Running head: MULTILEVEL MODELS

Multilevel Models for Multimethod Measurements

Joop Hox, Cora Maas
Utrecht University

Corresponding author:
J.J. Hox
Department of Methodology and Statistics
Utrecht University
P.O.B. 80140
NL-3508 TC Utrecht
the Netherlands
j.hox@fss.uu.nl
tel + 31 30 2539236 fax + 31 30 2535797

Abstract

This chapter discusses the use of multilevel modeling to assess the reliability and/or validity of measurement instruments. Multilevel modeling is most useful when the measurement instruments (e.g. questionnaire items) are defined using an arrangement that produces questions with distinct characteristics. Multilevel modeling can then be used to analyze the effects of person and question characteristics on the responses, and to assign scores on specific characteristics to individuals. Multilevel analysis is also useful if the goal is to measure the characteristics of a context that is shared by individuals (e.g. group characteristics). Multilevel modeling makes it possible to combine and compare information from different sources and at different levels. Finally, the use of multilevel models is considered for generalizability studies. Although multilevel models are useful, and simplify the analysis because they directly produce the necessary variance components, their use in generalizability theory is impeded by limitations of current multilevel software.

Multilevel Models for Multimethod Measurements[1]

Theoretical constructs used in social and behavioral science are often complex, and have an indirect relationship to the corresponding empirical observations. The distance between a theoretical construct and observable phenomena creates the problem that researchers must state explicitly how they plan to measure what they are theorizing about. To assure construct validity, the methodological advice is often given to measure each construct in more than one way (e.g., Hoyle, Harris & Judd, 2002, p35 & 78; Kerlinger, 1973, p462). Fiske (1971) advocates not only using multiple operationalizations of each construct, but also purposefully manipulating operationalizations to span different theoretical perspectives and modes of assessment. This raises questions about the convergence and discriminability of different constructs and measures, which underlies the development of the multitrait multimethod method (Campbell & Fiske, 1959).

This chapter focuses on using multilevel modeling to combine information from different sources, and to assess the reliability and validity of the resulting estimates. It starts with a brief introduction to multilevel analysis. Following this introduction, three measurement approaches are discussed where multilevel modeling is a valuable and effective analysis tool: facet design, assessing contextual characteristics, and generalizability theory. These approaches were chosen because they all, each in their own fashion, aim to incorporate information from several distinct sources in one measurement instrument. Each approach is explained using an example including an analysis of a small data set. The role of multilevel analysis in all three approaches is to assess the contribution of different sources of variance, both due to different traits and due to the specific measurement modes used, in designs where standard analysis methods encounter difficulties.

A Brief Introduction to Multilevel Analysis

Multilevel models are needed for the analysis of data that have a hierarchical or clustered structure. Such data arise routinely in various fields, for instance in educational research, where pupils are nested within schools, or in family studies with children nested within families.

[1] Unless stated otherwise, the data used in the examples are artificial data. Data sets used in the examples are available from the authors.

Clustered data may also arise as a result of the research design. For instance, repeated measures can be viewed as a series of measurements nested within individual subjects.

The models used in this chapter are multilevel regression models. The multilevel regression model assumes hierarchical data, with one response variable measured at the lowest level and explanatory variables at all existing levels. Conceptually the model is often viewed as a hierarchical system of regression equations. For example, assume we have data in J groups or contexts, and a different number of individuals N_j in each group. On the individual (lowest) level we have the dependent variable Y_{ij} and the explanatory variable X_{ij}, and on the group level we have the explanatory variable Z_j. Thus, we have a separate regression equation in each group:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}. \quad (1)$$

The β_j are modeled by explanatory variables at the group level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}, \quad (2)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}. \quad (3)$$

Substitution of (2) and (3) in (1) gives the single equation:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij} + u_{1j} X_{ij} + u_{0j} + e_{ij} . \quad (4)$$

In general there will be more than one explanatory variable at the lowest level and also more than one explanatory variable at the highest level. The assumptions of the multilevel regression model are that the residual errors at the lowest level e_{ij} have a normal distribution with a mean of zero and a variance σ^2_j. Usually, it is assumed that the groups have a common variance σ^2. The second level residual errors u_{0j} and u_{1j} are assumed to be independent from the lowest level errors e_{ij}, and to have a multivariate normal distribution with means of zero and variances σ^2_u. Other assumptions, identical to the common assumptions of multiple regression analysis, are fixed predictors and linear relationships. The estimators generally used in multilevel analysis are Maximum Likelihood (ML) estimators, with standard errors estimated from the inverse of the information matrix. These standard errors can be used to establish a confidence interval or test for significance. This is, in general, not correct for the variance components, because there the null-hypothesis is on the boundary of the parameter space (variances cannot be negative).

Therefore, variances are generally tested using a likelihood-ratio test or a chi-square test described by Raudenbush and Bryk (2002). Two different Likelihood functions are commonly used in multilevel regression analysis: Full Maximum Likelihood (FML) and Restricted Maximum Likelihood (RML) (cf. Raudenbush & Bryk, 2002; Goldstein, 1995). RML estimation is preferred when the interest is in estimating the variance components. For details on the statistical model and estimation techniques we refer to the literature (e.g., Goldstein, 1995; Hox, 2002; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

Facet Design

A useful device for the systematic definition of a theoretical construct is Guttman's facet design (Guttman, 1954). Facet design defines a universe of observations by classifying them with a scheme of *facets* with elements subsumed within facets. Facets are different ways of classifying observations, the elements are distinct classes within each facet. The universe of observations is classified using three kinds of criteria: (1) the population facets that classify the population, (2) the content facets that classify the variables, and (3) the common range of response categories for the variables. The facet design approach can be expressed graphically as follows.

$$[\,X\,] \times [\,A \times B \times \ldots N\,] \rightarrow R$$

In this representation, [X] is the population of objects (respondents, research participants), [A], [B]... [N] are content facets, and R is the common response range. Roskam (1990) emphasizes the importance of the response range, because it defines the domain of observations. Thus, if the range is defined as “correct/wrong by an objective criterion,” we are investigating intelligence behavior, and if the range is defined as “ordered as very positive/very negative toward that object,” we are investigating attitude behavior (Roskam, 1990, p.189).

For our present goal, we concentrate on the facet structure of the variables. The various content facets can be viewed as a cross-classification, analogous to an analysis of variance design that specifies the similarities and dissimilarities among questionnaire items. Each *facet* represents a particular conceptual classification scheme that consists of a set of elements that define possible observations. The content facets must be appropriate for the construct that they define. In selecting the most appropriate facets and elements, the objective is to describe all

important aspects of the content domain explicitly and unequivocally. For example, for many constructs it may be useful to distinguish a behavior facet that defines the relevant behaviors, and a situation facet that defines the situations in which the behaviors occur. An example of a facet design is Gough's (1985, p. 247) design for reasons for attending weight reduction classes. Gough defines the person facet [X] as “married women attending slimming groups”. There are two content facets: *source* and *motive*. The facet design can be summarized as: To what extend does the person [X] feel that *source* [S] led her to believe that she would achieve *motive* [M] if she lost weight, as rated using *Response* [R]. The source facet [S] has four elements: (1) own experience, (2) husband, (3) doctor, and (4) media. The motivation facet [M] has seven facets: (1) feel healthier, (2) feel fitter, (3) be more physically attractive, (4) have fewer clothing problems, (5) suffer less social stigma, (6) be less anxious in social situations, and (7) feel less depressed. The response range [R] is defined as a seven-point scale ranging from (1) “not really at all” to (7) “very much indeed.” In facet design, the facet structure is often verbalized by a mapping sentence, which describes the observations in one or more ordinary sentences. Figure 1 presents a mapping sentence for the reasons for attending weight reduction classes.

In this facet design the first facet (source) refers to the source of the belief, and the second facet (reason) refers to a specific consequence of losing weight. A facet design as given above can be used to generate questionnaire items. The [X] facet points to a specific target population of individuals. The source facet has four elements, the reason facet has seven, which defines $4\times7=28$ questions. For example, combining the first elements of the source and reason facets leads to the survey question “Did your own experience lead you to believe that your health would improve if you lost weight?” (Gough, 1985, p. 257).

--- Figure 1 about here ---

A facet design for a set of questions is a definition, which should not be judged in terms of right or wrong, but whether it leads to productive research. Facet design contains no general guidelines to determine the need for specific facets, it clearly assumes that we already have a good notion of the empirical domain under investigation.

Analysis of Facet Data

Facet design is part of a more general approach called facet theory, which uses the facet structure to generate hypotheses about similarities between items. Facet theory relies almost exclusively on producing low-dimensional geometric representations of the data, which are then interpreted in terms of the properties of the defining facets (c. Borg & Shye, 1995). Other approaches include confirmatory factor analysis (cf. Mellenbergh, Kelderman, Stijlen & Zondag, 1979). A problem with both approaches is that the analysis focuses on similarities between items, and attempts to relate characteristics of the facet design to these similarities. However, as Borg (1994) explains, the relationship between characteristics of the facet design and the geometric ordering of the ensuing items is weak at best. This can be illustrated with the idea of a confirmatory factor analysis of the reasons for slimming design. We have a source facet with four levels and a motivation facet with seven levels. Do we predict four + seven= eleven factors, or do we predict four × seven= 28 factors? Or should we assume that the facet design merely ensures the content validity of a one-dimensional instrument?

Figure 2 presents part of a (simulated) data set for 50 respondents responding to the 28 items generated by Gough's (1985) facet design.

--- Figure 2 about here ---

A classical reliability analysis produces a reliability coefficient (alpha) of 0.93. This is very high, but not unusual with facet data, because facet designs tend to produce items that are very similar in content and wording. A factor analysis (principal factors, eigenvalue>1, promax rotation) produces seven factors: four factors that are mostly based on the source facet, and three subsequent factors that are not readily interpretable.

Multilevel modeling of facet data takes a different viewpoint. The responses on the common response range are viewed as observations of what occurs when a specific person encounters a specific item. The goal of the multilevel analysis is to determine which item and person characteristics (as defined by the facet design) predict the outcome of this encounter. If all respondents respond to all items, a facet design produces cross-classified data, which can be handled by standard analysis methods such as Anova. However, a large facet design generates too many items to include them all in a single instrument. Older research (cf. Borg & Shye, 1995) typically solves this problem by taking a subsample from all possible items. However, modern computer-assisted data collection methods make it easy to present a different sample of questions to each respondent. In this case, the facet design produces multilevel data, with items

nested within respondents, with the response as the outcome variable, and person and item characteristics as predictors. The item characteristics are predictors at the lowest (item) level, and the person characteristics are predictors at the person level.

A multilevel analysis of the reasons for slimming design requires that the categorical source and motives facets be expressed as dummy variables. A multilevel analysis involving only these item characteristics shows that only the effects of the source dummies vary significantly across respondents; the effects of the motive dummies have no random variation at the respondent level. For the final model, it is convenient to include all four source dummies in the regression equation, so we can model the (co)variances of all regression coefficients of the source facet. Therefore, the intercept is no longer part of the equation. The seven motive elements are still represented by the usual set of 7-1=6 dummy variables. The final model can be expressed in equation 5:

$$Y_{ij} = ¥_1 S_{1ij} + ¥_2 S_{2ij} + ¥_3 S_{3ij} + ¥_4 S_{4ij} + ¥_5 M_{1ij} + ¥_6 M_{2ij} + ¥_7 M_{3ij} + ¥_8 M_{4ij} + ¥_9 M_{5ij} + ¥_{10} M_{6ij} + u_{1j} + u_{2j} + u_{3j} + u_{4j} + e_{ij}\ , \quad (5)$$

where S_1 to S_4 are dummies that indicate the four elements of the source facet, and M_1 to M_6 are dummies that represent the six elements of the motivation facet. The variances $\pm^2_{u1}$ to $\pm^2_{u4}$ of the person-level residual error terms u_1, u_2, u_3 and u_4 are significant (using a likelihood-ratio test, see Hox, 2002), which indicates that there is significant slope variation across persons for the sources 1) own experience, 2) husband 3) doctor, and 4) media. The variances of the regression slopes for the motivation dummies are not in the model because they were not significant, which means that there is no individual variation in the impact of the motivation facet.

This multilevel analysis produces several interesting estimates. Table 1 presents the regression coefficients and the variances for this model. The regression slopes for the item characteristics express overall differences between the item means related to the item content. The (significant) variances of the regression slopes for the predictors belonging to the source facet express differences between respondents in their sensitivity to item content coming from specific sources. The software HLM (Raudenbush, Bryk, Cheong & Congdon, 2000) calculates reliability estimates for the random slopes (for other software these must be hand-calculated using formulas presented in Raudenbush & Bryk 2002). The reliability estimates for the slope variation of s1, s2, s3 and s4 are 0.84, 0.83, 0.84 and 0.87, respectively. This means that

variations in sensitivity to reasons originating from different sources can be measured with sufficient precision.

--- Table 1 about here ---

The slopes of doctor and media and of self and husband correlate highly (0.93 and 0.89 respectively), but the other slopes are relatively independent (correlations lower than 0.61). If we need to use these measurements in a different context, we can estimate residuals or posterior means for the slopes. These are estimates of the slopes for the individual respondents. This is especially convenient if we want to use the slope estimates as predictors of person characteristics in another analysis. If we want to predict the slopes on the basis of person characteristics, a better strategy is to include these as person-level predictors in the analysis. In addition to the item characteristics, we have the person-level variable age. Since we have four slopes that vary across persons, we can use the respondents' age to predict these four slopes. Age is entered into the analysis centered on its grand mean; the model is presented in equation 6:

$$\begin{aligned} Y_{ij} = {} & ¥_1 S_{1ij} + ¥_2 S_{2ij} + ¥_3 S_{3ij} + ¥_4 S_{4ij} + ¥_5 M_{1ij} + ¥_6 M_{2ij} + ¥_7 M_{3ij} + ¥_8 M_{4ij} + ¥_9 M_{5ij} + ¥_{10} M_{6ij} \\ & + ¥_{11} S_{1ij} Age_j + ¥_{21} S_{2ij} Age_j + ¥_{31} S_{3ij} Age_j + ¥_{41} S_{4ij} Age_j \\ & + u_{1j} + u_{2j} + u_{3j} + u_{4j} + e_{ij} . \end{aligned} \qquad (6)$$

The estimates are presented in Table 1 next to the estimates of the previous model. The effects of age are not the same on all slopes. Sensitivity to reasons coming from the respondent herself and her husband increases with age, and sensitivity to reasons coming from the doctor or the media decreases with age.

In the example given above, the facets are characteristics of the questions, which is how facet design is commonly employed. However, the facet approach is very general, and can be extended for instance by expanding the person facet, denoted by [X] in figure 1, to include explicit definitions of important respondent characteristics. In addition, it is also possible to extend the response range by defining facets and elements for the responses. This is useful if there are multivariate outcomes, or if the response range is assessed by multiple persons such as independent raters. Analyzing facet data with multiple responses requires a multilevel model for multivariate outcomes, which is set up using a separate level for the multiple 130

outcome variables (Hox, 2002). The multilevel model used is similar to the model used for contextual measurement, a subject taken up in the next section.

Measuring Contextual Characteristics

The term ‘multilevel’ refers to a hierarchical data structure, which often consists of individuals nested within some social context, e.g. individuals within families or organizational contexts such as pupils in school classes. Individual outcome variables are viewed as influenced by both individual characteristics and characteristics of the higher-level units. In this perspective, measuring characteristics of these contexts is an important activity. Some of these characteristics may be measured directly at their natural level; for example, at the school level we can directly assess school size and school denomination, and at the pupil level intelligence and school success. In addition, we may move variables from one level to another, for instance by aggregation7. Aggregation means that the variables at a lower level are moved to a higher level, for instance by computing the school mean of the pupils' intelligence scores.

If the interest is in characteristics of the context, an approach often taken is to let subjects rate various characteristics of the context. In this case we are not necessarily interested in the subjects, they are just used as informants to judge the context. Such situations may arise in educational research, where pupils may rate school characteristics such as school climate, or in health research, where patients may be used to express their satisfaction with their general practitioner, or community research, where samples from different neighborhoods evaluate various aspects of the neighborhood in which they live. In these cases, we may use individual characteristics to control for possible measurement bias, but the main interest is in measuring some aspect of the higher-level unit (cf. Paterson, 1998; Raudenbush & Sampson, 1999; Sampson, Raudenbush & Earls, 1997).

A simple example concerns data from an educational research study by Krüger (1994), analyzed in more detail by Hox (2002). As part of the study, small samples of pupils from each school rated their school manager on six seven-point items that indicate a people-oriented approach toward leadership. There are ratings from 854 pupils within 96 schools, 48 with a male and 48 with a female school manager. Cronbach’s alpha for these six items is 0.80, which is commonly considered sufficient (Nunnally & Bernstein, 1994). However, this reliability estimate is difficult to interpret, because it is based on a mixture of school level and pupil level variance. Since all judgments within the same school are ratings of the same school manager,

within school variance does not give us information about the school manager. From the measurement point of view, we want to concentrate only on the between schools variance.

Reliability and Multilevel Measurement

Raudenbush, Rowan and Kang (1991) discuss the issues involved in multilevel measurement. One convenient way to model data such as these is to use a three-level model, with separate levels for the items, the pupils, and the schools. Using a model with no explanatory variables except the intercept, the variance between items is decomposed into variance components at the item-, pupil and school level. This model can be written as:

$$Y_{hij} = ¥_{000} + u_{0hij} + u_{0ij} + u_{0j}, \qquad (7)$$

where $¥_{000}$ is the intercept term and the subscript *h* refers to items, *i* to pupils and *j* to schools. The variances are in Table 2, using an obvious notation for the subscripts of the variance components.

--- Table 2 about here ---

In Table 2, $±^2_{item}$ can be interpreted as an estimate of the variation due to item inconsistency, $±^2_{pupil}$ as an estimate of the variation of the mean item score between different pupils within the same school, and $±^2_{school}$ as an estimate of the variation of the mean item score between different schools. The item level exists only to produce an estimate of the variance due to item inconsistency. The error variance in the mean of *p* items equals $±^2_e = ±^2_{item} / p$, which for the example data equals 0.141.

The pupil level internal consistency is given by $~_{pupil} = ±^2_{pupil}/(±^2_{pupil} + ±^2_{item}/p)$. For our example data $~_{pupil}$ is 0.71. This reflects consistency in the item scores from different pupils in the same schools. The internal consistency coefficient of 0.71 indicates that this variability is not random error, but that it is systematic. It could be systematic error, for instance response bias such as a halo effect in the judgments made by the pupils, or it could be based on different experiences of pupils with the same manager. This could be explored further by adding pupil characteristics to the model. The school level internal consistency is given by (Raudenbush et al., 1991, p. 312):

$$\sim_{school} = \pm^2_{school} \Big/ \left(\pm^2_{school} + \left(\pm^2_{pupil} + \pm^2_{item} / p \right) \Big/ n_j \right). \quad (8)$$

In equation (8), p is again the number of items in the scale, and n_j is the number of pupils in school j. Since the number of pupils varies across schools, the school level variability also varies. An indication of the average reliability can be calculated by using equation (8) with the mean number of pupils for n_j. In our example we have on average 8.9 pupils in each school, and the school level internal consistency is $\sim_{school} = 0.77$. The school level internal consistency coefficient indicates that the school managers' leadership style is measured with reasonable consistency.

The school level internal consistency depends on four factors: the number of items in the scale, the mean correlation between the items on the school level, the number of pupils sampled in the schools, and the intraclass correlation at the school level. The school level reliability as a function of these quantities is given by:

$$\sim_{school} = \frac{k n_j \circ_I \overline{r}}{k n_j \circ_I \overline{r} + \left[(k-1)\overline{r} + 1 \right](1 - \circ_I)}, \quad (9)$$

where $\overline{r}$ is the mean item intercorrelation at the school level, which can be estimated using the variances in the intercept-only model by $\overline{r} = \pm^2_{pupil} \Big/ \left(\pm^2_{pupil} + \pm^2_{item} \right)$. (The relationship between equation 8 and equation 9, which is based on the Spearman-Browne formula, is explained by Raudenbush *et al*., 1991).

Equation (9) shows that the internal consistency reliability can be improved by including more items in the scale, but also by taking a larger sample of pupils in each school. Raudenbush *et al*. (1991) demonstrate that increasing the number of pupils in the schools increases the school level reliability faster than increasing the number of items in the scale. Even with a low inter-item correlation and a low intraclass correlation, increasing the number of pupils to infinity will in the end produce a reliability equal to one, whereas increasing the number of items to infinity will in general not.

If we want to predict the evaluation scores of the school manager using school-level variables, for instance the experience or gender of the school manager, or type of school, we can simply include these variables as explanatory variables in the multilevel model. We can also

estimate the school managers' evaluation scores using the school level residuals. We can add pupil-level explanatory variables to the model, which would lead to evaluation scores that are conditional on the pupil-level variables. This can be used to correct the evaluation scores for inequalities in the composition of the pupil population across schools.

Multivariate Multilevel Measurement

Raudenbush *et al*. (1991) extend the measurement model by combining items from several different scales in one analysis. The constant in the multilevel model is then replaced by a set of dummy variables that indicate to which scale each item belongs. This is similar to a confirmative factor analysis, but with the restriction that the loadings of all items that belong to the same scale are equal, and that there is one common error variance. These are strong restrictions, and multilevel structural equation modeling (Hox, 2002) is both more flexible and less restrictive. However, multilevel structural equation modeling does not model raw scores, it is based on simultaneous analysis of a person-level and a group-level covariance matrix. Therefore, it does not produce estimated scores on the latent variables. Consider the following example of combining individual-level and group-level information. Assume we ask pupils in 100 classes to rate their teacher using the semantic differential method (Hoyle *et al*., 2002). In the semantic differential method, three factors 'evaluation,' 'activity' and 'potency' are assumed to underlie a set of bipolar rating scales. In our example, each teacher is rated by the students on a set of 3 items each for evaluation, activity and potency. In addition, the teachers rate themselves on the same set of nine items, using the same bipolar rating scale that runs from –4 to +4. This creates a multitrait multimethod structure, where the three semantic differential factors are the traits, and the teacher and students are the measurement methods. In addition, we have multiple raters for the student ratings, with in general a different number of student raters for each teacher.

The resulting data can be viewed as a multilevel structure, with nine items varying on both the pupil level and the teacher level, and nine items varying only on the teacher level. One convenient way to model data such as these is to use a multivariate multilevel model, with separate levels for the items, the pupils, and the schools. At the lowest level we have 18 items, which refer to three semantic differential scales for the pupils and three for the teachers. Thus, we create 6 dummy variables d_{1ij} to d_{6ij} variables to indicate the 3 scales × 2 types of raters, exclude the regression coefficient for the intercept from the model, but keep the lowest level

variance term to estimate the residual variance among the 18 items. Hence, at the lowest level we have

$$Y_{hij} = ®_{1ij} d_{1ij} + ®_{2ij} d_{2ij} + \ldots + ®_{6ij} d_{6ij} + e_{hij} . \quad (10)$$

At the pupil level we have

$$®_{pij} = \ _{pj} + u_{pij} , \quad (11)$$

and at the class/teacher level (the third level in the multivariate model), we have

$$_{pj} = ¥_{p} + u_{pj} \ . \quad (12)$$

By substitution, we obtain the single-equation version

$$\begin{aligned} Y_{hij} &= ¥_{1} \ d_{1ij} + ¥_{2} d_{2ij} + \ldots + ¥_{6} d_{pij} \\ &+ u_{1ij} d_{1ij} + u_{2ij} d_{2ij} + \ldots + u_{6ij} d_{pij} \\ &+ u_{1j} d_{1ij} + u_{2j} d_{2ij} + \ldots + u_{6j} d_{pij} + e_{hij} . \end{aligned} \quad (13)$$

The model described by equation (13) provides us with estimates of the six scale means, and of their variances and covariances at the pupil and class level. Since in this application we are mostly interested in the variances and covariances, Restricted Maximum Likelihood (RML) estimation is preferred to Full Maximum Likelihood (FML) estimation. Table 3 below presents the RML estimates of the covariances and the corresponding correlations at the pupil level and at the school level.

--- Table 3 about here ---

Table 3 shows that most of the variance is between classes. The variances and covariances at the class level are important for inspecting the convergent and discriminant validity of the measures. In fact, at the class level we have a multitrait multimethod matrix that consists of three traits and two methods. The pairwise correlations between the three methods measured through

pupils and teachers, respectively, is the validity diagonal. The correlations of 0.64-0.66 indicate a substantial convergent validity for these measures.

Generalizability Theory

The central issue in Generalizability (G) theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972) is the generalization from a sample of measurements to a universe of possible measurements. This universe is defined in terms of measurement conditions from which the observed measurements are a random sample. The question to be answered is how well measures taken in one condition can be generalized to other conditions. In other words, how well the observed scores correspond to the averages scores acquired under all possible conditions. In the classical true score model observed scores consist of two components, a systematic component, called the true score and a random error component. The reliability is then defined as the correlation between the observed and the true scores and all possible observed scores on this particular test. In generalizability theory, the variance of the measurements is divided into several different variance components. The generalizability coefficient based on this partition is defined analogous to the reliability coefficient: the true variance divided by the expected observed-score variance (Shavelson & Webb, 1991). The variance partition in generalizability theory requires a clear description of all relevant measurement conditions. These conditions are called *facets* (the terminology is similar to facet-theory, but there facets refer mostly to question formats, and in generalizability theory they refer mostly to measurement conditions). In the simplest case, there is only one facet. For instance, when students take a test of twenty multiple choice items at the end of a course, the examiner is not interested in the answers on these particular twenty items, but in the knowledge of the whole course content. From this perspective, the twenty items are a sample of all possible items. The *items* are the facet of the measurement. When all students answer the same twenty items, the design is *crossed*. This means that all students have the same conditions (items). When all students answer different items, the design is *nested*. Then, all students have different conditions.

Assume that the twenty items of the foregoing example are not multiple choice items, but behavioral observations. When these observations are coded by trained judges, the design becomes a two facet design. The observations are the first facet and the judges the second

facet. In this case we must generalize over both observations and judges to obtain an estimate of the true score we are interested in.

One Facet Crossed Design

To illustrate a one facet crossed design we created a small example with eight persons and four multiple choice items. The data are in Table 4.

--- table 4 about here ---

The person scores on the items are decomposed into four parts:

$$X_{pi} = « + («_p - «) + («_i - «) + (X_{pi} - «_p - «_i + «), \quad (14)$$

where X_{pi} is the score of person p on item i; μ is the grand mean, the expectation over persons and items; (μ_p-μ) is the person effect, the expectation of the persons score over items; (μ_i-μ) is the item effect, the expectation of the item difficulty over persons; and (X_{pi}-μ_p-μ_i+μ) is the residual, which includes both the interaction effect between items and persons and all error components. These two effects cannot be distinguished because we have only one observation for each person-item combination.

Each effect in equation (14), except the grand mean, has a distribution, with a mean of zero and a non-zero variance (Shavelson & Webb, 1991). Standard Anova estimates the mean squares as MS_{person} = 0.268, MS_{item} = 0.375, and $MS_{residual}$ = 0.232. The variance components are calculated from these mean squares (Shavelson & Webb, 1991):

$$£^2_{residual} = MS_{pi,error} \quad (15)$$

$$£^2_i = (MS_i - £^2_{pi,error}) / n_p \quad (16)$$

$$£^2_p = (MS_p - £^2_{pi,error}) / n_i . \quad (17)$$

The variance components are $£^2_p = 0.009$, $£^2_i = 0.018$ and $£^2_{residual} = 0.232$, which account for 3 percent, 7 percent and 90 percent of the variance.

The variance components themselves are unstandardized; therefore the interpretation uses the percentages. Three percent of the variance is associated with persons, seven percent with items and the remainder with the interaction and error.

The generalizability coefficient (G-coefficient) for the above example depends on the decisions one wants to make (Shavelson & Webb, 1991). For relative decisions, only the

variance component of the interaction between persons and items contributes to the measurement error. When this variance component is large, this means that the relative position of persons is different for the different items. Because all persons answer the same items, the item variance doesn't influence the relative position. In contrast, for absolute decisions, both the item variance and the variance of the interaction are important. In our example, the formulas for the estimated error variances are:

$$\hat{\pm}^2_{Abs} = \frac{\pm^2_i}{n_i'} + \frac{\pm^2_{residual}}{n_i'} \quad (18)$$

$$\hat{\pm}^2_{Rel} = \frac{\pm^2_{residual}}{n_i'}, \quad (19)$$

where n_i is the number of items. Calculating the error variances gives 0.06 for both the relative variance and the absolute variance (the estimates are equal due to rounding). The formula for the G coefficient for relative decisions is:

$$G - coefficient = \frac{\pm^2_p}{(\pm^2_p + \pm^2_{Rel})}. \quad (20)$$

Calculating this *G*-coefficient gives 0.09. The interpretation of this coefficient is analogous to the interpretation of the reliability coefficient in classical test theory. Because of the simple dataset used, we refrain from further interpretation.

The reliability-like index of dependability for absolute decisions is given by:

$$dependability = \frac{\pm^2_p}{(\pm^2_p + \pm^2_{Abs})}. \quad (21)$$

Calculating this index gives also 0.09. The interpretation of this coefficient is not exactly the same as the interpretation of the G-coefficient, but broadly speaking it has the same function. In both cases, the generalizability coefficient indicates to what extend the measurements converge across specific method facets, including possible interaction effects. The decision across which method effects we need to generalize, which leads to different generalizability coefficients, remains of course with the researcher.

One Facet Nested Design

For the description of a one facet nested design we created example data for eight persons on sixteen multiple choice items. The data are shown in table 5.

--- table 5 about here ---

The person scores on the items are decomposed into three parts:

$$X_{pi} = « + («_p - «) + (X_{pi} - «_p), \qquad (22)$$

where X_{pi} is the score of person p on item i; μ is the grand mean; $(\mu_p-\mu)$ is the person effect; and $(X_{pi}-\mu_p)$ is the residual. There is no separate term for the item effect. Because all students have answered different items, the item effect cannot be estimated and the item effect becomes part of the residual.

Each effect in equation (22), except the grand mean, has a distribution, with a mean of zero and a variance. The Anova estimates (persons are random, items not) are $MS_{person} = 0.348$ and $MS_{residual} = 0.188$, and the calculated variance components (see equations 15 and 17) are $£_p^2 = 0.080$ and $£_{residual}^2 = 0.188$. Thus, thirty percent of the variance is associated with persons, the remainder with items, the interaction and error.

Two Facet Designs

An example of a two-facet design is a design with assignments made by students which are each graded by a different judge. When all students make all assignments, and each judge evaluates one of the assignments, we have a two-facet crossed design. Assuming that we want to assign a single grade to the students, this constitutes a design with one trait and multiple methods, i.e. the cross-classification of assignments and judges. In this design the person scores are decomposed into seven parts: the grand mean, the effects from students, assignments and judges, and all the two-way interaction-terms. The three-way interaction and the error components cannot be distinguished. When all students make different assignments evaluated by different judges, we have a two-facet nested design. In practice, many designs are partly nested. For instance, all students make the same assignments evaluated by different judges, or each student makes his/her assignments evaluated by all judges, but the subset of assignments made is different for each combination of students and judges. For an elaborated description we refer to Shavelson and Webb, 1991.

Multilevel Models for Generalizability Analysis

Generalizability theory can be viewed as a special case of multilevel analysis. In the one facet nested design, the nesting structure is clear: the items are nested in the persons. In the one facet crossed design, the nesting structure is arbitrary: items can be seen as nested in persons, or persons nested in items. Both specifications lead to the same results. Because of the analogy with the nested design, we will use the specification structure of items in persons. In a two facet design the structure is more complicated, because of the large number of interaction effects. Although it can be specified as a cross-classified multilevel model (Goldstein, 1995), current software cannot analyze data of a realistic size and complexity.

The specification of a one facet nested design is straightforward. An intercept-only model is specified with two levels. At the lowest level we obtain a direct estimate of the residual variance, and at the second level a direct estimate of the person-level variance. These estimates are exactly the same as the variance components estimated before.

The one facet crossed design, where all persons respond to all items is specified as a three level intercept-only model. Although the analysis is set up using three separate levels, it should be clear that conceptually we have two levels, items nested in persons. The lowest level is added to estimate the residual variance, the item and person levels are 'dummy'-levels, with only one unit that covers the entire data set (cf. Hox, 2002). At the lowest level the items are represented by a full set of dummies. The fixed coefficients of these dummies are excluded from the model, but their slopes are allowed to vary at the second (item) 'dummy'-level. The covariances between these dummies are all constrained to zero, and their variances are all constrained to be equal. Thus, we estimate one variance component for the items. The specification of the third, the person level, is similar. At the lowest level we obtain a direct estimate of the residual variance, at the second level the item variance is estimated and at the third level the person variance. The estimates are exactly the same as the variance components estimated with Anova. Since the software specification for the multilevel approach requires as many dummy variables as there are subjects in the data set, it is clear that data of a realistic size and complexity pose severe difficulties.

A special case of the crossed facet designs is the situation in which persons only partially respond to the same items (see Table 6, as a special case of table 5). Analyzing these data as a crossed design with the Anova approach is not feasible, because of the empty cells in the observed dataset. Multilevel analysis of these data is straightforward. Following the same procedure as described above for the one facet crossed design, estimates for the variance components for the items, persons, and residual are obtained.

The variance components are estimated as $\hat{\sigma}^2_p = 0.006$, $\hat{\sigma}^2_i = 0.019$ and $\hat{\sigma}^2_{residual} = 0.239$. Two percent of the variance is associated with persons, seven percent with items, and the remainder with the interaction and error.

Conclusions

Multilevel models can be especially useful when measures are constructed according to a logic that confers specific characteristics to the measures. We discuss facet design as an example, but other systematic question construction approaches result in similar data. If the measures can be assigned values on specific variables, multilevel models can be used to analyze the effect of both person and question characteristics on the responses. For those question characteristics whose effects vary across persons, residuals or posterior means can be assigned to persons as 'scores' on these characteristics. A second area where multilevel models are useful for measurement is when contextual characteristics must be assessed. We discuss the example of pupils rating the school manager. Various multilevel models can be utilized to assess the reliability and validity of such ratings at specific levels of the hierarchy. Multilevel modeling is useful in generalizability theory only if the design results mostly in nested data sets; data sets with a large number of crossed facets lead to large cross-classified data sets which current multilevel software does not handle well.

The measurement procedures outlined above are based on classical test theory, which means that they assume continuous multivariate normal outcomes. Most test items are categorical. If the items are dichotomous, we can use logistic multilevel modeling. If there are two levels, the item level and the person level, multilevel logistic regression is equivalent to a Rasch model (Andrich, 1988; Kamata, 2001; Rost & Walter, this volume).

A nice feature of using multilevel models for measurement scales is that it automatically accommodates incomplete data. If some of the item scores for some of the pupils are missing, this is compensated in the model. The model results and estimated residuals or posterior means are the correct ones, under the assumption that the data are Missing At Random (MAR). This is a weaker assumption than the Missing Completely At Random (MCAR) assumption than is required with simpler methods, such as using only complete cases or replacing missing items by the mean of the observed items. The MAR assumption requires that the missing data are missing completely at random, conditional on the available observed data. Since items typically correlate

highly, the assumption that conditional on the available item scores any missed items are missing completely at random is reasonable. An interesting application is to assign different subsets of items to different subsets of persons by design. In this case the missingness is MCAR, and multilevel analysis provides a straightforward means to estimate the individuals' scores as the person-level residuals or posterior means for the intercept. The usual estimates in multilevel modeling are empirical Bayes estimates, shrunken towards the overall mean, which are equivalent to the true score in classical test theory (cf. Lord & Novick, 1968; Nunnally & Bernstein, 1994).

References

Andrich, D. (1988). Rasch-models for measurement. Newbury Park, CA: Sage

Borg, I. (1994). Evolving notions of facet theory. In: I. Borg & P.Ph. Mohler (Eds.), *Trends and perspectives in empirical social research*. Berlin/New York: De Gruyter, pp.178-200.

Borg, I. & Shye, S. (1995). *Facet theory: form and content*. Newbury Park, CA: Sage.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 546-553.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N., (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: John Wiley.

Fiske, D.W. (1971). *Measuring the concepts of personality*. Chicago: Aldine.

Goldstein, H. (1995). *Multilevel Statistical Models*. London: Edward Arnold/New York: Halsted.

Gough, G. (1985). Reasons for slimming and weight loss. In: D. Canter (Ed.), *Facet theory*. New York: Springer, pp. 245-259.

Guttman, L. (1954). An outline of some new methodology for social research. *Public Opinion Quarterly, 18,* 395-404.

Hox, J.J. (2002). *Multilevel analysis. Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hoyle, R.H., Harris, M.J. & Judd, C.M. (2002). *Research methods in social relations*. Thomson Learning.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38,* 79-93.

Kerlinger, F.N. (1973). *Foundations of behavioral research*. New York: Holt, Rinehart and Winston.

Krüger, M. (1994). Sekseverschillen in schoolleiderschap. Alphen a/d Rijn: Samson. [Gender differences in school leadership.]

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.

Mellenbergh, G.J., Kelderman, H. Stijlen, J.G. & Zondag, E. (1979). Linear models for the analysis and construction of instruments in a facet design. *Psychological Bulletin*, *86*, 766-776.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Paterson, L. (1998). Multilevel multivariate regression: an illustration concerning school teachers'perception of their pupils. *Educational Research and Evaluation*, *4*, 126-142.

Raudenbush, S.W & Bryk, A.S. (2002). *Hierarchical linear models*. Newbury Park, CA: Sage.

Raudenbush, S., Bryk, A., Cheong, Y.F. & Congdon, R. (2000). *HLM 5. Hierarchical Linear and Nonlinear Modeling*. Chicago: Scientific Software International.

Raudenbush, S.W., Rowan, B., & Kang, S.J. (1991) A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics*, *16*, 4, 295-330.

Raudenbush, S. W., & Sampson, R. (1999). Assessing direct and indirect associations in multilevel designs with latent variables. *Sociological Methods and Research*, *28*, 123-153.

Roskam, E.E. (1990). Formalized theory and the explanation of empirical phenomena. In J.J. Hox & J. de Jong-Gierveld (Eds.), *Operationalization and research strategy* (pp. 179-198).

Sampson, R., Raudenbush, S.W., & Earls, T. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, *227*, 918-924.

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability Theory: A Primer*. Thousand Oaks, CA: Sage Publications.

Snijders, T.A.B., & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

The extend that person [X] feel that

Source		*Reason*
her own experience her husband her doctor the media	led her to believe that she would	feel healthier feel fitter be more physically attractive have fewer clothing problems suffer less social stigma be less anxious in social situations feel less depressed

Response:

if she lost weight, is rated as → [not really at all; not very much; to a slight degree; to a fair degree; quite a lot; very much; very much indeed]

Figure 1. Mapping sentence for attending slimming classes

idnr	s1m1	s1m2	s1m3	s1m4	s1m5	s1m6	s1m7	s2m1	s2m2	s2m3
1	4	4	5	3	3	4	3	4	5	5
2	4	5	5	4	3	3	3	3	4	4
3	4	5	5	4	4	4	3	5	5	6
4	4	4	5	5	4	4	5	4	5	5
5	3	4	4	4	2	2	3	3	4	4
6	3	4	4	3	2	2	1	4	4	5
7	4	5	5	4	4	4	3	4	4	5
8	7	7	7	5	5	6	5	5	6	6
9	5	5	5	5	3	4	3	4	5	5
10	4	5	5	4	3	4	3	4	4	5

Figure 2. Part of data for reasons for slimming design.

Table 1. Multilevel analysis of reasons for attending weight reduction classes

Model: Only item characteristics				Model: Item characteristics + Age of respondent		
Regression slopes				*Regression slopes*		
Predictor	slope (s.e.)	*p*		slope (s.e.)	*p*	
Self	3.31 (.11)	.00		3.31 (.11)	.00	
Husband	3.42 (.10)	.00		3.42 (.10)	.00	
Doctor	3.30 (.11)	.00		3.30 (.10)	.00	
Media	2.91 (.12)	.00		2.91 (.11)	.00	
Health	0.96 (.07)	.00		0.96 (.07)		
Fitness	1.38 (.07)	.00		1.38 (.07)		
Attract.	1.76 (.07)	.00		1.76 (.07)		
Clothing	0.70 (.07)	.00		0.70 (.07)		
Stigma	0.24 (.07)	.00		0.24 (.07)		
Anxious	0.35 (.07)	.00		0.35 (.07)		
Age*Self		-		0.03 (.01)	.00	
Age*Husband		-		0.02 (.01)	.01	
Age*Doctor		-		-.02 (.01)	.01	
Age*Media		-		-.02 (.01)	.00	
Variances		χ^2 (*df*=49)	*p*		χ^2 (*df*=48)	*p*
Self	0.39	303	.00	0.31	235	.00
Husband	0.36	281	.00	0.31	249	.00
Doctor	0.40	308	.00	0.36	278	.00
Media	0.51	381	.00	0.44	320	.00
σ_e^2	0.53	-		0.52		

Table 2. Intercept and variances for school manager data (Data from Krüger, 1994)

Fixed part	Coefficient	s.e.
Intercept	2.57	0.05
Random part	**Variance component**	
$\pm^2_{school}$	0.179	
$\pm^2_{pupil}$	0.341	
$\pm^2_{item}$	0.845	

Table 3. Covariances and correlations of the semantic differential scales at the pupil and class level (simulated data).[a]

Covariances and correlations at the pupil level						
	1	**2**	**3**	**4**	**5**	**6**
1 Eval. Pup.	0.377	*.02*	*.04*	-	-	-
2 Act. Pup.	.007	.372	*-.02*	-	-	-
3 Pot. Pup.	-.013	-.006	.372	-	-	-
4 Eval Tch.	-	-	-	-	-	-
5 Act. Tch.	-	-	-	-	-	-
6 Pot. Tch.	-	-	-	-	-	-

Note: the italic entries in the upper diagonal are the correlations

Covariances and correlations at the class level						
	1	**2**	**3**	**4**	**5**	**6**
1 Eval. Pup.	.269	*.29*	*.01*	*.66*	*.14*	*-.04*
2 Act. Pup.	.069	.211	*.23*	*.22*	*.64*	*.06*
3 Pot. Pup.	.004	.061	.339	*-.01*	*.20*	*.64*
4 Eval Tch.	.441	.128	-.007	1.671	*.16*	*.10*
5 Act. Tch.	.090	.378	.145	.269	1.633	*.10*
6 Pot. Tch.	-.020	.028	.401	.134	.136	1.167

Note: the italic entries in the upper diagonal are the correlations

[a] Note: Item-level variance is 0.921

Table 4 Item scores for eight persons on four items

	Item			
Person	1	2	3	4
1	0	0	1	1
2	0	1	0	1
3	0	0	1	0
4	1	0	0	1
5	0	1	1	1
6	0	0	1	0
7	1	0	1	1
8	1	1	0	0

Table 5. Item scores from eight persons on sixteen multiple choice items

	Item															
Person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0	0														
2			0	1												
3					1	0										
4							0	0								
5									1	1						
6											1	1				
7													0	0		
8															1	0

Table 6 Item scores from four persons on eight multiple choice items with incomplete data

	Item			
Person	1	2	3	4
1	0	0	1	
2	0	1	0	
3		0	0	1
4	1	0	0	
5		1	1	1
6		0	1	0
7		0	1	1
8		1	0	0