MULTILEVEL COVARIANCE STRUCTURES FOR FAMILY DATA

J.J. Hox and M.E. Jacobs

When a research problem concerns the relationships between attributes of individuals and attributes of the families of which they are part, the research topic involves a multilevel problem. In multilevel research, the data structure in the population is hierarchical, and the sample data are viewed as a multistage sample from this hierarchical population. Thus, in family research, the population consists of families and individuals within these families, and the sampling procedure proceeds in two stages: first we take a sample of families, and next we take a sample of individuals within each family.

The theoretical concept behind multilevel research is that individuals interact with the social contexts to which they belong. Thus, individual persons are influenced by the families to which they belong, and the properties of these families are in turn influenced by the individuals who make up that family. Individuals and families are conceptualized as a hierarchical system, with individuals and families defining separate levels of this hierarchy. Naturally, such hierarchical systems can be observed at different hierarchical levels, and as a result may have variables defined at each level. This leads to research into the interaction between variables that describe the individuals and variables that describe the families.

In this book, Snijders describes several applications of the multilevel regression model to family data. Multilevel regression models are essentially a multilevel version of the familiar multiple regression model. As Cohen and Cohen (1983) and others have shown, the multiple regression model is very versatile. Using dummy coding for categorical variables, it can be used to analyze a wide variety of research problems. Snijders uses dummy coding to model both relationships and variance components within one comprehensive model.

In our contribution, we describe the application of covariance structure analysis to multilevel family data. The general statistical model for multilevel covariance structure analysis is quite complicated. This chapter describes a simplified statistical model proposed by Muthén (cf. Muthén, 1994), which can be estimated with conventional software such as Lisrel. We confine ourselves to multilevel exploratory factor analysis and a simple confirmative path model, but the extension of the approach to other models is straightforward (cf. Muthén, 1989; McDonald, 1994; Hox, 1994).

THE BASIC DECOMPOSITION MODEL FOR A HIERARCHICAL POPULATION

Suppose we have data from N individuals, divided into G family-groups. The individual data are denoted by $\mathbf{Y}_{ig}$ (subscripts i for individuals, i=1..N; g for groups, g=1..G). Cronbach and Webb (1975) have proposed to decompose the individual $\mathbf{Y}_{ig}$ into a between groups component $\mathbf{Y}_B = \bar{\mathbf{Y}}_{.g}$ and a within groups component $\mathbf{Y}_W = \mathbf{Y}_{ig} - \bar{\mathbf{Y}}_{.g}$. Thus, for each individual we replace the observed *Total* score $\mathbf{Y}_T = \mathbf{Y}_{ig}$ by its components: the group mean $\mathbf{Y}_B$ and the individual deviation from the group mean $\mathbf{Y}_W$. These two components have the attractive property that they are orthogonal (uncorrelated) and additive.

Multilevel structural models assume that we have a population of individuals that are divided into groups. If we decompose the population data we have, for the population covariance matrices:

In: J.J. Hox, B.F. van der Meulen, J.M.A.M. Janssens, J.J.F. ter Laak & L.W.C. Tavecchio (Eds.). (1995). *Advances in Family Research.* Amsterdam: Thesis Publishers.

$$\Sigma_T = \Sigma_B + \Sigma_W. \qquad (1)$$

Multilevel covariance structure models assume that the population covariance matrices Σ_B and Σ_W are described by separate models for the between families and within families structure.

Unfortunately, the sample between families covariance matrix S_B is not an unbiased estimate of Σ_B, and the sample within families covariance matrix S_W is not an unbiased estimate for Σ_W (Cf. Muthén, 1989). Thus, we cannot simply analyze S_B and S_W to find valid estimates of multilevel population models. But, as Muthén (1989) has shown, the sample *pooled within families* covariance matrix S_{PW} is an unbiased estimate of the population within groups covariance matrix Σ_W. As a result, we can estimate the population within group structure directly by constructing and testing a model for S_{PW}. Unfortunately, there is no simple estimator for the population between families covariance matrix Σ_B. Instead, the regular between families covariance matrix in the sample S_B is an estimator of the sum of two population matrices:

$$S_B = \hat{\Sigma}_W + c\Sigma_B, \qquad (2)$$

where c is a scaling factor based on the average family size.

EXPLORATORY ANALYSIS: A DIRECT SOLUTION

One solution to the problem that S_B estimates a combination of Σ_W and Σ_B is to estimate Σ_B by subtracting S_{PW} and S_B, as follows:

$$S^*_B = (S_B - S_{PW})/c = (\Sigma_W + c\Sigma_B - \Sigma_W)/c = \Sigma_B. \qquad (3)$$

Thus, S^*_B is a direct estimate of Σ_B. This direct approach has two important drawbacks, which make it virtually impossible to use it in confirmative factor analysis models. The first drawback is that S^*_B is rather unstable. The second drawback is that S^*_B, which is obtained by simply subtracting two covariance matrices, is not necessarily a proper covariance matrix. If S^*_B is converted into a correlation matrix, this will often show in correlations exceeding 1.00 or impossible correlation patterns. However, if we are willing to simply ignore these problems, an exploratory analysis of S^*_B or the corresponding correlation matrix may still be useful for exploration and scale construction.

As an example, we take the shortened version of the Leuven Family Questionnaire (LFQ, cf. Jacobs & Schoorl, 1993). This questionnaire consists of 42 items that form two non-overlapping scales: '*tension*' and '*connectedness*'. Our data set contains 420 individuals in 189 families. In each family both parents and one child (average age 12;8) filled in the questionnaire. In some families there was only one parent and one child. The children were all in the first year of secondary school.

Assume that we want to analyze the factor structure of the LFQ at both the family (between families) and individual (within families) level, with an aim to improve the scales by removing items that do not perform well at one or more level.

An exploratory factor analysis of the raw scores would analyze a mixture of the family and the individual level. Instead, we calculate S_{PW}, S_B, and S^*_B. For the individual level, we convert the pooled within families covariance matrix S_{PW} into the corresponding individual level correlation matrix R_{PW} and analyze this using ordinary component analysis. The results are in table 1 below.

Table 1. Varimax rotation within families components (individual level)

	tension	**connectedness**
V1	*.34*	-.13
V2	**-.34**	.17
V3	.06	**.44**
V6	-.12	**.31**
V10	*.32*	-.07
V12	-.08	**.34**
V13	*.44*	.00
V15	-.20	**.42**
V16	.21	*-.27*
V18	-.23	**.29**
V19	*.47*	-.14
V21	.02	**.46**
V22	.22	*-.28*
V23	**-.36**	.15
V25	*.56*	-.05
V27	-.26	.26
V28	*.41*	-.21
V33	-.03	**.24**
V34	*.39*	.20
V36	-.04	**.37**
V37	*.44*	-.38
V39	.10	**.59**
V41	-.09	**.29**
V42	.07	**.62**
V43	*.49*	-.20
V44	-.22	**.35**
V48	-.07	**.30**
V51	*.32*	-.24
V53	**-.32**	.19
V55	*.44*	-.33
V56	.01	**.22**
V57	-.20	**.46**
V60	*.49*	.18
V62	*.45*	-.03
V63	*.57*	.00
V64	*.46*	-.23
V65	*.63*	-.07
V66	*.39*	-.02
V67	*.46*	-.02
V68	*.59*	-.18
V69	.15	**-.39**
V73	.18	**-.32**

In Table 1 we have marked the items that belong to 'tension' by using italics, and items that belong to 'connectedness' with bold. Likewise, for these items we have used italics respectively bold typeface to mark the highest loadings. As is clear, most of the italics are in the first column, and most of the bolds in the second, which indicates that most of the items have their highest loading on the factor to which they belong.

At the family level, we convert the direct estimate of the family level covariances $\mathbf{S}^*_B$ into the corresponding correlation matrix $\mathbf{R}^*_B$ and again use ordinary component analysis to obtain an indication of the family level factor structure. However, we notice that the family level correlation matrix $\mathbf{R}^*_B$ is indeed not a proper correlation matrix, which implies that the corresponding covariance matrix $\mathbf{S}^*_B$ is not a proper covariance matrix either. Table 2 below shows part of the family level correlation matrix $\mathbf{R}^*_B$.

Table 2. Some entries in the family level correlation matrix

Var.	1	2	3	6	10	12	13	15	16
2	-.18								
3	-.52	.28							
6	-.71	.51	.44						
10	.50	-.19	-.72	-.80					
12	-.47	-.15	.18	.22	.02				
13	.26	-.33	-.63	-.88	.45	-.42			
15	-.37	.61	.91	.88	-.39	.46	-.69		
16	.57	-.15	-.55	-.18	.76	-.23	.71	-.56	
18	-1.39	1.05	1.41	.29	-1.25	.37	-1.09	.74	-1.73

There are 'smoothing' techniques to convert such matrices as in Table 1 into proper covariance or correlation matrices. Instead, we simply analyze the improper 'correlation matrix' $\mathbf{R}^*_B$ with component analysis.[1] The resulting solution is in Table 3 below. Note that, since the entries in the input matrix are not correlations, we cannot interpret the loadings in Table 3 as correlations either. Still, high values in Table 3 still indicate a relatively strong relationship, and we can interpret Table 3 analogous to an ordinary component analysis. Table 3 uses the same italic/bold notation to indicate the first component 'tension' and the second component 'connected.'

The pattern in Table 3 (family level) conforms reasonably to the theoretical structure, but it is somewhat less satisfactory as the pattern in Table 1 (individual level). If we compare Table 1 and Table 3, we observe that item 73 is problematic in both Table 1 (individual level) and Table 3 (family level) because it has a negative relationship with the component it is intended to measure. Some items, such as items 2 and 23 do not perform well at either level. Other items, such as item 27, perform well at one level but not at another. If our goal is item analysis and scale construction, we could remove such items and try to replace them.

[1]The advantage of component analysis over most factor analysis methods is that it is much more lenient as to the kind of matrix that is used. For the sake of consistency, we also used component analysis on the proper individual level correlation matrix $\mathbf{R}_{PW}$. A factor analysis on this matrix produces highly similar results.

Table 3. Varimax rotation between families components (family level)

	tension	**connectedness**
V1	*.63*	-.31
V2	**-.26**	.06
V3	**-.60**	.59
V6	**.43**	.43
V10	*.73*	-.16
V12	-.26	**.43**
V13	*.71*	-.28
V15	**.60**	.44
V16	*.93*	-.14
V18	**-1.66**	.03
V19	*.89*	-.35
V21	-.34	**1.09**
V22	*.85*	-.57
V23	.01	.07
V25	*.91*	.06
V27	.27	**.73**
V28	*.47*	*-.52*
V33	.04	**.50**
V34	*.88*	.01
V36	.02	**2.51**
V37	*.79*	-.33
V39	.00	**.85**
V41	-.03	.03
V42	.22	**.90**
V43	*.87*	-.36
V44	-.49	**.57**
V48	-.38	**.73**
V51	*.50*	-.30
V53	-.15	**1.12**
V55	*.99*	.03
V56	-.58	**.60**
V57	.15	**.87**
V60	*.54*	-.19
V62	*.84*	-.33
V63	*.67*	-.23
V64	*.92*	-.15
V65	*.70*	.11
V66	*.85*	.49
V67	*.65*	.21
V68	*.91*	-.34
V69	.74	**-.85**
V73	.20	**-1.11**

As we noted before, the direct estimate of the between families covariance or correlation matrix is not a particularly good estimate. Its use in the previous example is justified because the stated goal of the analysis was to identify items that predominantly load on the intended scale, for the purpose of scale construction. If our goal is more theoretical, we prefer to use confirmative models on the covariance matrices at both levels.

As noted above, the pooled within families covariance matrix $\mathbf{S}_{PW}$ is an unbiased estimate of the population within families covariance matrix Σ_W, and we can estimate the population within group structure by constructing and testing a model for $\mathbf{S}_{PW}$. But, if we want to model the between families structure, we cannot simply construct and test a model for $\mathbf{S}_B$, because $\mathbf{S}_B$ estimates a combination of Σ_W and Σ_B. Instead, we have to specify for $\mathbf{S}_B$ two models: one for the within families structure and one for the between families structure.

Muthén (1989) proposed to use the multigroup option of conventional covariance structure software to analyze these models simultaneously. The procedure is that we specify two 'groups,' with covariance matrices $\mathbf{S}_{PW}$ and $\mathbf{S}_B$ (based on N-G and G degrees of freedom).

The model for Σ_W must be specified for both $\mathbf{S}_{PW}$ and $\mathbf{S}_B$, with equality restrictions between both 'groups' (i.e., the within and between covariance matrix) to guarantee that we are indeed estimating the same model in both covariance matrices, and the model for Σ_B is specified for $\mathbf{S}_B$, with the scaling factor c built into the model.[1] The procedure is described in nontechnical terms in Muthén (1994,) using Muthén's Liscomp program; McDonald (1994) describes essentially the same procedure in different terms using a specialized computer program. Hox (1994) describes the modeling procedure in detail, and explains how to implement the models in Lisrel (Jöreskog & Sörbom, 1989).

Assume that we want to use our example data to test a simple path model, that states that the mean education of the parents and the family position, parent versus child, have an effect on both 'tension' and 'connectedness.' The scale scores are obtained by summing the item scores that belong to a scale. Family position is a dummy variable scored 0 for parents and 1 for children. It is an individual level variable. If we have families of widely different sizes, we may expect some family variation for this variable as well. Since in our example data almost all families consist of two parents and one child, we may expect that in our case there will be almost no family level variation for this variable. The mean education of the parents is a family level variable; by definition all members of a specific family will have assigned to them an identical value for this variable. Thus, for this variable we have by definition no within family variation. Variables that have no variance at one of the two levels are treated as variables with systematically missing values; in structural modeling this is a well-established procedure (cf. Jöreskog & Sörbom, 1989). As a result, these variables are only modeled on that level where they actually exist.

To compute the various covariance matrices we employ the program SPLIT2 (Hox, 1995). This program reads raw data, with family variables attached to the individual members, and computes the matrices $\mathbf{S}_{PW}$, $\mathbf{S}_B$, and $\mathbf{S}^*_B$, the appropriate degrees of freedom for the within and between group level, and the scaling constant c. Exhibit 1 below shows (part of) the output of the program for our example data.

[1]The model is only strictly valid in the *balanced* case, where all families have the same size. In the unbalanced case, we use an average group size (Muthén, 1989), and proceed as if the family sizes were equal. This *pseudo-balanced* solution is generally quite accurate (Muthén, 1990, 1994; McDonald, 1994; Hox, 1993).

Exhibit 1. Sample output of SPLIT2.

Variables are: *position, tension, connectedness, education.*

number of groups: 189 number of cases: 420
df within: 231 df between: 189

between model scaling constant and its square root: 13.06 3.61

Pooled Within Groups Correlations and Standard Deviations

	position	tension	connectedness	education
tension	.37			
connectedness	-.19	.13		
education	.00	.00	.00	
standard deviations	.59	10.80	6.25	1.00

Scaled Between Groups Correlations and Standard Deviations

	position	tension	connectedness	education
tension	.08			
connectedness	-.06	.20		
education	-.04	-.05	-.02	
standard deviations	.28	17.36	7.81	2.13

Rough Estimate of Between Groups Correlations and Standard Deviations

	position	tension	connectedness	education
tension	.00			
connectedness	.00	.28		
education	.00	-.06	-.04	
standard deviations	1.00	3.76	1.30	0.59

Estimates of Intraclass correlations for all Variables

	position	tension	connectedness	education
estimated intraclass corr.	-.06	.11	.04	1.00

The intraclass correlation in the last line of Exhibit 1. is an estimate of the proportion of family level variance. The intraclass correlation of the family variable 'education' is 1.00, which indicates that all of the variation of this variable is on the family level. The intraclass correlation of the dummy variable 'position' is negative, which indicates that the variation is less than we would actually expect on the basis of sampling variation. The reason is obviously the fact that we have chosen to interview only one child per household, which makes this variable almost a constant at the family level (the existing variation stems from the fact that in a few cases only one parent was available to be interviewed). Both cases are handled in the same way; the variable receives a value of one for the standard deviation and for the

diagonal of the correlation matrix, and a zero everywhere else. In our structural model, we must take appropriate measures to insure that this variable is omitted from the analysis.On the individual level, we have only one explanatory variable: 'position'. This translates into an extremely simple path model. Figure 1 below shows this model, with estimates obtained by analyzing S_{PW} only:

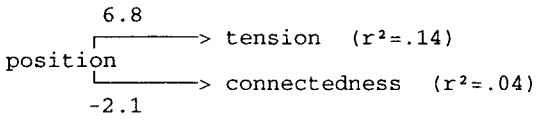

Figure 1. Example path model on individual level.

This model is a simple regression model with two dependent variables. Since it is a saturated model, the number of degrees of freedom is zero, and it cannot be tested. The regression coefficients tell us that children experience more tension and less connectedness than their parents, but the amount of variance that is explained is low. If we specify this model as a multigroup analysis for both S_{PW} and S_B, requiring that all corresponding parameter estimates must be equal for both the within and the between model, we obtain similar path estimates. This time we have six degrees of freedom to test the model, and the chi-square is 170.

Since the within families part of the model is saturated, *all* this lack of fit must derive from the omission of a between families model for S_B. The simplest between families model provides for family level variances, but no covariances. It amounts to the hypothesis that we have family level variance (corresponding to the positive intraclass correlations in Exhibit 1) but no significant covariance structure. The corresponding structural model has three degrees of freedom, and a chi-square of 19 (p=.00). Apparently, there is some covariance structure at the family level. This turns out to be a very simple model, with a single covariance between 'tension' and 'connectedness.' This model has one degree of freedom and a chi-square of 3.9 (p=.05). We could have reached the same conclusion by inspecting the rough estimate of the between groups correlations in Exhibit 1 and noting that there is only one meaningful correlation, which is between 'tension' and 'connectedness.' Families that experience more connectedness tend to experience less tension.

DISCUSSION

The analyses presented above are intended as examples, and we refrain from an extended substantive interpretation. Both examples highlight the methodological problems associated with a specific type of application. The first example is a scale construction problem. Typically, in such applications we have a large number of items, and only a few scales. Even in single level applications confirmative factor analysis is usually not the preferred method, because it will almost always reject the model without giving much information on how to improve the instrument. Instead, exploratory factor analysis or cluster analysis are routinely used. These methods can be used on the within families correlations, and also on the rough estimate of the between families correlations. In our example, we use component analysis

because this is has fewer assumptions than factor analysis, while still allowing us to easily compare the components at both levels (comparing two cluster solutions is more difficult). We should be aware that for the family level we are probably analyzing an improper matrix (even if this is not obvious from the values themselves,) but in scale construction we do not need statistically sanctioned ironclad proofs as much as sensible indications on how to proceed in our instrumental construction. If we need statistically correct verification of our results, we must use confirmatory analysis procedures, as in our second example.

Our second example is an extremely simple path model. Even so, it illustrates both the basic procedure and a typical methodological problem. The basic procedure is to establish a satisfactory model for the within groups (individual) part of the model. The reason is that we obtain our estimates for the between groups (co)variances after taking account of the within groups part of the model. If the within groups model is deficient, the between groups model will also be flawed. By using for the individual level not the complete pooled within groups covariance matrix, as in the direct rough estimate, but a parsimonious model for that matrix, we obtain more stable estimates for the family level of our model. After finding a model for the individual level, we specify models for the family level. In both specification searches, we may use the chi-square test and modification indices to suggest alterations of our models.

The path model also illustrates the problem of having variables that have no or almost no variance at one of the levels. This may be a variable that is defined at one level only, such as the family level variable 'mean parent education.' It may also be an empirical result, as with our variable 'position'. In both cases this variable receives a special treatment in the model. It is treated as a variable with values that are systematically missing in one of the groups, for which standard procedures exist in structural modeling.

We note that family position explains some variance in tension, but very little variance in connectedness. There are no correlated error terms, meaning that the covariance between tension and connectedness in Exhibit 1 is completely explained by position. On the family level this variable has no variance, and we have a simpler model that just specifies a covariance between tension and connectedness. Although the model is extremely simple, the example shows how the procedure works, and that it may lead to different models for within and between family processes.

REFERENCES

Cronbach, L.J. & Webb, N. (1979). Between class and within class effects in a reported aptitude x treatment interaction: a reanalysis of a study by G.L. Anderson. *Journal of Educational Psychology*, 67, 717-724.

Hox, J.J. (1994). *Applied multilevel modeling*. Amsterdam: TT-Publikaties.

Hox, J.J. (1995). Split2. Computerprogram, Department of Education, University of Amsterdam.

Jacobs, M.E. & Schoorl, P.M. (1993). De berekening van gezinsscores. De Leuvense Gezinsvragenlijst in Nederlands onderzoek [Calculating family scores. The Leuven Family Questionnaire in Dutch research]. In: J.J.F. ter Laak, L.W.C. Tavecchio & B.F. van der Meulen (eds): *Opvoeding in perspectief: theorievorming, onderzoek en hulpverlening* [Education in perspective: theory building, research, and assistance]. Groningen: Stichting Kinderstudies.

Jöreskog, K.G. & Sörbom, D. (1989). *LISREL 7. A guide to the program and applications*. Chicago: SPSS Inc.

McDonald, R.P. (1994). The bilevel reticular action model for path analysis with latent variables. *Sociological Methods & Research*, 22, 399-413.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, 54, 557-585.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22, 376-398.