Two Simple Models for Rater Effects

Dato N. M. de Gruijter
University of Leyden

In many examinations, essays of different examinees are rated by different rater pairs. This paper discusses the estimation of rater effects for rating designs in which rater pairs overlap in a special way. Two models for rater effects are considered: the additive model and a nonlinear model. An illustration with empirical data is provided.

Rater effects can be troublesome when different examinees are judged by different rater teams. For this reason it is important to obtain estimates of rater effects. When these are too large to be neglected, several courses of action are possible: (1) corrections can be made for differences between raters, (2) some of the essays can be rejudged, or (3) rater instructions can be discussed and improved.

Several approaches to the estimation of rater effects have been used. One approach is to implement a special estimation study in which a selection of essays is rated by all raters. Paul (1981) proposed Bayesian estimates of rater effects for such a crossed examinees × raters design. It is clear that such an analysis is only feasible with small examinee samples. A further disadvantage is that possible changes in rater standards between the estimation study and the final rating round cannot be detected.

In this paper the estimation of rater effects is discussed for a particular kind of rating design, namely one in which each examinee is rated by two independently judging raters and in which the above mentioned problems do not occur. Different rater pairs or teams should share raters in such a way that the teams cannot be divided into subgroups having no rater in common. In other words, each rater should be directly or indirectly linked or connected to each of the other raters. An example of such an overlapping design with four available raters and four teams is given in Figure 1.

In the general case there are n raters, numbered 1 through n, assigned in pairs to K different teams. Within each team one of the team members is arbitrarily given the first position. Now the rating design can be expressed in a pattern matrix $\mathbf{R}(2 \times K)$ with typical element r_{ik}, the number of the rater with position i in team k. For the design of Figure 1 this matrix could read

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

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 2, Spring 1984, pp. 213–218
© Copyright 1984 Applied Psychological Measurement Inc.
0146-6216/84/020213-06$1.55

Figure 1
An Overlapping Design with Four Teams and Four Raters

	Examinees			
Rater	Group 1	Group 2	Group 3	Group 4
1	xxx	xxx		
2		xxx	xxx	
3			xxx	xxx
4	xxx			xxx

The next two sections discuss two models for rater effects: the additive model and a simple nonlinear model.

The Additive Model

In the additive model the average difference between the ratings of the first and the second team member in team k, d_k, can be written as

$$d_k = \theta_{1(k)} - \theta_{2(k)} + f_k \quad , \qquad k = 1, \ldots, K \tag{1}$$

where $\theta_{i(k)}$ is the rater effect or leniency of rater r_{ik} ($i = 1,2$),
f_k is a residual with expectation zero and variance $N_k{}^{-1}\phi_k$,
N_k is the number of examinees assigned to team k, and
ϕ_k is the sum of the error variances associated with the individual raters in the team.

It is clear that the rater effects in Equation 1 are determined up to an additive constant. In order to eliminate the indeterminacy, Equation 1 can be rephrased in terms of relative rater leniences, that is the sum of the rater effects can be set equal to zero. This solution, which differs from the one proposed by Engelhard and Osberg (1983) for a similar estimation problem, makes it easy to obtain an expression for the variance-covariance matrix of estimated relative effects.

Due to the restriction on the rater effects, the last rater effect can be written as a function of the other effects, that is

$$\theta_n = -\sum_{j=1}^{n-1} \theta_j \quad . \tag{2}$$

Using this result, Equation 1 can be written in matrix terms as

$$\mathbf{d} = \mathbf{A}\boldsymbol{\theta} + \mathbf{f} \quad , \tag{3}$$

where $\boldsymbol{\theta}$ is a column vector with θ_1 through θ_{n-1},
$\mathbf{A}$ is a design matrix of order $K \times (n - 1)$, and
$\mathbf{f}$ is a column vector with K residuals.

Ordinary least squares estimates of the $n - 1$ θs are obtained from the minimization of

$$S = (\mathbf{d} - \mathbf{A}\boldsymbol{\theta})'\mathbf{N}(\mathbf{d} - \mathbf{A}\boldsymbol{\theta}) \quad , \tag{4}$$

where $\mathbf{N}$ is a diagonal matrix containing the group sizes N_j ($j = 1, \ldots, K$). The estimates of the first $n - 1$ parameters are given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{A}'\mathbf{N}\mathbf{A})^{-1}\mathbf{A}'\mathbf{d} \quad . \tag{5}$$

From Equation 5 an expression for the variance-covariance matrix of the $n - 1$ $\hat{\theta}$s can be obtained. Further, using Equation 2, $\hat{\theta}_n$ can be computed.

The additive model is computationally simple and straightforward, but unfortunately overly simplistic. A more general model is the linear model in which error and true score variances are allowed to vary between raters. De Gruijter (1984) described estimation in the linear model for multivariate normally distributed ratings and random assignment of examinees to rater teams. However, the linear model runs into problems at the lower and upper bound of the rating scale. Ultimately, only a nonlinear model can be satisfactory.

A Nonlinear Model For Rater Effects

With zero being the lowest and m the highest possible rating, plausible rating models are models in which the relationship between two rater scales is given by a function through the points with coordinates (0,0) and (m,m). A simple model satisfying this requirement is

$$\tau_{pj} = m \exp(\theta_p - b_j)/[1+\exp(\theta_p - b_j)] \quad , \qquad j = 1 , \ldots , n \tag{6}$$

where τ_{pj} is the expected or true score of examinee p for rater j,

θ_p is an examinee parameter, and

b_j is the effect of rater j, with $\Sigma b_j = 0$.

This model was proposed by Choppin (1982) for the analysis of rating data. It looks like the Rasch model, with the probability from this model replaced by an expected score, and for this reason the model was called an extension of the Rasch model. Clearly, the true scores for all raters are defined on a scale from zero to m.

From Equation 6 the relation between the true score for rater j and the true score after elimination of the rater effect can be obtained as

$$\tau_{pj} = \varepsilon_j \tau_p/[1 - m^{-1}\tau_p(1 - \varepsilon_j)] \quad , \tag{7}$$

or

$$\tau_p = \varepsilon_j^{-1} \tau_{pj}/[1 - m^{-1}\tau_{pj}(1 - \varepsilon_j^{-1})] \quad , \tag{8}$$

where τ_p is the true score after elimination of the rater effect and $\varepsilon_j = \exp(-b_j)$. When b_j is known, the scale transformation from Equation 8 can be used to correct scores for the effect of rater j.

Choppin presented an estimate of the difference between two rater effects, $h_k = b_{1(k)} - b_{2(k)}$, where $b_{i(k)}$ is the effect of rater i ($i = 1,2$) in team k. This estimate is

$$\hat{h}_k = \ln\{[\Sigma x_{p2(k)}(m - x_{p1(k)})]/[\Sigma x_{p1(k)}(m - x_{p2(k)})]\} \quad , \tag{9}$$

where the $x_{pi(k)}$ are the ratings given by the raters in team k. Estimates of $\hat{b}_j$ ($j = 1 , \ldots , n$) can be obtained from the K values $\hat{h}_k$, using the least squares procedure that was discussed in the previous section.

A second procedure to obtain the estimates of $\hat{b}_i$ is based on the fact that application of Equation 8 eliminates the rater effect. This equation can also be used with observed scores. The transformed or rescaled observed scores $x'_{p1(k)}$ and $x'_{p2(k)}$ with

$$x'_{pi(k)} = \varepsilon_{i(k)}^{-1} x_{pi(k)}/[1 - m^{-1}x_{pi(k)}(1 - \varepsilon_{i(k)}^{-1})] \qquad i = 1,2 \tag{10}$$

will generally lie close together. Parameter estimates $b_j = -\ln\varepsilon_j$ can be obtained for which the sum of squared differences,

$$F = \sum_k \sum_{p(k)} (x'_{p1(k)} - x'_{p2(k)})^2 \quad , \tag{11}$$

obtains its minimum. Estimates from the first procedure based on Equation 9 can be used as starting values in the iterative minimization of F.

An Example

For an analysis of rater effects, two ratings of an essay question on a scale from zero to ten were available for 949 examinees. Seventy-five examinees had both ratings equal to zero, most likely because they did not answer the question at all. These examinees were eliminated from the analysis in order to avoid underestimation of rater effects in the additive model.

Table 1 shows the rating design, the final group sizes, and the average ratings for both members of each rating team. The parameter estimates for the additive and the nonlinear model are given in Table 2. Finally, the variance-covariance matrix of the $\hat{\theta}$s, estimated under the assumption of equal ϕ_k, is given in Table 3.

From the θs in Table 2 and the estimated variance-covariance matrix (Table 3), it is clear that there are real differences between raters. The largest estimated difference is nearly one point on the 10-point rating scale.

Table 1
Number of Examinees Per Team (N), First Rater (r_1), Mean Rating for First Rater ($\bar{x}_1$), Second Rater (r_2) and Mean Rating for Second Rater ($\bar{x}_2$)

Group/Team	N	r_1	$\bar{x}_1$	r_2	$\bar{x}_2$
1	33	8	6.65	5	5.79
2	134	1	6.44	2	6.99
3	120	2	7.03	3	7.35
4	16	3	7.59	8	7.47
5	119	3	6.87	6	6.87
6	123	4	6.48	5	6.20
7	138	5	5.60	7	6.49
8	76	6	6.57	8	6.59
9	39	7	6.96	8	6.40
10	50	7	6.19	3	6.55
11	26	7	6.87	4	6.38

Table 2
Estimated Rater Effects

Rater	Rater Leniency θ (Additive Model)	$b(1)$ (Eq.9)	$b(2)$ (Eq.11)	θ*(τ=6.67)
1	-0.46	0.25	0.21	-0.49
2	0.09	-0.02	-0.01	0.01
3	0.41	-0.20	-0.17	0.36
4	-0.28	0.11	0.13	-0.29
5	-0.58	0.26	0.25	-0.59
6	0.31	-0.15	-0.16	0.34
7	0.32	-0.15	-0.17	0.36
8	0.18	-0.09	-0.09	0.19

Table 3
Estimated Variance-Covariance Matrix of the Effects in the Additive Model (× 1000) Under the Assumption of Equal Residual Variances ϕ_k

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
θ_1	16							
θ_2	8	10						
θ_3	1	2	5					
θ_4	-7	-6	-4	12				
θ_5	-7	-6	-3	6	8			
θ_6	-2	-1	2	-3	-3	7		
θ_7	-5	-4	-2	3	4	-2	6	
θ_8	-4	-3	-1	-1	0	2	0	7

The results from the nonlinear analysis are in agreement with those of the additive model. This agreement becomes more conspicuous with the examination of the impact of rater effects in the nonlinear model on true scores given a score level $\tau = 6.67$, a value close to the mean rating. In the last column of Table 2, shifts in true scores, θ^*, are given for this score level (with the estimates of b based on the minimization of F). These are close to the rater effects obtained from the additive model. For more extreme values of τ, the results for the two models must diverge.

The question of fit arises with respect to the nonlinear model. Fit was examined by plotting the transformed scores for both raters in a team against each other and then looking for whether the scores were spread around the diagonal line. Most of the plots looked satisfactory, but some plots suggested that the model was too simple to adequately represent the effects of all raters.

Discussion

Two simple models for the analysis of rater effects have been discussed within the context of a particular rating design. The additive model might be used when no accurate estimates of rater effects are needed, for example, when the results from an analysis are to be used only in a discussion on the impact and origin of differences between raters.

When more accurate results are needed, for example, for intended score corrections, the additive model is clearly inadequate. In such cases the more realistic nonlinear model might be used. Unfortunately not much is known about the properties of the estimates in this model. The jackknife (Mosteller & Tukey, 1968) and bootstrap methods (Efron, 1979) give possible solutions, while a first impression of the variance-covariance matrix of estimates could be formed on basis of the corresponding matrix in the additive model. Finally, it should be noted that more general nonlinear models can be defined in connection with the minimization criterion F.

References

Choppin, B. H. (1982). The use of latent trait models in the measurement of cognitive abilities and skills.

In D. Spearritt (Ed.), *The Improvement of Measurement in Education and Psychology*. Melbourne: Australian Council for Educational Research.

De Gruijter, D. N. M. (1984). The estimation of examiner effects in designs with overlapping examiner teams. *Kwantitatieve Methoden*, *13*, 148–155.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*, 1–26.

Engelhard, G., & Osberg, D. W. (1983). Constructing a test network with a Rasch measurement model. *Applied Psychological Measurement*, *7*, 283–294.

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *Handbook of Social Psychology*, (Vol. 2, 2nd Ed.). Reading MA: Addison-Wesley.

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

Acknowledgments

The author thanks the staff of political science from the University of Leyden for use of their data.

Author's Address

Send requests for reprints or further information to Dato N. M. de Gruijter, University of Leyden, Educational Research Center, Boerhaavelaan 2, 2334 EN Leyden, The Netherlands.