KM 33(1990)
pag 67 - 88

Factorial Surveys; Multilevel by Design

J.J. Hox[1], G.G. Kreft[2], P.L.J. Hermkens[3]

Summary

Factorial surveys constitute a specific technique for introducing experimental designs in sample surveys. Respondents are presented descriptions (vignettes) of a constructed world, in which important factors are built in experimentally. Using balanced designs well known from the multivariate experimental tradition, it is possible to build in a relatively large number of factors and levels.

Factorial surveys are mostly used in research on social judgments. Within this context, the normal hypothesis is that such judgments are consistent on the individual level, but not totally idiosyncratic. In the analysis of these judgments, it is important to determine the influence of both the vignette and the respondent variables. Analysis models for this type of data should reflect the fact that factorial surveys produce data pertaining to two distinct levels: the individual level and the vignette level. Such models are available, and are general known as multilevel analysis models.

This article discusses the properties of factorial survey designs and some analysis models which address the multilevel aspects of the data. An example is presented using data on judgments on the fairness of incomes.

---

[1]*Faculty of Educational Sciences, University of Amsterdam*
[2]*Faculty of Psychology, University of California, Los Angeles*
[3]*Faculty of Social Sciences, University of Utrecht*

Correspondentieadres: J.J. Hox, Faculteit PAOW,
Prinsengracht 227
1015 DT Amsterdam
Tel (31)(20)(5253315)
EARN: A716HOX @ HASARA11

A characteristic attribute of many problems in social research is that both the variables of interest and the explanatory variables may be realized at a number of different levels.

As a rule, in multilevel research the data have a hierarchical nature because they reflect the properties of some hierarchical system in the social world. For instance, in most countries the educational system has a hierarchical structure; pupils are collected in social units, classes and schools, which are in turn part of larger units, such as school districts, and so on. Consequently, the multilevel problem has received much attention in educational research, where 'outcome' variables (such as school career, examination results, drop out) are often observed at the pupil level, and explanatory variables are sought at both the pupil level (e.g. IQ, parent's education) and the class or school level (e.g. resources, class size, teacher's attributes) (Burstein, 1980; Kreft, 1987).

In the examples given above, the multilevel features of the data are forced upon the investigator by the hierarchical nature of the social system studied. There is, however, another class of multilevel problems, in which the multilevel structure is not so much given by the data, but designed into the research by the researchers. One example is found in experimental psychology, for instance in experiments where subjects react to a series of stimuli. In this case the multilevel nature of the data is no real problem. The stimuli are varied according to an experimental design, and important subject variables such as age or sex are also under experimental control. Since all variables are (nearly) orthogonal, and the number of subjects is generally small, repeated measures analysis of variance can readily be used (cf. Kirk, 1969; Hays, 1973).

The multilevel problem becomes more prominent in designs where elements of experimental design are used in survey research. This occurs in facet research (cf. Guttman, 1954; Canter, 1985; Roskam, 1987), in methodological studies on questionnaire design (cf. Saris, 1987), and in factorial surveys (cf. Rossi & Nock, 1982).

Building an experimental design into a survey is a demanding exercise, which will be undertaken only when the research problem clearly demands it. A good example is found in the research projects reported by Saris (1987), where the central hypothesis is that the responses to attitude items depend on an unobserved response function, which is approximately linear and which may be different for each respondent. Saris c.s. report a number of studies, where the format of the questions is carefully manipulated, and for each format individual response functions are estimated by linear regression. In a second step, the parameters of the individual response functions are analyzed, generally by analysis of variance or multiple regression analysis.

Clearly, the problem here is a problem of statistical design, rather than a conceptual problem. The link between theory and analysis is straightforward. In this article, we will discuss the multilevel problems that occur in this type of research within the context of factorial surveys, which is more general than that of methodological research. We will use a data set analyzed previously by Hermkens (1986), both with a single level model and with a two step procedure. This example also highlights an important difference between multilevel models as they are generally used in educational research, and multilevel models in survey research. Because in factorial surveys the multilevel structure is there by design, the explanatory variables at the different levels are (nearly) orthogonal, which eliminates the multicollinearity problems often found in educational research. However, since the data modeled are based on repeated measures collected in one session, the chances are that the dependencies in the data are substantial.

THE FACTORIAL SURVEY

Factorial surveys are a specific application of experimental design in survey research (Rossi & Nock, 1982). In a factorial survey, the respondents are presented questions based on descriptions of a hypothetical, or constructed, world. These descriptions, sometimes called vignettes, contain a sketch of a hypothetical object (e.g. a person or situation). The respondent is asked to rate the objects on a specified dimension. The descriptions in the vignettes are constructed systematically, based on factors thought to be relevant to the judgment process. For example, in a study of the social status of households, Rossi (1974: 175) used vignettes in which the educational level and occupation of both husband and wife were systematically varied. A typical vignette from such a study would be:

Husband	Wife
Architect	Artist
College	High School

The factorial survey approach combines the advantages of an experimental design with the advantages of the survey tradition. A large number of respondents can be sampled, following one of the well known survey sampling designs (Kish, 1987). These respondents can be presented with interview questions in which the factors underlying the vignettes are orthogonally varied, following one of the well known experimental designs (Kirk, 1969). Generally, the respondents each judge a different subset of all possible vignettes; using random sampling of vignettes or a balanced design it is possible to build in a large number of factors and levels.

Factorial surveys have been used in research on judgments of complex social objects, in which the vignettes vary on a great number of dimensions. Within this context, it is assumed that individuals select a relatively small number of factors to base their judgments on, and that individuals follow consistent rules to make those judgments. These rules may be partly idiosyncratic, but it is assumed that social judgments will be at least partly socially determined. That is: both individual characteristics and vignette variables are assumed to have an effect on the judgments.

ANALYSIS MODELS FOR FACTORIAL SURVEY DATA.

Single level analysis

The analysis method mostly used in factorial survey research is analysis of variance or multiple regression analysis. The factors underlying the vignette are the independent variables, and the judgment of the vignette is the dependent variable. Combining all judgments from all respondents, an analysis is made to determine the effect of the vignette factors on the judgments. This analysis is followed by an analysis which includes respondent characteristics. This is normally done either by repeating the analysis on different subsets of the sample, or by including respondent characteristics in the regression equation.

In the simplest case, respondents are assumed to be completely interchangeable, and only vignette characteristics are entered in the regression equation:

$$Y_{ij} = a + \Sigma\, b_l V_{li} + e_{ij} \qquad [1]$$

where Y_{ij} is the judgment of person j on vignette i

a is the intercept for the sample

b_l is the coefficient for vignette attribute l

V_{li} is the value of vignette attribute l for vignette i

e_{ij} is the error term

This model can be expanded to include respondent characteristics and interactions between vignette and respondent characteristics, as follows:

$$Y_{ij} = a + \Sigma\, b_l V_{li} + \Sigma\, c_k R_{kj} + \Sigma\, d_{lk} V_{li} R_{kj} + e_{ij} \qquad [2]$$

where Y_{ij} is the judgment of person j on vignette i
a is the intercept for the sample
b_l is the coefficient for vignette attribute *l*
V_{li} is the value of attribute *l* for vignette i
c_k is the coefficient for respondent attribute k
R_{kj} is the value of attribute k for respondent j
d_{lk} is the coefficient for the interaction between vignette attribute *l* and respondent attribute k
e_{ij} is the error term

Here, the beta coefficients for the vignette characteristics reflect effects that are common to all respondents in the sample. Rossi and Anderson (1986: 21) designate these as the social component of the judgments. The beta coefficients for the respondent characteristics predict the judgment thresholds of different subjects. The interaction terms reflect idiosyncratic judgment rules of different subsets of the sample. Another version of this model would be the analysis of covariance model, which omits the interaction terms and models a different judgment threshold for each respondent.

Two step analysis

A different approach to include respondent characteristics in the analysis is the two step procedure. In this procedure a separate multiple regression equation is computed for each respondent:

$$Y_{ij} = a_j + \Sigma\, b_{jl}V_{li} + e_{ij} \qquad [3]$$

where Y_{ij} is the judgment of person j on vignette i
a_j is the intercept for individual j
b_{jl} is the regression coefficient for respondent j of vignette attribute *l*
V_{li} is the value of vignette attribute *l* for vignette i
e_{ij} is the error term

This creates as many regression equations as there are respondents. Since the regression equation is computed for each separate individual, the regression parameters refer to his or her judgment process only. The individual R^2 indicates the amount of structure in the judgment process; individuals making random judgments would have a R^2 close to 0. In most applications, the term $1-R^2$ is used, indicating the amount of random error in the judgments of individual respondents.

The second step in the procedure is to treat the individual vectors of regression parameters as characteristics describing the respondents that have generated them. For instance, Rossi and Anderson suggest factor analysis (Q-method) or cluster analysis on the individual regression parameters as a means of grouping respondents. Individual differences in the judgment process can be modeled by taking the regression parameters from the first step, and using them in a second regression analysis, in which they are regressed on respondent characteristics. Typically, in the second step, each first step parameter would be taken as a dependent variable in a separate regression analysis, but it is also possible to use a simultaneous procedure such as LISREL.

Statistical considerations

The regression models above are a specific case of the fixed effect linear model. Parameters are point estimates, estimated with a standard error. If the standard error is small, compared to the value of the parameter, we conclude that the estimated value corresponds with a significant effect. This significance test is based on the assumptions of the model and is only valid in so far as these assumptions are approximately true, or at least not too badly violated. The assumptions of the model are: cases are independently sampled, from normally distributed populations, differing only in their means. Unmeasured effects and measurement errors are included in the error term e_{ij}, which has a mean of zero and a variance equal to s^2, which is the same for all subjects. The covariance between individual error terms is assumed to be zero.

As a consequence of the nesting of the vignettes within respondents, the data in a vignette design does not follow the assumptions of the traditional linear model. In fact, most of the assumptions above will be violated. The most critical assumptions are that error terms are uncorrelated and units of analysis are independently sampled. Since a number of judgments is given by each single subject, unmeasured subject variation will to an unknown extend cause correlated error terms within individuals. The assumption of independent sampling is violated, because vignettes judged by the same respondent will have values for respondent variables which are necessarily exactly equal.

Both point estimates and standard errors will be biased. The standard errors will be underestimated, particularly for the respondent variables. Even the signs of the regression coefficients may be misleading ((Kreft, 1987; Kreft & De Leeuw, 1988).

The two step procedure, also known as the 'slopes-as-outcomes' approach, consists of a micro model (e.g. a regression analysis for the individual respondents) and a macro model, which relates the parameters of the micro models (generally regression coefficients and error variances) to the macro level variables. Thus, within respondents regression coefficients for vignette variables are predicted by respondent variables in the macro model. The error variances, or, conversely, the percentage of variance explained, can be included to give information about differences in the amount of stochastic error.

The basic problem of the two step procedure is the use of identical assumptions and an identical error distribution in both steps, whereas the distribution of the parameter estimates in the second step is fully determined by the assumptions of the first step. If the usual linear model assumptions are true in the first step, they will not be true in the second step. The error structure will generally be quite different from the error structure assumed by the linear model. Some specific problems with the slopes-as-outcomes approach are (Raudenbush & Bryk, 1986; Kreft, 1987): 1) regression coefficients have a greater sampling variability than means; 2) the sampling variance of regression coefficients is dependent on both the sample size (in our case the number of vignettes used, which may vary across respondents due to missing values), and the error term for the individual respondent (which is certain to vary across individual respondents). Consequently, there is no simple sampling distribution for these parameters. What we have here is a mixture of as many sampling distributions as there are respondents in the sample. Using ordinary standard errors in the second stage is not only inefficient because of incomplete specification (Tate & Wonbundhit, 1983), but also wrong (De Leeuw & Kreft, 1986; Kreft, 1987).[4]

Random coefficient models

The model presented here for the analysis of factorial surveys is a fairly general model, which is known under various names, such as the 'random coefficient' model (De Leeuw & Kreft, 1986), and 'variance decomposition' or 'variance component' model (Aitkin & Longford, 1986). The variance component model, off course, is well-known. To highlight the fact that we are using a specific case of this model, one involving nested data, we will use the term 'random coefficient model' throughout this article.

---

[4] In a technical appendix Rossi & Anderson (1982) note that the error structure of the two step model may pose a problem. However, they assume that the ordinary linear model is robust against the type of violations incurred in a two step analysis of a factorial survey design.

If we substitute the respondent level model in the vignette level model we find:

$$Y_{ij} = a_{j,0} + \Sigma\, a_{j,1}R_j + \\ + \Sigma\, b_{j,0}v_{ij} + \Sigma\Sigma\, b_{j,1}R_j v_{ij} + \\ + D_{j,1} + D_{j,2}v_{ij} + E_{ij} \qquad [4]$$

This model is similar to the simple multiple regression model in which all parameters are specified at the vignette level. The important difference is the complicated error structure, incorporating disturbance terms for each level and the $D_{j,2}v_{ij}$ disturbance term.

EXAMPLE I, COMPARING SINGLE LEVEL, TWO STEP, AND MULTILEVEL MODELS

In this example we will use data from a study by Hermkens (1984A, 1984B) on judgments about the fairness of the distribution of incomes. For simplicity, in the following example only a small selection of the variables is used: education and household size at the vignette level, and education, household size, and income at the respondent level. Also, vignettes describing single person households and single respondents were omitted from the analysis. This results in a sample of 107 respondents, each judging 16 vignettes.

Results multiple regression analysis

The results of an ordinary multiple regression analysis, combining vignette and respondent variables in one analysis, are in Table 1 and Table 2. Table 1 presents the results of adding respondent and interaction variables to the vignette variables:

Table 1. Summary of stepwise regression analyses

Level	Cum. Per. Var.	'p'
Vignette	18.5	.00
Respondent	22.4	.00
Interaction	22.5	.47

The interaction term in Table 1 is clearly not significant. Since including the interaction variables in the multiple regression results in high multicollinearity, the regression coefficients in Table 2. were computed without the interaction terms.

Table 2. Results multiple regression analysis

Dependent Var.: Fair income. R = .47; 'df' = 5/1585

Variable	Level	b	beta	'p'
V-Edu	Vignette	218	.39	.00
V-Size	Vignette	106	.15	.00
R-Edu	Respondent	19	.04	.13
R-Size	Respondent	-83	-.12	.00
R-Inc	Respondent	.21	.15	.00

The results in Table 1. and Table 2. are clear. Vignette variables explain most of the variance. Respondents assign higher incomes to vignettes which describe earners with a high education and a large household. Respondent education is not important. Respondents with relatively large households and income assign higher incomes. There is no interaction, which implies that the judgment process itself does not depend on respondent education, household size, or income.

Results two step approach

In the first step 107 separate regression equations were computed, one for each respondent, predicting the assigned incomes by vignette education and household size. The results are summarized in Table 3:

Table 3. Summary of regression analyses

Coefficient	Mean	St. Dev.	Z
Intercept	1250	185	27.0
b-Edu	172	185	3.7
b-Size	231	153	6.0

If we take the distribution of the individual parameters to be normal, and compute the standard normal test statistic Z, all means differ significantly from zero.

In the second step, the individual regression parameters are predicted by the respondent variables education, household size, and income:

Table 4. Second step regression analysis

Predicting:	b-Edu		b-Size		Intercept	
Predictor	beta	'p'	beta	'p'	beta	'p'
R-Edu	.04	.71	-.06	.56	.09	.40
R-Size	.08	.42	-.07	.47	-.14	.15
R-Inc	.27	.00	-.13	.18	-.03	.76
	R = .29		R = .17		R = .18	

One interaction seems to be significant: respondents with a high income have steeper slopes for the vignette variable education. The nonsignificant effects for the intercept imply that the respondent variables education, household size, and income, do not influence the assigned income.

The results from the two step analysis are, again, clear. Unfortunately, they are inconsistent with the results obtained in the combined multiple regression analysis (cf. Table 2).

Results random coefficient model

For the random coefficient model we will use the following notation:

ve: vignette education
vs: vignette household size

re: respondent education
rs: respondent household size
ri: respondent household income

Subscript i is used for vignettes; j is used for respondents; bold, upper case letters are used for random variables; lower case letters are used for fixed constants.

The vignette level model is:

$$\mathbf{Y}_{i,j} = \mathbf{A}_j + \mathbf{B}_j ve_{i,j} + \mathbf{C}_j vs_{i,j} + \mathbf{E}_{i,j}. \qquad [5]$$

Respondent level model:

$$\mathbf{A}_j = a_{j,0} + a_{j,1}re_j + a_{j,2}rs_j + a_{j,3}ri_j + \mathbf{D1}_j, \qquad [6]$$
$$\mathbf{B}_j = b_{j,0} + b_{j,1}re_j + b_{j,2}rs_j + b_{j,3}ri_j + \mathbf{D2}_j, \qquad [7]$$
$$\mathbf{C}_j = c_{j,0} + c_{j,1}re_j + c_{j,2}rs_j + c_{j,3}ri_j + \mathbf{D3}_j. \qquad [8]$$

If we substitute the respondent level model in the vignette level model we obtain:

$$Y_{i,j} = a_{j,0} + a_{j,1}re_j + a_{j,2}rs_j + a_{j,3}ri_j + \\ + b_{j,0}ve_{i,j} + b_{j,1}re_jve_{i,j} + b_{j,2}rs_jve_{i,j} + b_{j,3}ri_jve_{i,j} + \\ + c_{j,0}vs_{i,j} + c_{j,1}re_jvs_{i,j} + c_{j,2}rs_jvs_{i,j} + c_{j,3}ri_jvs_{i,j} + \\ + D1_j + D2_jve_{i,j} + D3_jvs_{i,j} + E_{i,j}. \quad [9]$$

This model is similar to the multiple regression model presented above in formula [2]. The estimated coefficients, corresponding to the coefficients in formula [9], are given in Table 5:

Table 5. Results full random coefficient model

Coefficient	Estimated value	Significance
a0	2002	*
a1	-12	
a2	-151	*
a3	.05	
b0	-7	
b1	5	
b2	19	
b3	.07	*
c0	108	*
c1	.70	
c2	13	
c3	-.02	
d1	483	*
d2	177	*
d3	75	*
e	679	-

The full model has many parameters which do not differ significantly from zero. A more specific model may fit almost as well. Such a model may be chosen based on apriori statistical considerations, on substantive knowledge of the subject area or the research design, and a posteriori by assessing the goodness-of-fit of the model. A model fitting procedure, based on the full set of variables used in the study, is discussed in the next section.

EXAMPLE II: MULTILEVEL MODELS

In this example we will again use data from the study by Hermkens (1984A, 1984B, 1986) on judgments about the fairness of the distribution of incomes. In this study, hypothetical households were described by combinations of attributes noted in previous research on distributive justice. The following attributes were used: occupation, educational level, source of income, sex, household size, and age. 127 respondents were presented with a random sample of 24 vignettes, and were requested to assign the income which they judged to be 'fair' for this specific household. Some of the respondent variables were: occupation, educational level, marital status, sex, number of persons in the household, age, and income.

As stated before, a number of different models may be chosen from the general class of random coefficient models. It should be observed, however, that the choice of model will generally have implications for the quantitative assessment of the effects estimated. Therefore, models should be chosen carefully, and the choice should be based as much as possible on substantive knowledge of the subject area or the design of the study. In factorial surveys the usual assumption is that individual judgments are composed of an individual, idiosyncratic component, and a social component. The social component of judgments can be modeled by aggregating judgments of respondents and estimating regression coefficients for vignette attributes. The individual component of social judgments includes several different effects (Rossi & Anderson, 1982). An important distinction is that between judgment threshold and judgment process. Subgroups or individuals may differ in their thresholds; in the terminology of the random coefficient model this means that subjects have different intercepts. Differences in the judgment process imply variations in the slopes of the regression equations for subgroups or individuals.

Following Hermkens (1986) vignettes describing married and singles were analyzed separately, using the VARCL program (Longford, 1986). Table 6. presents the result for the full model, with fixed effects estimates.

Table 6. All variables fixed, random intercept.

Variable	Estimate	St. Error	
G. MEAN	892		
V-age	3.2	1.34	*
V-size	111.5	42.49	*
V-edu	21.4	55.24	
V-paid work	229.0	39.49	*
V-sources inc.	309.7	23.04	*
V-SES	.6	5.64	
V-size x R-size	12.1	8.64	
V-edu x R-size	5.0	10.44	
V-SES x R-size	.7	1.07	
V-size x R-inc	-.0	.02	
V-edu x R-inc	.0	.02	
V-SES x R-inc	.0	.00	
R-sex	-24.6	111.71	
R-married	116.4	45.22	*
R-edu	-40.71	43.18	
R-paid work	-109.1	121.12	
R-SES	45.9	46.71	
R-age	-.6	3.81	
R-size	-153.0	55.11	*
R-inc	.1	.11	
R-fair income	-13.8	52.61	

In Table 6. four variables are significant at the vignette level: age, household size, receiving pay for work (versus receiving money from social security), and number of income sources (e.g. two wage earners in the household). Apparently, our respondents assign larger incomes to older people with larger households. There are no interaction effects, and only two judge characteristics have an effect: being married or not, and the respondents own household size. The variance components for the model in Table 6 are:

Table 7. Variance component estimates (cf. Table 6).

Level	Variance	Sigma
Vignettes	469843.4	685.45
Respondents		
G. Mean	184393.8	429.41
Deviance	26757.58	

The next model, in which we do not assume that the slopes are equal for all judges, shows partially different results. In this model some variables are dropped, and random intercepts are specified for three vignette characteristics: household size, SES, and education. Six interactions are tested with these vignette characteristics and the respondents characteristics household size and income. The slope of vignette age is fixed. This model has a deviance of 26830. The comparable model with all slopes fixed and only a random intercept has a deviance of 27034. With a difference of 204, with only nine more parameters to estimate, the random model fits the data much better than the fixed slopes model. The results of the random model are reported below in Table 8:

Table 8. Mixed model, random intercept.

Variable	Estimate	St. Error	
G. MEAN	2085		
V-age	-2.0	1.26	
V-size	93.2	50.67	
V-edu	13.2	64.36	
V-SES	.1	5.89	
V-size x R-size	12.4	10.21	
V-edu x R-size	13.8	12.44	
V-SES x R-size	.1	1.14	
V-size x R-inc	-.0	.02	
V-edu x R-inc	.0	.03	
V-SES x R-inc	.005	.0023	*
R-size	-140.8	63.72	*
R-inc	-.1	.13	

In the random model all vignette effects disappear. The only significant effects are the respondents household size and an interesting interaction effect of the income of the respondent on the SES of the vignette. The higher the income of the respondents is, the steeper the slope for the effect of vignette SES on assigned income. Respondents with a lower income are more egalitarian in this respect than respondents with a higher income.

DISCUSSION

A comparison of the results of the different models above leads to the conclusion that they give results which would lead to quite different substantive conclusions. One obvious recommendation is that if one uses factorial surveys, one should recognize that this involves a complex hierarchical research design. The statistical model used to analyze the results should reflect this complexity, and be specified as such. The data reported here clearly show that in this case one cannot rely on the 'robustness' of a simplified model.

The design of a factorial surveys leads to hierarchical, or nested data. In principle, such data can readily be analyzed by standard ANOVA techniques (cf. Kirk, 1969; Searle, 1971; Hays, 1973). However, in practice, problems will arise. Since in factorial surveys each respondent judges his/her individual set of stimuli, the design is not a simple repeated measurements factorial design, but a nested design. Also, following the survey tradition, the sample size is generally large. As consequence, standard ANOVA programs will generally not do to analyze the data. The analysis of hierarchical data is complicated, and only recently computer programs have become available, which are both statistically correct and sufficiently powerful to be useful in actual research (Raudenbush & Bryk, 1986; Aitkin & Longford, 1986; DeLeeuw & Kreft, 1986).

Random coefficient models seem to be a good choice for this kind of data. They are parsimonious and based on realistic assumptions about the structure of the data. Both main effects and interactions can be tested for significance. If models are nested, the significance of their difference can be tested in the usual way by a chi-square statistic. The partitioning of the variance in between respondents and within respondents/between vignettes will give information on the feasibility of improving the model by including other variables. For instance, the significance of the respondent level disturbance terms d1 to d3 in the first example suggests ample room for improvement of the model. In this case, we know this to be true, because the data come from a vignette study in which more vignette variables were manipulated than have been included in this example.

The conclusion that random coefficient models are a natural choice for this type of data, does not mean that they are the only possibility. For one thing, it is attractive to apply the principles set forth by Cronbach et al. (1972) to develop indices for the generalizability of factorial survey results over different sources of error variance. Coefficients of generalizability or replicability have been applied to a variety of designs (Mellenbergh, 1977; Mellenbergh & Saris, 1982; DeLeeuw & Hox, 1983); in a factorial survey variance components estimated by multilevel procedures should be used to calculate these coefficients.

Other models that could be used to analyze factorial survey data include item response models, such as the linear logistic trait model (Fisher, 1974) or the models used in multiple matrix sampling of test items (Bock, Mislevy & Woodson, 1982; Mislevy, 1983). However, the restrictive nature of most item response models makes them less attractive for our purpose. Also, they are generally geared toward problems of test equating, while in factorial surveys one is generally interested in the size and direction vignette and respondent effects.

Since in factorial surveys the data structure is multilevel by design, the explanatory variables at the different levels are not correlated. In contrast to educational research, multicollinearity is not a problem. Additionally, in multilevel models reported in educational research, the amount of variance explained by the 'higher' levels is usually small. In our application, the amount of variance on the respondents level is sizable. This points to another specific feature of this type of data; the respondent variance, which is not controlled in our design, is large compared to the vignette variance. Finally, the large difference between the single level analysis and the multilevel analysis in the first example indicates highly correlated errors. In the multilevel analysis, this is correctly modeled. However, if the covariance between the repeated measures is high, the power for uncovering vignette effects may be very low. Factorial survey design has often been used to combine a large number of vignette characteristics in one study. This may be good in exploratory research, but for theory testing it is probably advisable to use a large number of vignettes together with a small number of vignette characteristics.

REFERENCES

Aitkin, M.A. & Longford, N.T. (1986). Statistical modeling issues in school effectiveness studies. **Journal of the Royal Statistical Society A**, 149, 1-43.

Bock, R.D., Mislevy, R.J. & Woodson, C.E. (1982). The next stage in educational assessment. **Educational Researcher**, 11, 4, 11-16.

Burstein, L. (1980) The analysis of multilevel data in educational research and evaluation. In: D. Berlinger (ed.) **Review of research in education** Washington: American Educational Research Association

Canter, D. (ed.) (1985) **Facet theory**. New York: Springer.

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

DeLeeuw, E.D. & Hox, J.J. (1983). Using coefficients of replicability to estimate the sensitivity of ratings of odours. In: **Multivariate Analysis in the Social and Behavioural Sciences, 1983**. Amsterdam: SISWO.

DeLeeuw, J. & Kreft, I. (1986). Random coefficient models for multilevel analysis. **Journal of Educational Statistics**, 11, 57-86.

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

Hays, W.L. (1973). **Statistics for the social sciences**. New York: Holt, Rinehart & Winston.

Hermkens, P.L.J. (1983). **Oordelen over de rechvaardigheid van inkomens** (Judgments on the fairness of incomes). Ph.D. dissertation, University of Utrecht.

Hermkens, P.L.J. (1984). Rechtvaardig geachte inkomens (Incomes judged fair). Paper, working group on 'Income and Reward', Amsterdam, may 25.

Hermkens, P.L.J. (1986). Fairness judgments of the distribution of incomes. **The Netherlands' Journal of Sociology**, 22, 1, 61-71.

Kirk, R.E. (1969). **Experimental design: procedures for the behavioral sciences**. Belmont, Calif.: Brooks/Cole.

Kish, L. (1987). **Statistical design for research**. New York: Wiley.

Kreft, I. (1987). **Models and methods for the measurement of school effects**, Amsterdam: Ph.D. Dissertation, University of Amsterdam.

Kreft, I. & De Leeuw, E.D. (1988). The seesaw effect, a multilevel problem. **Quality and Quantity**, 22, 127-137.

Mislevy, R.J. (1983). Item response models for grouped data. **Journal of Educational Statistics**, 8, 271-288.

Mellenbergh, G.J. (1977). The replicability of measures. **Psychological Bulletin**, 84, 2, 378-384.

Mellenbergh, G.J. & Saris, W.E. The research design and model for a generalizibility study. **Proceedings 2nd Sociometric Congress.** Amsterdam: Sociometric Research Foundation.

Raudenbush, S. & Bryk, A.S. (1986). A hierarchical model for studying school effects. **Sociology of education**, 59, 1-17.

Roskam, E.E. (1987) Methodische aspecten van de vragenlijst. In: J. de Jong-Gierveld & J. van der Zouwen (eds.) **De vragenlijst in het sociaal onderzoek** (The questionnaire in social research). Deventer: Van Loghum Slaterus.
Rossi, P.H. & Anderson, A.B. (1982). The factorial survey approach: an introduction. In: P.H. Rossi & S.L. Nock (eds.) **Measuring social judgments**. Beverly Hills: Sage.
Saris, W.E. (ed) (1987) **Variation is response function: a source of measurement error**. Amsterdam: Sociometric research Foundation.
Searle, S.R. **Linear Models**. New York: Wiley, 1971.
Tate, R.L. & Wongbundhit, Y. (1983). Random versus nonrandom coefficient models for multilevel analysis. **Journal of Educational Statistics**, 8, 103-120.

Ontvangen: 27-10-1988
Geaccepteerd: 18-11-1989