

Chapter 29

On the Latent Structure of the Beck Depression Inventory (BDI): Using the „Somatic“ Subscale to Evaluate a Clinical Trial

Ulrich Frick, Jürgen Rehm and Uta Thien

Estimate GmbH, Augsburg
Addiction Research Foundation, Toronto
Biometric Center for Therapeutic Studies, München

1. Introduction

Keller & Kempf (this volume, chapter 30) give an outline of the present discussion on psychometric properties of the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, et al., 1961). The overwhelming majority of papers on depression scales used classical test theory and correlation techniques. These methods are rarely complemented by applications of the dichotomous Rasch model (Bech, Gram, Dein, et al., 1975; Maier & Phillip, 1993). Usually, depression scales aimed at measuring severity of illness do not fulfill the criteria of the dichotomous Rasch model (see section 1.1 in chapter 1). The statement holds for expert rating scales as well as for self-administered questionnaires like the BDI (Bouman & Kok, 1987; Frick, Rehm, & Weber, 1991; Keller & Kempf, this volume). Some authors therefore propose to divide the complete set of BDI-items into homogeneous subscales. Bouman & Kok (1987) in this context reported a satisfactory Rasch scale „SOMATIC“ consisting of 8 items. A new look to solve the problem of heterogeneity of the BDI is given by Keller & Kempf (this volume) who divided depressive patients into groups of „scalable“ and „unscalable“ persons by means of a hybrid model (model (49) in chapter 1).

Our approach is different: We had to analyze a large behavioral intervention study (de Jong-Meyer, Hautzinger, Rudolf, et al., in press; Hautzinger, de Jong-Meyer, Treiber, et al., in press) comparing the efficacy of different treatment strategies (antidepressive medication and psychotherapy). Within this study, substantial hypotheses concerning somatic aspects of depression were written down in the study protocol a priori. Based on the results of Bouman & Kok (1987), the operational definition of „somatic aspects“ was given by the respective 8 BDI items (see below). Thus, our focus is not on constructing a meaningful new subscale of the BDI, but instead on the interpretation of the latent structure of these 8 „somatic“ items.

We applied latent class analysis as a tool for this substantial problem because a cross validation of Bouman & Kok (1987) came to the result that neither a dichotomous Rasch, nor a polytomous Mokken scale (see section 1.6 in chapter) could be regarded satisfactory in explaining the latent structure of our sample data (details are not presented in this article but available from the first author). This inhomogeneity could have different reasons: Measuring somatic depressiveness either could be meaningful solely on a categorical scale level. Or, on the other hand, the response formats of several items may not fulfill the conditions of an

interval scale. Estimating LCA-models enabled us to simultaneously check both possible reasons for inhomogeneity.

2. Methods and sample

Two samples were used: Sample I consisted of $n=441$ depressive patients (mean age: 41.2 years $s.d.=10.8$; 32.5% female; 67.5% male) who were measured on the BDI during their diagnostic entrance procedure at 6 different treatment centers. Of this group, $n=274$ patients (= sample II) were randomized into two treatment arms, either receiving antidepressive drugs alone ($n=112$) or a combination of medication and psychotherapy ($n=162$). Patients not randomized participated in a different study protocol carried out at 2 of the 6 treatment centers. Sample II was evaluated again with the BDI at discharge from therapy 12 weeks later.

Sample I was used to estimate the parameters of different latent class models and to choose an adequate psychometric model. For sample II, a manifest classification into latent classes was realized for both measurement points using the parameters of sample I. Notice that these were „admittance parameters“. This is an important difference to the data situation in Keller & Kempf (this volume). In their sample a single person could contribute twice to parameter estimation, when both admittance and discharge BDI data were available.

The BDI subscale SOMATIC comprises 8 items with such diverse symptoms as crying spells, irritability, sleep disturbance, fatigability, loss of appetite, weight loss, somatic preoccupations and loss of libido. Each item is constituted by four alternative categories which are each described by a short sentence. The first alternative always is presenting a state entirely free of the depressive symptoms in question. The three following sentences depict - according to Beck's theory - increasing problem severity concerning the respective symptom (for an example, see 3.1).

Our study focusses on the assumption of ordinality within each item. Thus, we started our analyses without any recoding of the categories. To select the appropriate LCA model for this type of scale and data we ran analyses for each type of threshold specification described for ordinal items (Rost, 1988). We refer to the nomenclature of LACORD (model 1 to model 8) to indicate the models. The mathematical background for possible parameter restrictions is presented in formulas (39) and (40) of chapter 1. Comparisons among models were achieved by using the respective BIC values (Read & Cressie, 1988). All models were estimated by EM-algorithms using the programs LACORD (Rost, 1990) for LCA models and WINMIRA (von Davier, 1994) for the Partial Credit Model. In order to avoid local optima of the likelihood function, 4 different starting values for the estimations procedure were chosen per model. For each LCA model type the number of latent classes was specified to range within 1 to 6.

3. Results

Table 1 shows the results of a first analysis assessing the assumption of within item ordinality under the hypothesis that the BDI subscale SOMATIC could represent a continuous latent trait (partial credit model; see formula (14) in chapter 1). Clearly this assumption is violated

for the items „crying spells“, „irritability“, and „sleep disturbances“. Location parameters in Table 1 correspond to σ_i in formula (17) of chapter 1.

<i>Item</i>	<i>Threshold 1</i>	<i>Threshold 2</i>	<i>Threshold 3</i>	<i>Location</i>
crying spells	1.613	-1.765	1.371	1.291
irritability	1.373	-0.524	0.038	0.888
sleep disturb.	1.320	0.008	0.076	1.404
fatigue	1.883	-0.184	-0.911	0.787
loss of appetite	0.046	-0.257	-1.303	-1.514
weight loss	-1.086	-0.950	-0.785	-2.821
som. preoccupat.	0.507	-0.470	-0.825	-0.789
loss of libido	0.957	0.041	-0.172	0.826

Table 1: Threshold parameters of partial credit model for BDI SOMATIC

Thus we recurred to latent class models to check whether estimation of threshold parameters in different latent classes could solve this problem. A brief inspection of Table 2 indicates that model 4 shows the best fit. In the case of only one latent class, models 5 to 8 with class-specific parameters do not differ from model 1 to 4

Number of latent classes	Number of LC model (LACORD)							
	1	2	3	4	5	6	7	8
1	9834	11567	8881	8788	-	-	-	-
2	8738	8803	8706	8623	8746	8833	8745	8694
3	8702	8706	8690	8602	8720	8714	8747	8711
4	8697	8779	8663	8601	8687	8742	8724	8817
5	8718	8814	8685	8631	8710	8816	8819	8901
6	8735	8787	8705	8660	8754	8846	8848	9012

Table 2: Goodness of fit according to BIC index for psychometric models by number of latent classes

Other goodness of fit criteria (AIC-index or LR-comparisons) showed the same result that model 4 is most adequate. Model 4 defines thresholds which may vary from item to item but which are constant for all latent classes. The fact that each item requires a specific latent metric of its thresholds is not surprising since the verbal anchors were item- and category-specific (see below). Thus it should not be expected that models for constant response formats (see formula (17)) or equidistant thresholds within each single item (see formula (18) in chapter 1) could fit the data. Four classes emerged as the best solution, but the difference to a three class solution was only minimal.

3.1 Are the thresholds ordered?

After selecting the model type the thresholds were inspected to check whether the categories of each item were ordered as required by the theoretical assumption of increasing severity levels. Table 3 lists threshold parameters for the 4 classes solution (interpretation of the 3 classes solution will draw the same conclusions).

The items „crying spells“, „irritability“, „sleep disturbance“ and „weight loss“ deviate from the ordering assumption. For all 4 items the conditional transition from third to fourth category is less „difficult“ than the conditional transition between second and third answer.

„Crying spells“ had the following response alternatives: Category 0: „I don't cry any more than usual“; category 1: „I cry more now than I used to“; category 2 „I cry all the time now. I can't stop it“; category 3: „I used to be able to cry but now I can't cry at all even though I want to“. The patients of the sample examined did not answer according to the a priori defined severity order of these sentences. The loss of crying ability was not rated the most severe symptom of somatic depression.

<i>Item</i>	<i>Threshold 1</i>	<i>Threshold 2</i>	<i>Threshold 3</i>
crying spells	1.06	-2.14	1.09
irritability	1.03	-0.81	-0.22
sleep disturbance	0.77	-0.43	-0.34
fatigability	2.07	-0.39	-1.68
loss of appetite	1.57	-0.09	-1.48
weight loss	0.24	-0.28	0.04
som. preoccupat.	0.68	-0.22	-0.47
loss of libido	0.76	-0.28	-0.48

Table 3: Partial credit thresholds for latent class model 4 (original categories)

An analogous phenomenon could be found with the item „irritability“, where the last alternative also referred to the loss of irritability reactions. Again, this loss did not mark the most severe verbalization of the prespecified metric of irritability. The response alternatives of the item „sleep disturbance“ are as follows: Category 0: „I can sleep as well as usual“; C1: „I can't sleep as well as in former times“; C2: „I wake up 1-2 hours earlier than usual and find it hard to get back to sleep“; C3: „I wake up several hours earlier than usual and can't get back to sleep“. Note that this wording is a re-translation of the German version into English and not the original wording of Beck. Kammer (1983) tried to improve the scale properties of that item by slightly changing the original alternatives. Despite the improved wording, the patients' responses did not confirm the intended order.

So far, the response pattern could be seen as contradictory to established psychiatric theories, but alternative psychiatric theories of depression exist. The reversal pattern in the „weight loss“ item, seemed more surprising where the response alternatives were simply reflecting larger amounts of weight losses (<2kg, 2-5kg, 5-8kg, > 8kg). Here, given a patient loses more than 5kg of weight, it seems „easier“ to lose more than 8kg than 5-8kg. A possible first explanation states that the relationship between this somatic symptom and depressive state may be U-shaped. Secondly, reporting this symptom might be differentially biased according to degree of depression (more tendencies to exaggerate symptoms at medium severity grades of depression). Thirdly, the intention of the weight loss may have caused problems: As in the German version only unintentional weight losses are to be scored, a dichotomous item on the intention to lose weight was inserted into the questionnaire. This item possibly may have altered the cognitive processes of patients' evaluations of their weight losses.

3.2 How to re-establish the order of categories?

This section suggests some methodology to „repair“ the items in order to re-establish order. This is done to allow subsequent substantial analysis of the theoretical hypotheses. Guiding principle of our suggestions were methodological considerations that improvements on the measurement level are inseparably linked to changes on the theoretical level (cf. Gadenne, 1984). Thus, when trying to „repair“ a depression scale, one has to consider theoretical aspects of depression as well as measurement aspects. This procedure will be exemplified discussing all 4 items with ranking deviations of their categories.

The LCA-analogue of the Partial Credit Model (model 4) has per se (Verhelst & Verstralen, 1991) no a priori specification of ordering item categories. It was for theoretical reasons, that the 0-1-2-3 ranking of the item categories was tested. But additional 23 possibilities of ordering categories for each item also exist. Indeed, for three out of four problematic items one can find category rankings that fulfill the criterion of decreasing thresholds.

For „crying spells“ the orders 0-1-3-2 and 2-3-1-0, for „sleep disturbance“ the combinations 0-2-1-3 and 3-1-2-0, and for „weight loss“ the combinatoric possibilities 2-0-1-3, 2-1-0-3, 3-0-1-2, and 3-1-0-2 all display decreasing thresholds parameters with exactly the same corresponding category probability parameters (π_{ixg} in formula (36) of the introductory chapter) and the same likelihood for the model in the given data set. To choose one of those combinatoric possibilities seems to be a strictly empiricistic procedure that would result in item metrics that cannot convince theoretically. Moreover this „repair procedure“ could not be applied for the item „irritability“ in our example. Here, none of the 24 ranking combinations yields decreasing threshold parameters.

Instead we proceeded guided by substantial psychiatric reasoning (for the role of theory in scale validation, see also Gittler, 1986) and discussed different ways of agglomerating categories in order to combine substantially equivalent symptoms into one category. For „crying spells“ and „irritability“ it was argued, that the loss of the respective ability reflects some sort of a depressive status of a patient comparable to an increased frequency of the respective behaviour. This means that the loss of „crying spells“ and/or „irritability“ is judged as less severe than the permanent occurrence of the respective behaviour. Thus, a trichotomization of the categories into the order 0-(1=3)-2 was the first solution strategy.

Basis for our recoding of the sleeping item were numerous psychological findings that people's reports about sleeping and sleeping problems are highly unreliable (Schubert, 1978). Thus, any statements involving only general terms like „several hours“ are suspect to be classified intuitively under a vague category of sleeping problems. On the other hand, a clear statement involving „1-2 hours“ may be more reliably answered, since this response alternative may serve as an anchor (Rehm & Strack, 1994). Therefore, we decided to trichotomize the item „sleeping disturbances“, too, using the succession 0-(1=3)-2 as severity grading.

If the alternatives do not reflect degrees of severity, a categorization into two states (symptom present/absent) seems appropriate. Such an argumentation was used for the item „weight loss“. Weight loss surely can be considered as a symptom of depressive states. The amount of reported weight loss, however, is subject to many other considerations. For example, social desirability plays an important role here, dependent on age and sex of the

respondent. Additionally, it can be doubted, whether depressive patients can correctly classify a weight loss as intended or not. Dichotomizing 0-(1=2=3) was therefore decided as „repair mechanism“ for the item „weight loss“.

The application of the above described repair mechanisms led to monotonously decreasing threshold parameters for each item („crying spells“: 2.15, -2.15; „irritability“: 1.39, -1.39; „sleeping disorders“: 0.31, -0.31; „weight loss“ now has only location parameters). Again, four classes constituted the best solution for the number of latent classes.

3.3 On the interpretation of the „repaired“ solution

A comparison of the two solutions in Figure 1 demonstrates the consequences of repairing items. Whereas the lower part displays the location parameters of the 4 latent classes before repairing, the „corrected“ solution with a strictly ranking of the categories is shown in the upper part of Figure 1.

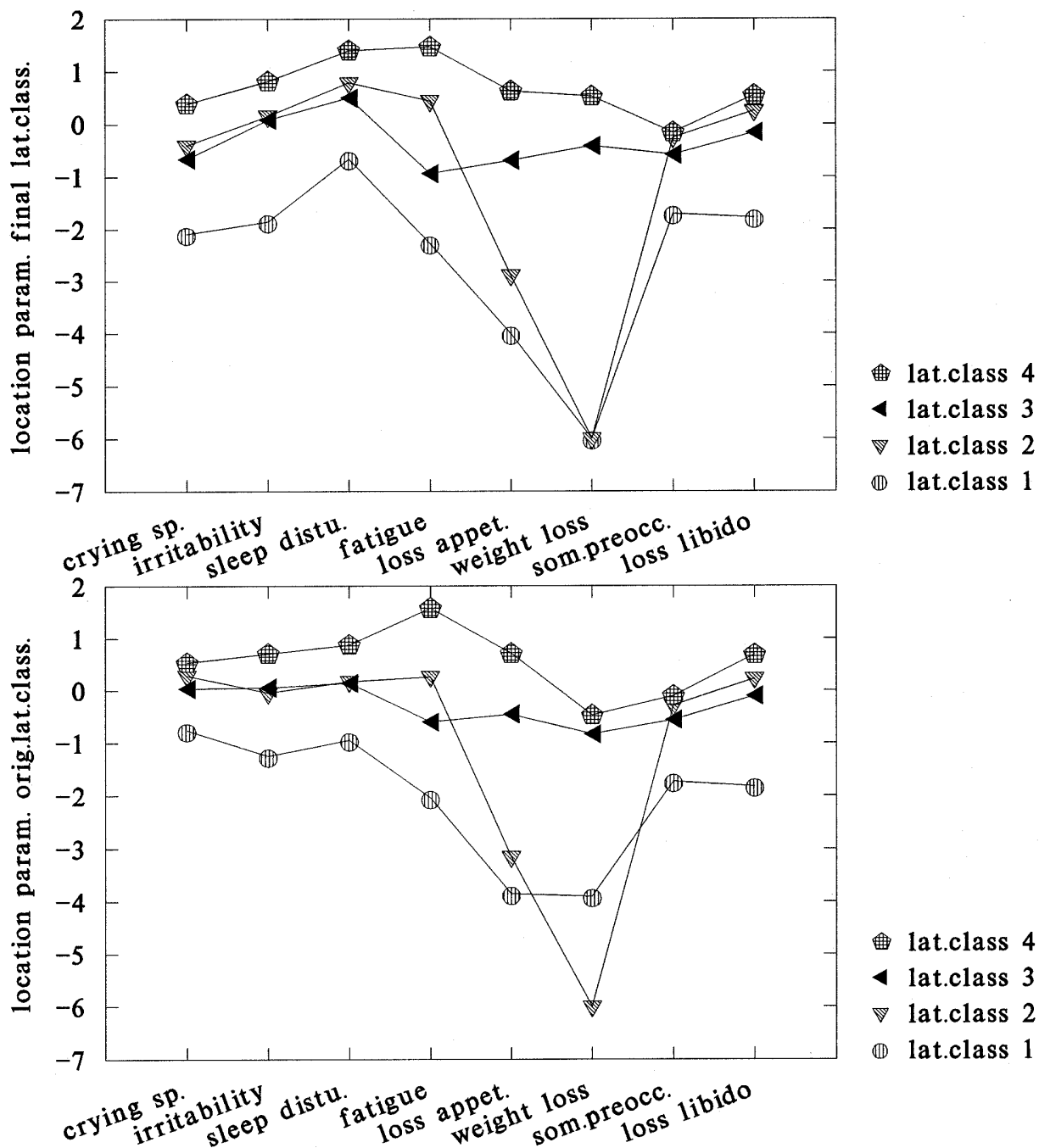


Figure 1: Location parameters of latent class model 4 with original and „repaired“ categories

The most important difference concerns the reduction of intersections between classes. These intersections were mostly caused by violations of the a priori order of item alternatives. In the „corrected“ solution almost an ordinal scale concerning the person parameters (see section 2.1 of the introductory chapter) was achieved. Class 4 displayed the highest values, class 1 the lowest values on all items. For 6 out of 8 items the ranking of the latent classes was perfect, since for those items class 2 without exception showed higher values than class 3. However, this pattern reversed for the two items concerning eating behaviour „loss of appetite“ and „weight loss“. Thus, a strict ordinal ranking of severity of

somatic-depressive states was not possible. Two different patterns of moderate severity emerged and thus still prohibit the utilization of latent trait models.

3.4 What have we gained for the analysis of the clinical trial?

Traditional scoring of the SOMATIC subscale and analysing the results of the clinical trial in sample II by means of an ANOVA (one between subjects factor = kind of treatment received; one repeated measurement factor = admission vs. discharge) would have resulted in a non-significant difference between treatment arms, and a significant diminution of „severity“ of somatic depression at discharge from therapy. But when summing up the items of a heterogenous scale the absence of a treatment effect could have simply been a consequence of the misspecification of a latent trait model for the BDI subscale SOMATIC, with no relation to „real“ treatment effects.

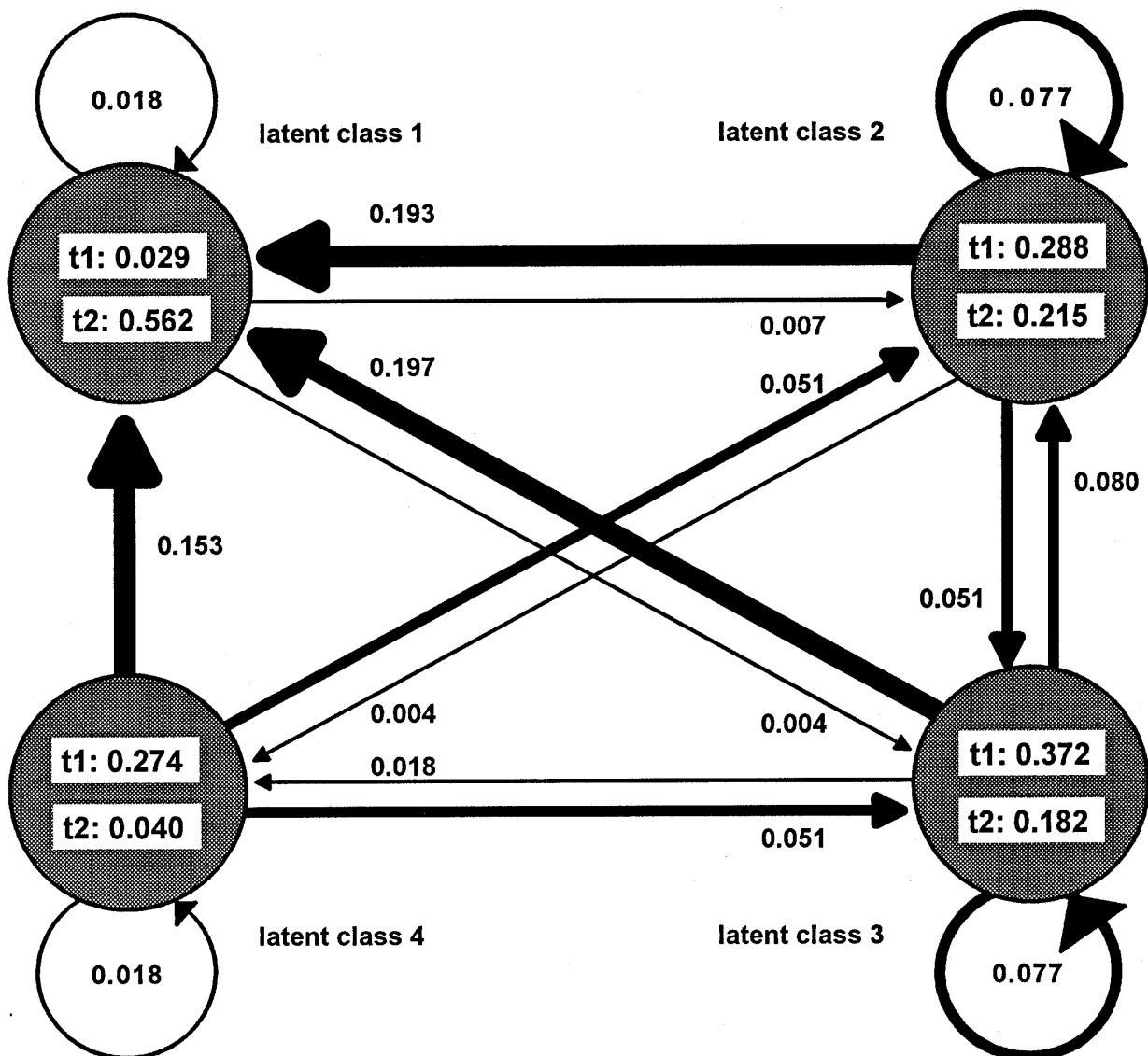


Figure 2: Class membership of 274 depressive patients before (t1) and after (t2) treatment

By deciding on the modal value of the class probability $p(g|x_v)$ (see formula (37) in the introductory chapter) the patients of sample II were classified to one of the four latent classes for both time points. The cross tabulation of the pre- versus post-state of the $n=274$ patients could not disclose any differential treatment effect for the course of depression, either. But in Figure 2 one can see some aspects of both treatment groups that could not be seen by means of an ANOVA. Numbers in Figure 2 denote sample fractions.

The more or less symptom free latent class 1 grows in size from 2.9% at admission (t_1) to 56.2% at discharge (t_2). But recurrence from somatic depression is not a constant reduction of symptoms for all patients. Instead, more than about half of the patients in the symptom loaden categories move to the symptom free state 1. Neither this effect nor the considerable amount of "stayers" in latent classes 2 and 3 could have been detected by linear models: 27% of class 2 members (=7.7% of total sample II) and 21% of class 3 members (=7.7% of total sample II) do not change. Also the symptom shift between the two "medium" states (22% of class 3 move to class 2; 18% of class 2 to class 3) would have remained undetected.

4. Discussion and conclusion

LCA revealed which items failed to reach an ordinal level of measurement. After having applied „repair mechanisms“ that were guided by psychiatric and psychological reasoning, still some problems were found in regard to the unidimensionality of the SOMATIC scale. The two items with respect to eating behaviour were mainly responsible for this heterogeneity.

Some cautious remarks concerning the repair mechanisms are necessary. Whereas problems of ordinality in a priori fixed response categories must be empirically tested for each sample, „repairing“ cannot be the solution in every problematic situation. Especially in cases when underlying theories are not well developed, a slightly misspecified scale could nevertheless be brought into use reporting only the anomaly for further discussion.

Psychometrics can only contribute to validity if the underlying theoretical concepts are consistent and reflect reality. Psychiatric knowledge sets the upper boundaries of the validity of the respective tests. Quite a lot of work seems necessary in this respect for the field of measurement of depression.

The decision for changing the original scale by collapsing categories and/or reversing categories should always be cross-validated in separate samples. Concerning our example of the SOMATIC scale of the BDI such a cross-validation is planned together with Ferdinand Keller using his data set (chapter 30).

To conclude, for the purposes of an applied statistician who has to make the best of the data already collected, LCA seems to be the least unsatisfying way to develop a scoring rule for substantive interpretation of non homogeneous psychometric tests.

References

- Bech, P., Gram, L.F., Dein, E., Jakobsen, O., Vitger, J., & Bolwig, T.G. (1975). Quantitative rating of depressive states. *Acta psychiat. scand.*, 51, 161-170.
- Beck, A.T., Ward, C.H., Mendelson, M., Moch, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Arch. of General Psychiatry*, 4, 53-63.
- Bouman, T.K., & Kok, A.R. (1987). Homogeneity of Beck's Depression Inventory (BDI): Applying Rasch analysis in conceptual exploration. *Acta psychiat. scand.*, 76, 568-573.
- de Jong-Meyer, R., Hautzinger, M., Rudolf, G., Strauss, W., & Frick, U. (in press). Combined cognitive-behavioral and pharmacotherapy in endogenously depressed inpatients and outpatients. *Archives of General Psychiatry*.
- Frick, U., Rehm, J., & Weber, A. (1991). *How to validate endpoint measures in behavioural intervention trials?* revised version of a paper presented at the Joint Meeting of the Society for Clinical Trials and the International Society for Clinical Biostatistics, July 1991, Brussels.
- Gadenne, V. (1984). *Theorie und Erfahrung in der psychologischen Forschung*. Tübingen: Mohr.
- Gittler, G. (1986). Inhaltliche Aspekte bei der Itemselektion nach dem Modell von Rasch. *Z.f. exper. angew. Psychol.*, 33, 386-412.
- Hautzinger, M., de Jong-Meyer, R., Treiber, R., Rudolf, G., & Thien, U. (in press). Wirksamkeit kognitiver Verhaltenstherapie, Pharmakotherapie und deren Kombination bei nicht-endogener, unipolarer Depression. *Z. f. klinische Psychologie*, 26.
- Kammer, D. (1983). Eine Untersuchung der psychometrischen Eigenschaften des deutschen Beck-Depressionsinventars (BDI), *Diagnostica*, 29, 48-60.
- Maier, W., & Philipp, M. (1993). *Reliabilität und Validität der Subtypisierung und Schweregradmessung depressiver Syndrome*. Berlin/Heidelberg/New York: Springer.
- Reed, T.R.C. & Cressy, N.A.C. (1988). *Goodness of fit statistics for discrete multivariate data*. New York: Springer.
- Rehm, J., & Strack, F. (1994). Kontrolltechniken. In: T. Herrmann & W. Tack (Eds.) *Methodologische Grundlagen der Psychologie (Enzyklopädie der Psychologie, Themenbereich B, Serie 1, Band 1)* 1994. p. 508-555.
- Rost, J. (1988). Rating scale analysis with latent class models. *Psychometrika*, 53: 3, 327-348.
- Rost, J. (1990). *LACORD - Latent class analysis for ordinal variables. Manual*. Kiel: Institut für die Pädagogik der Naturwissenschaften.
- Schubert, F.-Ch. (1978). *Einschlafverlauf und Einschlafstörungen. Aktuelle und habituelle Formen*. Frankfurt/Bern: Lang .
- Verhelst, N.D., & Verstralen, H. (1991). The partial credit model with non-sequential solution strategies. Cito: *Measurement and Research Department Reports*, 91-5.
- von Davier, M. (1994). WINMIRA: A program system for analyses with the Rasch Model, with the Latent Class Analysis and with the Mixed Rasch Model. Manual. Kiel: Institut für die Pädagogik der Naturwissenschaften.