Data Set 2.

In sum, the conclusions established for Data Set 1 were replicated for Data Set 2, although not as clear for the correspondence with the cut-off classes and the original Saltus model, but in a clearer way for the differences in discriminations being explained away in the multiplicative Saltus model. Again, the depressed group seems more homogeneous than the nondepressed.

Data Set 3

In the DIMCAT analysis of Data Set 3 the same data analysis strategy was followed as for the previous two data sets. First, the model fit of the 2PLGR model was compared to that of the QUAN-HET model. Both the AIC and the BIC were found to be smaller for the QUAN-HET model (47025.85 and 47238.89, respectively) than for the 2PLGR model (47373.47 and 47569.24, respectively). The analyses were continued allowing item discrimination differences across the latent classes (QUAL1-HET model) which turned out to be a better fitting model ($AIC = 46932.78$, $BIC = 47116.77$) than the one with equal discriminations. The model with item location differences across the latent classes also yielded a better model fit than the QUAN-HET model based on both the AIC (46817.36) and the BIC (47116.77). When both item discrimination differences and item location differences were included in the analysis (QUAL1&2-HET model), it was found to be the best fitting model as far as the AIC is concerned, whereas the QUAL2-HET is the best fitting as far as the BIC is concerned ($AIC = 46795$, $BIC = 47180.78$). The homogeneous models yielded worse fit statistics. Hence, based on these results, the conclusion can be the same as for the other two data sets. The summary of the fit indexes appear in Table 4. The majority class size is 1586, while the minority class size is 754.

Assuming the QUAL1&2-HET model as the best fitting one (the item locations and item discriminations appear in Figures 7 and 8, respectively), the interpretation of the resulting classes is not as straightforward as in the previous two studies (and a similar result was obtained with the QUAL2-HET model). The same interpretation of the classes did not work for Data Set 3. The class assignments based on the latent class analysis and the cutpoint of 20 showed no correspondence at all, with a κ value as low as .02. Hence, it can be concluded that the two classes found in the latent class DIMCAT analyses have no link with severity of depression. This finding is not very surprising taking into account

Table 4: Goodness of fit indexes of the DIMCAT models for Data Set 3.

Model	AIC	BIC
QUAL1&2-HET	46795	47180.78
QUAL1-HET	46932.78	47232.19
QUAL2-HET	46817.36	47116.77
QUAN-HET	47025.85	47238.89
QUAL2-HOM	48533.59	48827.25
QUAN-HOM	48578.65	48791.69
2PLGR	47373.47	47569.24
Saltus	46879.21	47184.38

the mean value (8.00) and the standard deviation (8.28) of the CES-D scores in this sample. These values show that people in the elderly sample have a substantially lower score on average than people in the other two samples, and also there is a lower variance in this sample. A substantially lower proportion of this sample scores above the cutpoint of 20, than in the other two data sets. It seems that the prevalence of depression in the elderly sample is very low. When, for the sake of completeness, the Saltus model was estimated nevertheless, it turned out to be inferior to both the QUAL1&2-HET and the QUAL2-HET models on both the AIC and the BIC.

From an inspection of the proportions of the rating-scale responses, the two classes resulting from the latent class DIMCAT analyses seem to represent more a specific response style than depression. The response style in question can be interpreted as a one-sided extremity response, with a relatively higher tendency in the first class than in the second to respond in the extreme category on the higher end of the scale, and less so in the second highest.

In order to investigate the response style interpretation, an additional model was

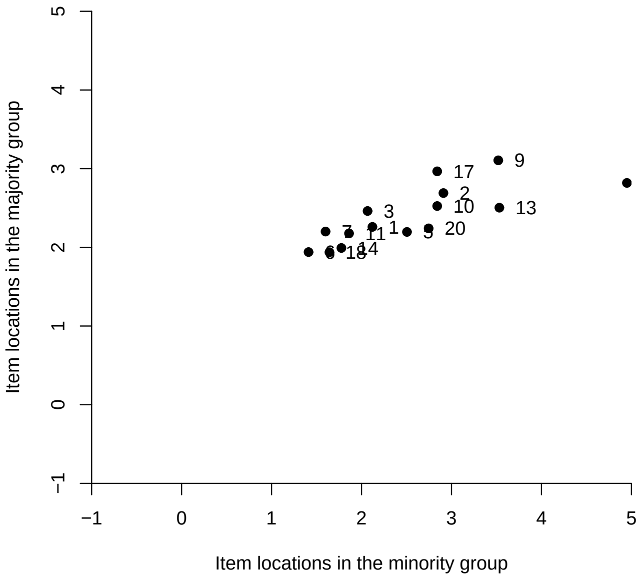

Figure 7: Scatter plot of the item locations for Data Set 3.

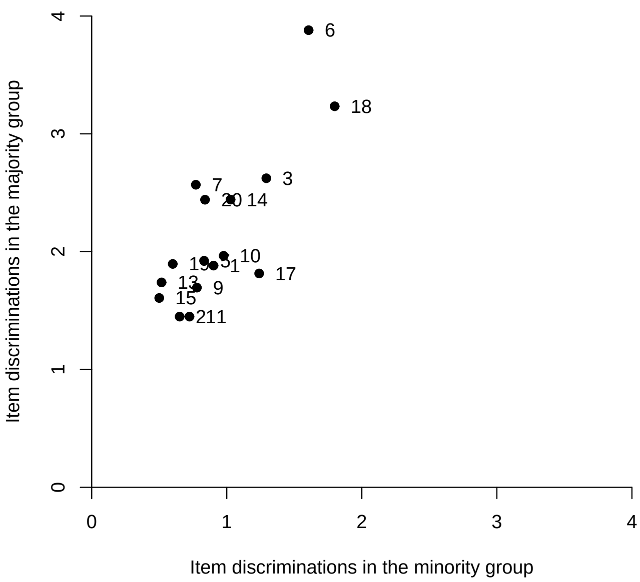

Figure 8: Scatter plot of the item discriminations for Data Set 3.

fitted. In this model, the distances between the response category thresholds are not class invariant but class specific instead, while the locations and the discriminations are the same for the two classes. This model is called the QUAL3-HET model because the class differences concern a third kind of parameter, the response category thresholds. This new model yields by far the best model fit ($AIC = 46724.9$ and $BIC = 46949.46$) in comparison with the other models. Because the response scale has four response categories, there are three category thresholds, a threshold between 0 and 1, between 1 and 2, and between 2 and 3. Upon an inspection of the threshold estimates, it turned out that the distances between the category thresholds are larger in the majority class than in the minority class. This can perhaps be explained by a difference in variance, but on the other hand, the second distance between the 1/2 threshold and the 2/3 threshold, is relatively smaller in the majority class than in the minority class (.6/1.8 versus .3/.7). This corroborates the response style interpretation. It was also investigated for this latter model if the resulting latent classes correspond to depression severity, but it was found that the classes showing different response styles do not show any relationship with depression.

As it may happen that the difference in response style is more salient in this sample than a difference in depression, the depression class may perhaps appear only as a third class. For this reason also a 3-class latent class model (QUAL1&2-HET) was estimated, but no convergence of the estimation procedure was reached.

Discussion

In two out of the three studies, a class of depressed respondents was found, and in both cases the structure turned out to be of the *dimensional categories* type (upper left in Figure 1). Depression, as measured by the CES-D, seems to be a category, but one with a dimension within the category expressing systematic differences in the degree of being depressed, although the quality of being depressed is different depending on the category.

In the third sample, no evidence was found for a depressed category, perhaps because the respondents were not depressed or have not expressed much depression.

One should realize that depression, as measured here, is primarily a state variable given the instructions of the CES-D and that most likely only a mild form of depression is identified in the analysis, although a systematic variance was found in the depressed class and thus also more extreme forms may be part of the class. Perhaps depression as investigated here, is rather a mood than a psychopathological syndrome. The findings contradict in some way the dominant view that depression is dimension-like. On the one hand, depression is dimensional in that depressed persons have a higher level of depression symptoms, and in that they differ within the category of the depressed with respect to how much depressed they are (a dimension within the category). On the other hand, depression is categorical because the degree of being depressed is qualitatively different for the depressed and the nondepressed. Being depressed appears to be primarily a matter of affect. The four indicators ("I could not shake off the blues", "I felt depressed", "I felt lonely", "I felt sad") are relatively more characteristic of the depressed than of the nondepressed. This is not a surprising result, but it has not been shown before in studies on the categorical/dimensional nature of depression.

In previous studies on the categorical or dimensional nature of depression, one has not used the CES-D. It is evident that the nature of the structure (categorical or dimensional) may depend on the indicators under consideration. For example, if the depressed affect items had not been included, then the conclusion would have been that depression is dimensional. However, the conclusions are in line indeed with the conclusions of Santor and Coyne (2001). There are several differences between the methodologies applied in the current paper and in Santor and Coyne's paper (parametric vs. nonparametric IRT, latent vs. manifest classes, CES-D vs. HRSD[4]), but the basic findings in both studies are that (1) symptom severity is not sufficient to account for the differences between de-

[4]Hamilton Rating Scale for Depression (Hamilton, 1960)

pressed and nondepressed individuals, and that (2) some symptoms behave in a different way among depressed persons, meaning that the measure under investigation measures a different construct among depressed and nondepressed. The differently behaving symptoms in Santor and Coyne's study included core symptoms of depression such as depressed mood and anhedonia/impairment, a result which is similar to the results of the current study, where also the depressed affect indicators (which are core symptoms of depression) formed a differently behaving group of indicators. In addition, Santor and Coyne found a number of other indicators that behaved differently, including for instance suicide, middle insomnia and hypochondriasis.

Because both, the inventory (CES-D) and the method (DIMCAT) differ from previous studies, the result may be due to either of both. The Santor and Coyne (2001) results may be considered as convergent evidence obtained with a different inventory, but a conclusion regarding DIMCAT would be stronger if also the taxometric methods would be applied to the CES-D data. If the result is dimensional, then it must be concluded that the difference with most previous studies is caused by a difference in methodology. Then it is clear that not the inventory is the cause but that a different method, one which includes various aspects of category-likeness, is therefore able to find categorical features in the structure which tend to go undetected by taxometric methods. If the results of the taxometric methods is categorical, then it must be concluded that the difference is due to a different inventory being used and that both approaches converge. A more definite conclusion would require that DIMCAT is used also for the other inventories taxometric methods have been used for. However, the present study concentrates on the CES-D, and also the taxometric methods will be used in the following, in order to understand better the obtained results.

Taxometric analyses

The taxometric approach consists of a family of methods that all aim at revealing the underlying categorical vs. dimensional nature of different phenomena. The most well-known of these methods are MAXCOV and MAMBAC. Each of the taxometric procedures require the variables to fulfill certain requirements. Based on these requirements, indicators are selected for the application of the procedures.

Indicator selection

The most important requirements are indicator validity and within-class indicator correlation. Indicator validity expresses how well a certain indicator separates the two classes. The indicator validity or indicator separation is generally defined in taxometrics as *Cohen's d* statistic. Meehl (1995) recommended $d = 1.25$ as a threshold of indicator validity. As for within-class indicator correlation he suggested .30 as an upper limit. To calculate either indicator validity or within-class indicator correlation, the cases have to be assigned to putative classes. In this paper the procedure was as follows. First, indicator validity was calculated using both the original cutpoint of 16 and the increased cutpoint of 20, and the one that yielded larger indicator separation was used for the selection of the appropriate indicators. Second, the within-class indicator correlations were investigated using only the indicators that showed sufficient indicator separation.

There were two types of indicator variables employed in the taxometric analyses in the present paper. On the one hand, the original *CES-D items* that proved to have the desirable indicator properties (indicator validity and intercorrelation) were used without any modification. The specific items selected using this indicator selection procedure will be reported in the Results section for all the three data sets separately. On the other hand, *composite scores* were used that were created by summing the CES-D indicator variables with a similar content. Three such composite indicator variables were created (corresponding to the item groups described by Radloff (1977), except for the positive affect item group): the depressed affect indicators ("I felt that I could not shake off blues even with the help from my family or friends", "I felt depressed", "I felt lonely", "I had crying spells", "I felt sad"); the somatic and retarded activity indicators ("I was bothered by things that usually don't bother me", "My appetite was poor", "I felt that everything I did was an effort", "My sleep was restless", "I talked less than usual", "I could not get going"); and the interpersonal indicators ("People were unfriendly", "I felt that people dislike me"). The composite indicator variables created this way were also subject to criteria of indicator validity and intercorrelation. As the positive affect items of the CES-D were not employed in the DIMCAT analyses, they were also excluded from the taxometric analyses.

Two taxometric procedures, MAXCOV and L-Mode, were applied and will be described briefly. In fact, a third taxometric procedure, MAMBAC (Mean Above Minus Below A Cut; Meehl & Yonce, 1996) was also applied, but because the results of the MAMBAC analyses were very similar to those of the MAXCOV analyses (providing no further information) MAMBAC is not included in the manuscript.[5] The use of multiple (at least two) taxometric procedures is a very important feature of the taxometric approach, as the taxometric methodology does not rely on significance tests or information criteria but on consistency checks such as the consistency of the results across different taxometric procedures.

MAXCOV

In MAXCOV (MAXimum COVariance; Meehl, 1973; Meehl & Yonce, 1996), three indicator variables are required: one input indicator and two output indicators. The cases are sorted into subsequent sections along the input indicator and in each section the covariance of the two output indicators is calculated and plotted. In case of a taxonic

[5]The MAMBAC results are available upon request from the first author.

underlying structure, the resulting MAXCOV graph is expected to show a clear peak, in contrast with a flat MAXCOV graph to be expected when the underlying structure is dimensional. When there are more than three appropriate indicator variables present, MAXCOV can be repeated with all possible triplets of indicators.

It may be a problem for MAXCOV to use the items as input indicators because the rating scales have only four points so that only four intervals are available for the covariance plot. To overcome this difficulty the following modification of the original MAXCOV procedure is applied when using the CES-D items as indicators: two items are assigned the role of output indicators and the remaining items are summed to form a *summed input indicator* which has more points than just one item to sort the cases.

In all cases, overlapping intervals of the input indicator were used as described by Waller and Meehl (1998) for the MAXEIG (MAXimum EIGenvalue) method (that can be considered as the multivariate generalization of the MAXCOV). The overlapping windows are obtained by taking a group of cases with the lowest value(s) on the input indicator variable, then another group of cases overlapping with the previous group to a predefined extent, and so on. The advantage of using overlapping windows to sort the cases is that many more intervals of the input indicator are obtained, and the resulting graph is easier to interpret without using additional smoothing techniques. Hence, the potential problem of smoothing down a genuine peak when using smoothing techniques can be eliminated (Ruscio, 2000).

L-Mode

L-Mode (Latent Mode; Waller & Meehl, 1998) is a relatively new and simple procedure. In L-Mode a common factor analysis is performed using all the available indicator variables and the factor scores of the first factor are plotted. When there is an underlying categorical structure, the resulting L-Mode plot of factor scores shows bimodality; in case of an underlying dimensional structure, the plot of the factor scores shows unimodality. For the L-Mode, no modification of the procedure is required using either of the indicator types described earlier. For the L-Mode, all items can be used without further restrictions given that the first factor is a dominant factor (as for our data).

Using empirical sampling distributions to help interpret taxometric results

The correct and objective interpretation of taxometric graphs often proves to be fairly difficult, as the resulting graphs are often not clearly indicative of either a taxonic or a dimensional underlying structure. The interpretation, however, can be aided with the help of simulating artificial data (J. Ruscio, Ruscio, & Meron, in press), that very closely model the properties of the empirical data under investigation. Two types of data sets, a taxonic and a dimensional, are simulated in a way to reproduce the important distribution properties (e.g., mean, variance, skew, and correlation of the indicators) of the original data with a high precision. These two kinds of data sets, with known underlying structure are then subjected to the same taxometric analyses as the empirical data, hence, the results obtained using the empirical data can be compared to the results obtained using simulated data sets which have the same properties the empirical data have, but the underlying structure is known.

The results of the taxometric analyses of the simulated comparison data aid the interpretation of the results of the empirical data in two ways. First, the taxometric graphs of the empirical data and the taxometric graphs of the simulated taxonic and simulated dimensional data can be compared by visual inspection. Second, an objective measure can be applied, the Comparison Curve Fit Index (CCFI)[6], which shows which type of simulated data graph is closer to the empirical data graph. The CCFI value ranges from 0 to 1, where 0 means a dimensional structure, 1 indicates a taxonic structure, whereas

[6] $CCFI = \frac{Fit_{dim}}{Fit_{dim}+Fit_{tax}}$, with $Fit = \sqrt{\frac{\sum_{n=1}^{N}(y_{ne}-y_{ns})^2}{N}}$, where y_{ne} and y_{se} denote the nth data point in the graph of the empirical data and the simulated data, respectively, and N is the number of data points on either of the graphs.

.5 means equal support for both structures (J. Ruscio, Ruscio, & Meron, in press).

The taxometric analyses were carried out using John Ruscio's (2006) taxometric package which offers a flexible way to set up several taxometric methods and offers the possibility to utilize simulated comparison data to help interpret the results. Unfortunately, according to John Ruscio (personal communication, October 18, 2006), the CCFI does not work well in case of the L-Mode procedure and for that reason it is not included in his taxometric program code.

Results

Data Set 1

Indicators

CES-D items. As the first step, indicator variables were selected following the procedure described above. The indicator validities were larger for all indicator variables when the respondents were assigned to putative classes using a cutpoint of 20. Of the 20 CES-D indicator variables five ("I could not shake off the blues", "I felt depressed", "I felt lonely", "I had crying spells", "I felt sad", "I could not get going") turned out to have an indicator separation above the suggested threshold of 1.25. All of these five are depressed affect indicators. Unfortunately, the indicators chosen this way correlate with each other to a relatively high degree within the putative classes, reaching correlations as high as .61. Because of the high within-class correlations these indicator variables are not appropriate for taxometric analyses; therefore, the items were not used as indicator variables.

Composite scores. Composite scores of the original CES-D indicator variables were formed in a way described above, and their appropriateness for taxometric analyses was investigated. Indicator separations again were larger when using a cutpoint of 20, and exceeded 1.25 for all three composite scores. The within-class correlations were moderate and fell well below the suggested .30 limit. Hence, the composite indicators could be considered as appropriate for taxometric analyses.

MAXCOV

Since the CES-D items seemed to be inappropriate for taxometric analyses, the analyses were performed using only the composite scores. The MAXCOV graphs are depicted in Figure 9. For the sake of simplicity only averaged graphs (averaged over the input, output, output triplets) are presented, and this is the case also for all further taxometric analyses. The individual graphs do confirm the conclusions to be drawn from the averaged graphs.

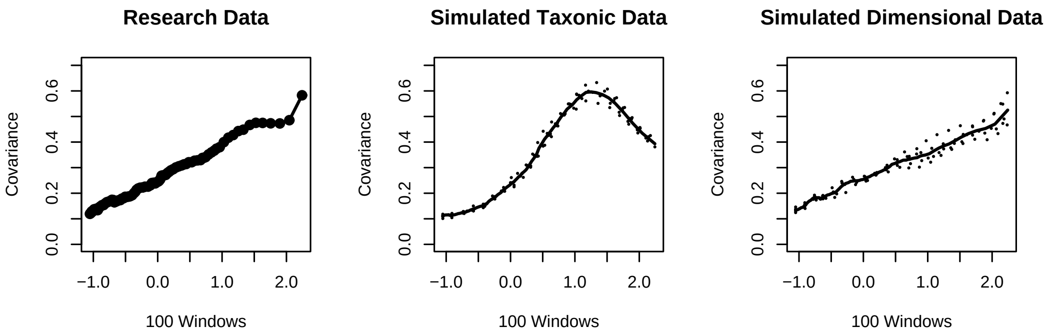

Figure 9: MAXCOV curves for Data Set 1 (left panel), simulated taxonic data (middle panel) and simulated dimensional data (right panel).

The plot of covariances of the empirical data shows a marked elevation on the right side of the graph, but does not show a genuine peak. This result can have two explanations. First, this marked elevation on the right side of the curve may suggest a taxonic structure with a very low base rate. Second, it may be the consequence of a skewed distribution in all the (composite) indicator variables. When the indicators are positively skewed, the expected change of the expected flat curve (expected when the structure is dimensional) is that the right tail of the MAXCOV curve is shifted upwards. This is exactly the kind of function that can be seen in Figure 9. Since all the CES-D items as well as the composite scores are positively skewed (the mean skew is .88 and .90, respectively), the second explanation seems to be a reasonable one.

The inspection of the MAXCOV curves of the simulated taxonic and the simulated dimensional data (middle panel and right panel of Figure 9, respectively), shows that the structure is dimensional, since the simulated taxonic data show a clearly peaked curve, whereas the simulated dimensional data show an upward sloped curve, very similar to the curve resulting from the actual empirical data ($CCFI = .18$). Based on these MAXCOV results it can be concluded that depression appears as a continuous phenomenon rather than as a separate category.

L-Mode

When evaluating the L-Mode graph of the empirical data (depicted in the left panel of Figure 10) it can be concluded that it is unimodal, suggesting a dimensional underlying structure. As for MAXCOV, the plotted factor scores show a different picture in case of the simulated taxonic data as compared to the simulated dimensional data (shown in the middle and right panel of Figure 10, respectively). The plot of the empirical data is most similar to that of the simulated dimensional data. For reasons explained earlier, the CCFI was not used.

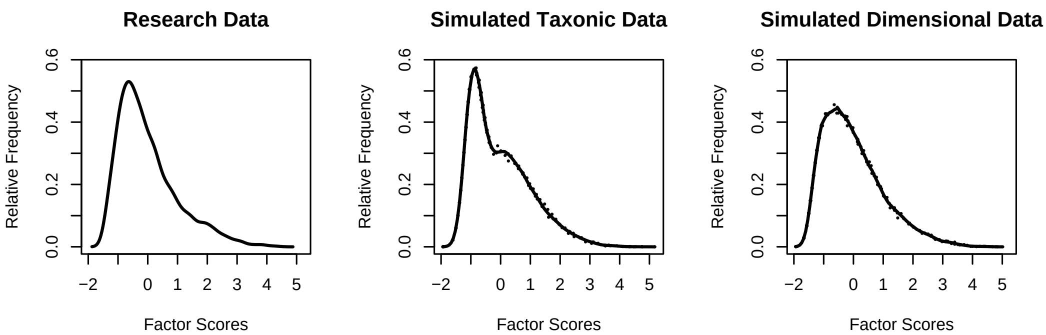

Figure 10: L-Mode plots for Data Set 1 (left panel), simulated taxonic data (middle panel) and simulated dimensional data (right panel).

Data Set 2

Indicators

CES-D items. Of the sixteen CES-D items considered, here seven had an indicator separation of 1.25 (defined as Cohen's d) or larger, four of which belong to the depressed affect group. For these seven indicator variables, within-class correlations were derived (within the putative classes based on a cutpoint of 20). It was found that correlations above the .30 threshold could be avoided if three indicators were excluded from the taxometric analyses. The four remaining items used as indicators were: "I could not shake off the blues", "I thought my life had been a failure", "I felt lonely", and "I could not get going".

Composite scores. The same three composite scores were created in the same manner as in Data Set 1. All three composite scores had sufficient indicator separation, and the within-class indicator correlations were below .30, and thus acceptable.

MAXCOV

CES-D items. As mentioned earlier, when employing CES-D items, composite input indicators were used to overcome the problem of the four-point rating scale being too short to reliably sort cases. In the averaged covariance plot (shown in the left side of Figure 11) no clear peak is present, which implies that there is no underlying taxon, and that depression is a dimensional construct. This finding is corroborated by the covariance plots of the simulated taxonic and simulated dimensional data sets (depicted in the middle and right panel of Figure 11, respectively), of which the plot of the simulated dimensional data is considerably closer to that of the empirical data than the plot of the simulated taxonic data ($CCFI = .16$). The right-end elevation in the plot of the original data can be attributed again to the positively skewed indicators.

Composite scores. The findings obtained with the CES-D items were replicated with the composite CES-D scores as well. The averaged plots of the actual empirical data as well

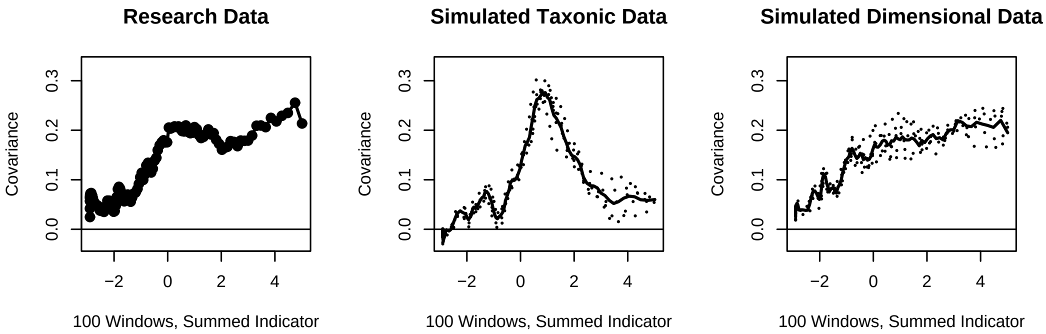

Figure 11: MAXCOV graphs for Data Set 2 (left panel), simulated taxonic data (middle panel) and simulated dimensional data (right panel), using CES-D items.

as the simulated taxonic and simulated dimensional data (Figure 12) are very close to their counterparts stemming from the analyses of the original CES-D items, and again the dimensional nature of the structure is confirmed ($CCFI = .24$).

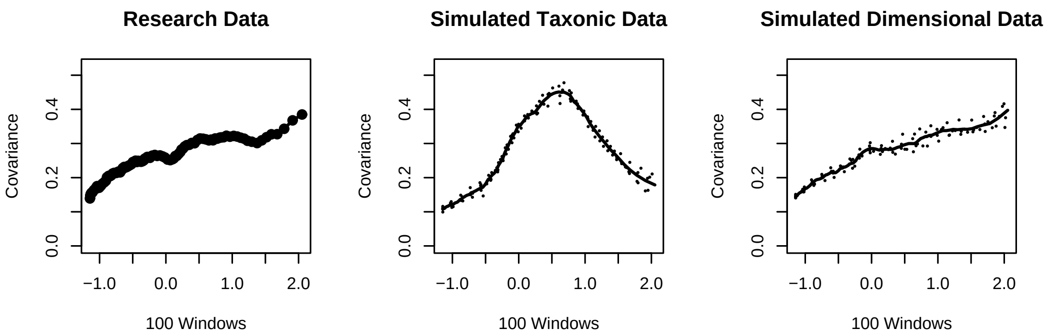

Figure 12: MAXCOV graphs for Data Set 2 (left panel), simulated taxonic data (middle panel) and simulated dimensional data (right panel), using composite cores.

L-Mode

CES-D items. The plotted factor scores of the L-Mode analysis show a unimodal distribution (left panel of Figure 13), which is suggestive of a dimensional underlying construct, similarly to the finding obtained with the MAXCOV analyses. The distribution of the factor scores is markedly positively skewed as the indicators are. The finding of a dimensional underlying construct is corroborated again by the simulated comparison data, of which the plotted factor scores of the simulated dimensional data are nearly identical with those of the empirical data, whereas the plot of the simulated taxonic data shows a modest bimodality.

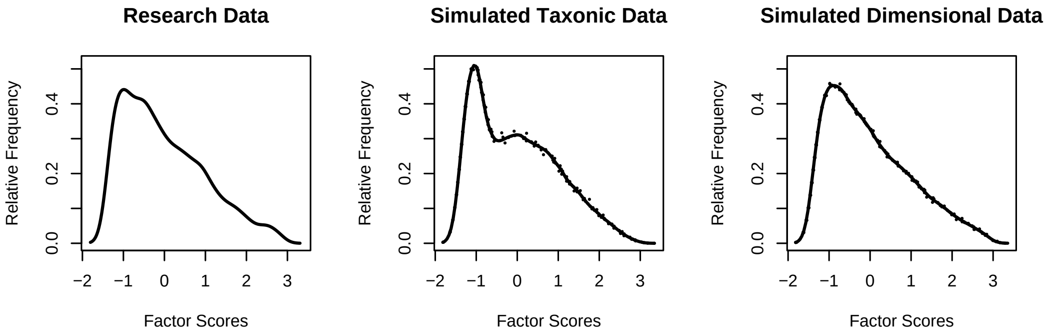

Figure 13: L-Mode graphs for Data Set 2 (left panel), simulated taxonic data (middle panel) and simulated dimensional data (right panel), using CES-D items.

Composite scores. The L-Mode analyses using the composite CES-D scores did not result in any new findings, but confirm previous findings of a dimensional underlying structure. The factor score distributions of the empirical data, the simulated taxonic data, and the simulated dimensional data (depicted in the left, middle and right panel of Figure 14, respectively) are each nearly identical with those stemming from the L-Mode analyses of the CES-D items, although the bimodality of the simulated taxonic data is more pronounced than for items as indicators.

Data Set 3

Indicators

CES-D items. Six CES-D items were found to have an indicator separation of more than 1.25. Two of these six indicators were excluded because of too high correlations. The four remaining indicators that proved to be appropriate for the taxometric analyses were: "I

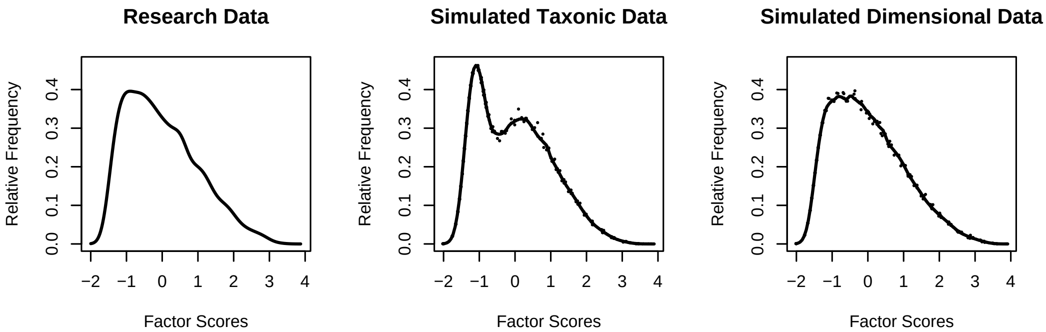

Figure 14: L-Mode graphs for Data Set 2 (left panel), simulated taxonic data (middle panel) and simulated dimensional data (right panel), using composite scores.

felt that everything I did was an effort", "My sleep was restless", "I felt lonely" and "I felt sad".

Composite scores. Three composite indicator variables were created based on the item groups described by Radloff (1977). One of these three composite indicators, the interpersonal, turned out not to have sufficient indicator separation ($d = 0.75$). The remaining two indicators were not sufficient to accomplish the taxometric analyses; hence, for Data Set 3 no composite scores were submitted to taxometric analyses. This failure is not unexpected given the lower mean and variance, and the DIMCAT results obtained for Data Set 3.

MAXCOV

The basic pattern (Figure 15, left panel) is similar to what was seen in Data Set 1 and Data Set 2, that is, a curve with an elevation towards the right end, but the curves of both, the simulated taxonic and simulated dimensional data (Figure 15, middle and right panel, respectively) are quite close to the curve of the empirical data, and not very different from one another. The result from the MAXCOV analyses is rather unclear ($CCFI = .34$).

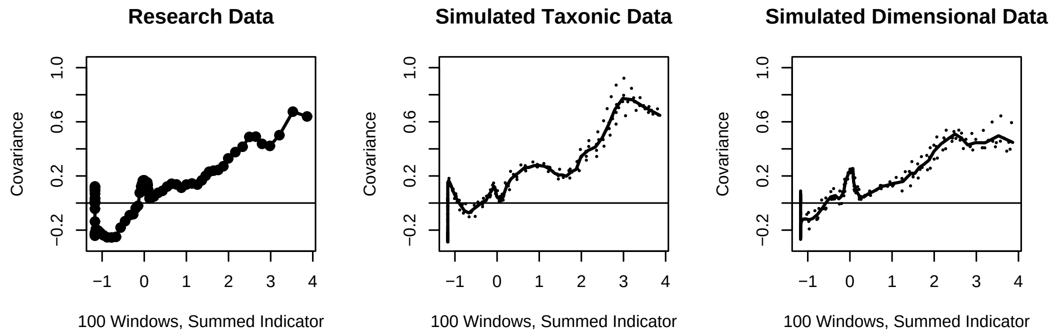

Figure 15: MAXCOV graphs for Data Set 3 (left panel), simulated taxonic data (middle panel) and simulated dimensional data (right panel), using CES-D items.

L-Mode

The software failed to accomplish the L-Mode analysis, because one or more items did not have any variation in at least one of the conjectured classes.

Discussion

The results of the taxometric analyses show a rather homogeneous picture of a dimension-like structure, rather than a category-like structure. This finding could be replicated across two large, independent samples, while the taxometric methodology did not seem appropriate for the third sample, perhaps because of the relative absence of depression in the latter sample. Because of the low degree of depression, it is not unexpected that no clear results are found.

As it was already mentioned in the introduction, the underlying nature of depression has been investigated in several studies. Most of these studies were based on a taxometric approach, and resulted in the conclusion that depression is dimension-like (e.g., Hankin, Fraley, Lahey, & Waldman, 2005; Beach & Amir, 2003; Franklin, Strong, & Greene, 2002; A. M. Ruscio & Ruscio, 2002; J. Ruscio & Ruscio, 2000). These findings are corroborated by the taxometric approach used for the CES-D) in the present manuscript. The CES-D seems to behave in a way that is similar to the other inventories when a taxometric analysis is used.

However, the studies investigating the more specific question whether subtypes of depression exist, came to the conclusion that subtypes do exist. Three previous studies (Ambrosini, Bennett, Cleland, & Haslam, 2002; Haslam & Beck, 1994; Grove et al. 1987) found evidence of taxonicity within the class of depressed. In all three of these studies samples consisting exclusively of depressed subjects were analyzed. Beach and Amir's finding that depression is taxonic when a specific set of indicators is used to investigate the underlying structure (see later) is somewhat different in that these authors were investigating the underlying nature of depression in a general sample consisting of both depressed and nondepressed subjects. The present study is not set up to investigate subtyping, and the data are not really appropriate for this purpose either.

General discussion

An interesting finding when comparing the results of the DIMCAT and the taxometric analyses is that the underlying structure of the CES-D data was found to differ depending on the approach that was used. This divergence of the results can have several possible explanations.

First of all, the definition of category-likeness is different across the two approaches. In taxometrics, the difference between categorical and dimensional may be seen as referring to the differentiation between smooth and abrupt quantitative differences in the DIMCAT approach. In fact, in taxometrics the distinction should rather be seen as one between dimensional (upper right in Figure 1) with smooth differences and also dimensional but with abrupt differences or categories on a dimension (bottom right in Figure 1 which expresses an abrupt difference as well). As Waller and Meehl (1998, p.9) formulated:

"Thus, the convenient taxonomy taxonic-vs.-dimensional should, strictly speaking, read 'taxonic-dimensional vs. dimensional only'" . That is, when the difference between two conjectured classes is abrupt, a categorical or taxonic underlying structure is concluded, while a smooth quantitative difference between the conjectured classes is suggestive of a dimensional underlying structure. In DIMCAT, a more elaborated distinction between category-likeness and dimension-likeness is made, and also a distinction between several forms of category-likeness is made. The distinction between abrupt and smooth quantitative differences is an important one, but it does not tell the whole story.

Second, when performing the taxometric and the DIMCAT analyses not the same indicators have been used. The obvious reason was that not all items met the requirements. Unfortunately, only a fraction of the original indicators managed to fulfill the requirements. It is important knowing that different indicator sets may lead to different conclusions.

The effect of the particular indicator set used in the taxometric analyses was nicely illustrated by Beach and Amir (2003). Using indicators measuring distress from the Beck Depression Inventory (BDI; Beck, Steer, & Brown, 1996), A. M. Ruscio and Ruscio (2002), J. Ruscio and Ruscio (2000) and also Beach and Amir report a dimensional structure for depression (or maybe distress) but Beach and Amir report a taxonic structure when using indicators from the BDI that measure somatic symptoms that can be associated with depression (indicators of the Involuntary Defeat Syndrome). It is also emphasized by De Boeck et al. (2005) that "the results may depend on the indicators considered" (p. 144.). Because of the differences in the indicator sets used, and differences between the aspects of the structures that have been looked at in the two approaches, the divergence of the results does not necessarily imply that the results contradict one another.

An important consideration when drawing conclusions based on the CES-D Scale is: what does the CES-D really measure? Some experts (e.g., Coyne, 1994) question that a construct measured by screening tools such as the BDI or the CES-D corresponds to clinical depression. They argue that these screening tools measure distress but not depression. According to these authors, the differences cannot be reduced by considering these tools as measures of a milder form of depression. They claim instead that there is a conceptual difference between the construct measured by the different screening instruments and clinically diagnosed depression. Santor et al. (in press) reviewed the most widely used measurement tools and they found considerable differences in how these tools operationalize depression. They concluded that the CES-D may contain a number of indicators, such as the interpersonal indicators, that are not likely to be very specific of depression.

There have been several attempts in the past to find an appropriate cutpoint for the CES-D scale, but there is still no widely accepted value. Santor and Coyne (1997) prefer a different path to refine the CES-D, namely shortening the questionnaire. They argued that although it is often assumed that the individual indicators of these tools indicate equally well how severely depressed a certain person is, it is apparently not true, as a smaller group of indicators appears to be considerably more influential than other indicators, some of which can even be considered to be marginal for the assessment of the construct. Furthermore, in a current study, Santor et al. (in press) also pointed out that the CES-D is not representative of how the currently used measurement tools operationalize depression (see above). They also argued that some of the items in the CES-D do not seem very relevant in distinguishing depressed and nondepressed individuals. Our finding of a small subset of items that do distinguish the two classes clearly better than the other items do, does support their conclusions.

The analyses presented in this paper have several limitations. First of all, in the DIMCAT analyses there were only one-class and two-class models considered (except for the third sample). The reason why the number of latent classes was restricted is that the focus of this study was to find out if depression (in comparison to normality) can be described as a dimensional construct rather than as a categorical structure. In the present study, DIMCAT was used in a confirmatory way (two expected classes), rather than in an exploratory way (any number of classes). Since the class assignments provided by the latent class DIMCAT analysis were in good agreement with the class assignments based on a cutpoint of 20 of the CES-D total score, the two-class model seems satisfactory. However, including further latent classes in the analyses could allow to explore and possibly further refine the underlying structure of depression into a subtype structure, although a sample with only depressed persons would make for a better approach.

Second, only a specific measure of depression, the CES-D, was considered in the present studies. Although it can also be considered as a strength, since a measure was investigated which has not yet been studied in this context before (at least the authors are not aware of any such studies), it would also be of interest to investigate other measures, such as the BDI, for example, to see what kind of category-likeness depression would show when the data were analyzed with the DIMCAT methodology. It would be particularly interesting in the light of Beach and Amir's (2003) findings with different indicator sets.

Third, as far as the taxometric analyses are concerned, a limiting factor was the rather limited number of appropriate indicators. Ideally, all relevant depression indicators should be incorporated in a study, so that the limited coverage can no longer be used as a possible explanation of why results are different from what was expected.

In sum, based on the results obtained with the CES-D, one cannot conclude that depression is basically dimension-like. Further studies are required in order for a more definite conclusion to be drawn.

An additional finding, unrelated to the issue of categorical versus dimensional, is that qualitative differences may relate also to the response style. Whether qualitative differences refer to psychopathological groups or to response styles, the conclusion must be that scores cannot always be compared, because the dimension they rely on may be different from one class of persons to another. This is an important measurement issue, one that deserves more attention and can be dealt with using a mixture model approach as used in the present manuscript.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood theory. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–381). Budapest: Akadémiai Kiadó.

Ambrosini, P. J., Bennett, D. S., Cleland, C. M., & Haslam, N. (2002). Taxonicity of adolescent melancholia: a categorical or dimensional construct? *Journal of Psychiatric Research, 36*, 247–256.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders.* (4th ed.). Washington, DC: Author.

Beach, S. R. H., & Amir, N. (2003). Is depression taxonic, dimensional, or both? *Journal of Abnormal Psychology, 112*, 228–236.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory Manual* (2nd ed.). San Antonio, TX: Psychological Corporation.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental test Scores*. Reading MA: Addison-Wesley.

Coyne, J. C. (1994). Self-reported distress: Analog or ersatz depression? *Psychological Bulletin, 116*, 29–45.

De Boeck, P., Wilson, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review, 112*, 129–158.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. London: Erlbaum.

Flett, G. L., Vredenburg, K., & Krames, L. (1997). The continuity of depression in clinical and nonclinical samples. *Psychological Bulletin, 121*, 395–416.

Franklin, C. L., Strong, D. R., & Greene, R. L. (2002). A taxometric analysis of the MMPI-2 depression scales. *Journal of Personality Assessment, 79*, 110–121.

Gotlib, I. H., Lewinsohn, P. M., & Seeley, J. R. (1991). Symptoms versus a diagnosis of depression: differences in psychosocial functioning. *Journal of Consulting and Clinical Psychology, 63*, 90–100.

Grove W. N., Andreasen, N. C., Young, M., Endicott, J., Keller, M. B., Hirschfeld, R. M. A., et al. (1987). Isolation and characterization of a nuclear depressive syndrome. *Psychological Medicine, 17*, 471–484.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Mental Science, 105*, 985–987.

Hankin, B. L., Fraley, R. C, Lahey, B. B., & Waldman, I. D. (2005). Is depression best viewed as a continuum or discrete category? A taxometric investigation of childhood and adolescent depression in a population-based sample. *Journal of Abnormal Psychology, 114*, 96–110.

Haslam, N., & Beck, A. T. (1994). Subtyping major depression: A taxometric analysis. *Journal of Abnormal Psychology, 103*, 696–704.

Herche, J., & Engelland, B. (1996). Reversed-polarity items and scale unidimensionality. *Journal of the Academy of Marketing Science, 24*, 366–374.

Lenzenweger, M. F., & Korfine, L. (1992). Confirming the latent structure and base rate of schizotypy: A taxometric analysis. *Journal of Abnormal Psychology, 101*, 567–571.

Meehl, P. E. (1972). Specific genetic etiology, psychodynamics and therapeutic nihilism.

International Journal of Mental Health, 1, 10–27.

Meehl, P. E. (1973). MAXCOV-HITMAX: A taxometric search method for loose genetic syndromes. In P. E. Meehl (Ed.), *Psychodiagnosis: Selected papers.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of personality, 60*, 117–174.

Meehl, P. E. (1995). Bootstraps taxometrics. Solving the classification problem in psychopathology. *American Psychologist, 50*, 266–275.

Meehl, P. E., & Yonce, L. J. (1994). Taxometric analysis: I. Detecting taxonicity with two quantitative indicators using means above and below a sliding cut (MAMBAC procedure). *Psychological Reports, 74*, 1059–1274.

Meehl, P. E., & Yonce, L. J. (1996). Taxometric analysis: II. Detecting taxonicity using covariance of two quantitative indicators in successive intervals of a third indicator (MAXCOV procedure). *Psychological Reports, 78*, 1091–1227.

Muñoz, R. F., Lenert, L. L., Delucchi, K., Stoddard, J., Perez, J. E., Penilla, C., et al. (2006). Toward evidence-based Internet interventions: A Spanish/English Web site for international smoking cessation trials. *Nicotine & Tobacco Research, 8*, 77–87.

Muraki, E. (1990). Fitting a polytomous item response model for Likert-type data. *Applied Psychological Measurement, 14*, 59–71.

Muthén, L. K., & Muthén, B. O. (2006). M*plus* 4.0 [Computer software]. Los Angeles, CA: Muthén & Muthén.

Oakman, J. M., & Woody, E. Z. (1996). A taxometric analysis of hypnotic susceptibility. *Journal of Personality and Social Psychology, 71*, 980–991.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401.

Rodebaugh, T. L., Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L., & Rapee, R. M. (2004). More information from fewer questions: The factor structure and item properties of the original and brief Fear of Negative Evaluation Scale. *Psychological Assessment, 16*, 169-181.

Ruscio, A. M., & Ruscio, J. (2002). The latent structure of analogue depression: Should the Beck Depression Inventory be used to classify groups? *Psychological Assessment, 14*, 135–145.

Ruscio, J. (2000). Taxometric analysis with dichotomous indicators: The modified MAXCOV procedure and a case removal consistency test. *Psychological Reports, 87*, 929–939.

Ruscio, J. (2006). *Taxometric Programs in R.* [Computer software and documentation]. Retrieved June 27, 2006, from http://www.taxometricmethod.com.

Ruscio, J., & Ruscio, A. M. (2000). Informing the continuity controversy: A taxometric analysis of depression. *Journal of Abnormal Psychology, 109*, 473–487.

Ruscio, J., & Ruscio, A. M. (2004). Clarifying boundary issues in psychopathology: The role of taxometrics in a comprehensive program of structural research. *Journal of Abnormal Psychology, 113*, 24–38.

Ruscio, J., Ruscio, A. M., & Meron, M. (in press). Applying the bootstrap to taxometric analysis: Generating empirical sampling distributions to help interpret results. *Multivariate Behavioral Research.*

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100–114.

Santor D. A., & Coyne J. C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. *Psychological Assessment, 9*, 233–243.

Santor D. A., & Coyne J. C. (2001). Evaluating the symptomatology between depressed and nondepressed individuals. *Journal of Abnormal Psychology, 110*, 216–225.

Santor, D. A., Gregus, M., & Welch, A. (2006). Eight decades of measurement in depression. *Measurement: Interdisciplinary Research and Perspectives, 4*, 135–155.

Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9*, 367–373.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Strube, M. J. (1989). Evidence for the type in type A behavior: A taxometric analysis. *Journal of personality and Social Psychology, 56*, 972–987.

Taylor, J. O., Wallace, R. B., Ostfeld, A. M., & Blazer, D. G. (1998). Established populations for epidemiologic studies of the elderly, 1981-1993: [East Boston, Massachusetts, Iowa and Washington Counties, Iowa, New Haven, Connecticut, and North Central North Carolina] [Computer file]. 3rd ICPSR version. Bethesda, MD: National Institute on Aging [producer], 1997. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

Trull, T. J, Widiger, T. A., & Guthrie, P. (1990). Categorical versus dimensional status of borderline personality disorder. *Journal of Abnormal Psychology, 99*, 40–48.

Tuerlinckx, F., & Wang, W. C. (2004). Models for polytomous data. In P. De Boeck, & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 75–109). New York: Springer.

van der Maas, H. L. J., & Molenaar, P. C. M (1992). Stagewise cognitive development:

An application of the catastrophe theory. *Psychological Review, 99*, 395–417.

van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological review, 98*, 3–53.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua.* London: Sage.

Wilson, M. (1989). Saltus: A psychometric model for discontinuity in cognitive development. *Psychological Bulletin, 105*, 276–289.

Appendix A

The Center for Epidemiologic Studies Depression Scale

Instructions for questions: Below is a list of the ways you might have felt or behaved. Please tell me how often you have felt this way during the past week.

(0) Rarely or none of the time(less than 1 day)

(1) Some or little of the time (1–2 days)

(2) Occasionally or a moderate amount of time (3–4 days)

(3) Most or all of the time (5–7 days)

During the past week:

1. I was bothered by things that usually don't bother me.
2. I did not feel like eating; my appetite was poor.
3. I felt that I could not shake off the blues even with help from my family or friends.
4. I felt that I was just as good as other people.
5. I had trouble keeping my mind on what I was doing.
6. I felt depressed.
7. I felt that everything I did was an effort.
8. I felt hopeful about the future.
9. I thought my life had been a failure.
10. I felt fearful.
11. My sleep was restless.
12. I was happy.
13. I talked less than usual.
14. I felt lonely.
15. People were unfriendly.
16. I enjoyed life.
17. I had crying spells.
18. I felt sad.
19. I felt that people dislike me.
20. I could not get "going".