KM13 (1984)
pag 148-155

The estimation of examiner effects in designs with overlapping examiner teams

Dato N.M. de Gruijter[*]

Summary

Examiners may differ in level and scale if they have to judge the quality of papers. This fact is particularly troublesome when papers of different examinees are judged by different examiner teams. In this paper the estimation of examiner effects is discussed for judgemental designs in which teams have examiners in common. An illustration with empirical data is provided.

Introduction

Sometimes papers of different examinees are judged by different teams of examiners. When in such a case examiners differ - they may, for example, differ in leniency - it matters to the examinees which team has judged their work. For that reason it seems useful to estimate examiner effects. When some of the effects are large, one can

- correct for them
- have some of the answers rejudged
- discuss and improve the instructions for examiners.

Here the estimation of examiner effects will be discussed when the teams consist of two independently judging examiners. It is assumed that the teams overlap in a special way: the teams should be composed in such a way that they cannot be divided into subgroups of teams having no examiner in common. So, each examiner is directly or indirectly linked to each of the other examiners. An example of such an overlapping design

[*] Bureau Onderzoek van Onderwijs, Rijksuniversiteit Leiden,
Boerhaavelaan 2, 2334 EN Leiden, tel. 071-148333, tst. 5392.

- a balanced incomplete block design with four examiners and four teams - is given in Figure 1.

		examinees			
		Group 1	Group 2	Group 3	Group 4
	1	xxx	xxx		
examiner	2		xxx	xxx	
	3			xxx	xxx
	4	xxx			xxx

Figure 1. An overlapping design with four pairs of examiners from a pool of four examiners.

In the general case there are n examiners, who are assigned to K different teams. In each team one of the examiners is arbitrarly chosen as the first team member. The judgemental design can be laid down in a matrix R(2xk) with typical element r_{ik}, the number of the examiner to whom position $i(i{=}1,2)$ in team $k(k{=}1,...,K)$ has been assigned. For the design of Figure 1 this matrix could read

$$R = \begin{pmatrix} 1 & 1 & 2 & 3 \\ 4 & 2 & 3 & 4 \end{pmatrix}.$$

Two models are discussed, the additive model (model A) and the linear model under the assumption of random assignment of examinees to teams (model B). So, it is assumed that the observed judgments are not so extreme as to cause disturbing ceiling and floor effects. A nonlinear model will be discussed elsewhere.

Model A

Model A can be written as

$$1) \qquad x_{pi(k)} = \tau_{p(k)} + \theta_{i(k)} + e_{pi(k)},$$

where $x_{pi(k)}$ is the judgment given by examiner r_{ik} to the script of examinee p in group k (assigned to team k), $\tau_{p(k)}$ the true score of the script, $\theta_{i(k)} = \theta_j$ the effect of examiner $j=r_{ik}$ and $e_{pi(k)}$ a random error. In order to obtain a unique solution for the θ's in the additive ANOVA decomposition, the restriction that the sum of all effects equals zero, i.e.

$$2) \quad \sum_{j=1}^{n} \theta_j = 0,$$

is introduced.

From Equation (1) one obtains

$$3) \quad d_k = x_{.1(k)}-x_{.2(k)} = \theta_{1(k)}-\theta_{2(k)} + e_{.1(k)}-e_{.2(k)}$$

where the dot indicates averaging over the examinees within group k. Least squares estimates of the θ's can be obtained from the minimization of

$$4) \quad S = \sum_{k=1}^{K} N_k(d_k-\theta_{1(k)}+\theta_{2(k)})^2,$$

where N_k is the number of examinees in group k, w.r.t. the θ's under restriction (2).

Model B

In this section it is assumed that the groups of examinees can be considered as random samples from a population of examinees. In this population the examiners are characterized by their means, μ_j, true score variances, β_j^2, and error variances, ϕ_j. It is assumed that the joint population distribution of judgements for examiners $1(k)$ and $2(k)$ of team k (k=1,...,K) is a bivariate normal distribution, with variance-covariance matrix

$$5) \quad C_k = \begin{pmatrix} \sigma_{11(k)} & \sigma_{12(k)} \\ \sigma_{12(k)} & \sigma_{22(k)} \end{pmatrix}$$

$$= \begin{pmatrix} \beta^2_{1(k)}+\phi_{1(k)} & \beta_{1(k)}\beta_{2(k)} \\ \beta_{1(k)}\beta_{2(k)} & \beta^2_{2(k)}+\phi_{2(k)} \end{pmatrix} .$$

This model can be regarded as a submodel of LISREL with structured means (Sörbom, 1982).

Using an iterative estimation procedure, one can obtain maximum likelihood estimates of the μ's, β's and ϕ's by the maximization w.r.t. the parameters of the log likelihood

$$6)\quad \log L=-\tfrac{1}{2}\sum_k N_k\{\log|C_k|+[\sigma_{22(k)}\overset{*}{s}_{11(k)}-2\sigma_{12(k)}\overset{*}{s}_{12(k)}+\sigma_{11(k)}\ \overset{*}{s}_{22(k)}]/|C_k|\}$$

where

$$\overset{*}{s}_{ij(k)}=s_{ij(k)}+(x_{.i(k)}-\mu_{i(k)})(x_{.j(k)}-\mu_{j(k)})\qquad (i,j=1,2)$$

with observed variances $s_{11(k)}$ and $s_{22(k)}$, and observed covariances $s_{12(k)}$ (cf. Jöreskog, 1970, p. 240). This can be achieved by maximizing log L w.r.t. all $\mu_{i(k)}$, $\beta_{i(k)}$ and $\phi_{i(k)}$ under the constraint that μ's, β's and ϕ's corresponding to the same examiner are equal. The model may, however, not be identified without further restrictions on the parameters; this is, for example, the case in the design of Figure 2.

Using further equality constraints, one can define various submodels like:

I. $\beta_j = \beta$ for all j

II. $\beta_j = \beta$
$\phi_j = \phi$ for all j

III. $\beta_j = \beta$
$\phi_j = \phi$ for all j.
$\mu_j = \mu$

The model without restrictions is the model of congeneric measurements, model I is the model of essentially τ-equivalent measurements and model III is the model of parallel measurements.

Instead of maximizing log L, one can minimize the function

$$7)\quad F = 2\log L-\sum_k N_k\{\log(s_{11(k)}s_{22(k)}-s^2_{12(k)})+2\},$$

where the second part comes from the minimum of the likelihood without any constraints within and between groups (cf. Jöreskog, 1970, equation 9). The minimum of F is asymptotically chi-squared distributed under the model assumptions, with degrees of freedom $5K-m$, m being the number of free model parameters. So, it is possible to obtain an indication of the adequacy of a model.

An empirical example

The estimation procedure can be illustrated with the help of data from an examination for a freshmen course in Internal Medicine.[*] Part of this examination was open-ended and judged by a group of ten examiners according to the design in Tabel 1. Examinees were randomly assigned to groups. It was decided to analyze the data from a subset of questions with a total score range from 0 to 30 and a subgroup of 153 examinees who answered at least five of the six questions.

Table 1. Team composition and numbers of examinees

group	examiners	number of examinees
1	1,2	17
2	2,3	12
3	3,4	14
4	4,5	16
5	5,6	16
6	6,7	16
7	7,8	15
8	8,9	17
9	9,10	16
10	10,1	14

The means, variances and covariances are given in Table 2.

[*] The author wishes to thank the staff of Internal Medicine of Leyden University, especially Dr. H.J.M. van Rossum, for making the data available and E.A. Bakker for his assistance with the data analysis.

Table 2. Statistics for the subgroups or teams

group	$x^{*}_{.1}$	$x_{.2}$	s_{11}	s_{22}	s_{12}
1	10.500	10.559	8.735	5.879	6.441
2	11.167	9.417	8.681	9.785	7.618
3	7.393	10.143	10.006	8.980	8.194
4	11.250	10.750	8.312	13.812	9.156
5	9.375	10.375	19.703	8.484	11.766
6	9.437	9.063	9.246	11.809	8.973
7	10.200	11.133	12.293	12.216	11.140
8	11.029	10.176	8.896	7.322	7.715
9	8.831	9.187	10.371	13.246	11.207
10	9.750	9.964	3.312	3.767	1.991

* Index 1 relates to the first examiner according to Table 1, index 2 to the second examiner

First, a model A analysis was performed. The resulting $\hat{\theta}$'s are given in Table 3.

Table 3. Estimates of θ for two models

examiner	$\hat{\theta}$(A)	$\hat{\theta}$(B;with prior)
1	0.1	0.2
2	0.0	0.2
3	-2.0	-1.8
4	0.5	0.6
5	-0.2	-0.1
6	0.7	0.5
7	0.1	-0.1
8	0.9	0.8
9	-0.1	-0.3
10	0.1	0.0

Next, several model B analyses were done. An analysis with equal β's and ϕ's resulted in a significant chi-square (χ^2=62.64,df=38). An analysis in which only the ϕ's were allowed to vary, resulted in a satisfactory fit (χ^2=28.12, df=29). The results were unsatisfactory for another reason, however: some of the $\hat{\phi}$'s were equal to zero, the lower bound which the computer program imposed on the error variances. This result probably is due to the small sample sizes involved (see also Boomsma 1983). Therefore the model B analysis was repeated with a common prior for the ϕ's. A normal prior with mean μ_0 and variance ϕ_0 was introduced for the γ_j=log ϕ_j. A uniform prior was chosen for μ_0 and a nearly non-informative $\chi^2(\lambda,\nu)$ prior for ϕ_0 with λ=0.1 and ν=0 (Paul, 1981). For this specification the $\hat{\phi}$'s strongly regressed to a common value: the estimates ranged from 1.06 to 1.31. Estimates $\hat{\theta}_j=\hat{\mu}_j-\hat{\mu}$. are given in Table 3.

The results of the analyses, presented in Table 3, are quite similar. This is not amazing since both analyses are based on the additive model, the second analysis differing from the first in that random sampling of examinees is assumed and differences between error variances explicitly are dealt with. The results in Table 3 indicate that one of the examiners, examiner 3, is severe in comparison to the other examiners.

Discussion

In this paper two approaches to the estimation of examiner effects for overlapping designs have been discussed: a quick and dirty approach for the additive model and a more sophisticated approach for a more general class of models. The latter approach seems not without problems when sample sizes are small, unless prior information is used. Here, only a common prior for the error variances has been used, but prior information on other parameters could be added in a way similar to the approach of Paul (1981) for a crossed design.

The results of analyses like the above could be very useful in a discussion of examiner instructions. In that case a simple estimate of examiner effects for all questions should be available, when more than one question is involved. The model A approach seems promising in that respect.

References

Boomsma, A. *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality*. Doct.Diss. Groningen, 1983.

Jöreskog, K.G. A general method for analysis of covariance matrices, *Biometrika*, 1970, *57*, 239-251.

Paul, S.R. Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology*, 1981, *34*, 213-223.

Sörbom, D. Structural equation models with structured means. In K.G. Jöreskog (Ed.) *Systems under indirect observation*. Part I. Amsterdam: North-Holland, 1982.

Ontvangen: 23-9-1983