APPLICATION OF JUDGMENT ANALYSIS TO INTERRATER COMPARISONS

Linda E. Kapuniai

University of Hawaii

Joe H. Ward, Jr.

University of Texas at San Antonio

David H. Crowell

University of Hawaii

Michael J. Light and Rodney B. Boychuk

University of Hawaii

Joan E. Hodgman

USC Medical Center

ABSTRACT

A multiple regression method is presented for comparing the bases of two raters' judgments. This technique, which has been referred to as judgment analysis or policy capturing, is described for judgments of two nurses. In the example presented, judgments of future infant performance were derived from the nurse's scoring of infants' behavior on the Brazelton Neonatal Behavioral Assessment Scale. Brazelton dimension scores served as predictors of future performance in a test of differences between the policies (criteria) of the two nurse-raters. Sample data illustrate the technique but do not constitute a direct test of the data since the two nurse's ratings were actually on two different sets of infants. If the ratings had been on the same babies or identical samples of babies, the technique would have revealed, first, that the two nurses based their judgments primarily on one Brazelton dimension, interactive processes; and second, that one nurse consistently rated the babies' future performance at a higher level than did the other nurse. This technique has potential application to evaluation of rating criteria for training of observers or judges and in other problem solving areas such as conflict resolution.

Subjective predictions of progress and objective assessments of behavior are frequently required in many programs and projects. Consistency and accuracy of these observations are important issues in evaluating the judgments of different individuals or policies in relation to patterns of attributes (Fisher, 1983; Most & Starr, 1983). When assessing these judgments several questions often arise, for example, which of the many observations contributed the most to the overall judgment? Or, more importantly, if more than one observer is involved, to what extent did the raters rely on the same criteria as the basis for their predictions.

This paper presents a general statistical method for comparing the observations and determining the bases of the judgments of two individuals. The method is applied to observations on the current status and judgments of future capabilities of newborn infants. The observations were made by two nurses in the process of conducting the Brazelton Neonatal Behavioral Assessment Scale, BNBAS, (Brazelton, 1973). Judgment of the infant's future performance was made after completion of the BNBAS assessment. To illustrate the method, the nurse's judgments are treated as if they were rating the same infants. The two raters' judgments are then compared in terms of the regression weights associated with BNBAS dimension scores (Als, Tronick, Lester & Brazelton, 1977) derived from the original BNBAS observations. These scores represent the following dimensions: 1. Interactive processes: capacity to respond to social stimuli through orientation, cuddling and consoling; 2. Motoric processes: ability to maintain good tone, control motor behavior and integrate actions; 3. Organizational processes: ability to modulate states of consciousness in interactions with the environment primarily by shutting out aversive stimuli; and 4. Physiological reaction to stress: stability in response to stress. Dimensions 1-3 are scored as follows: 1 for good, 2 for average, and 3 for worrisome or deficient performance. Dimension 4 is coded either 1 (good) or 0 (bad).

The general model used as a basis for analysis of degree of rater similarity is a multiple linear regression using least-squares estimation of the regression weights,

$$Y=w_1U + w_2X_1 + \ldots + w_3X_2 + \ldots + w_kX_k + E, \qquad (1)$$

where w_1, w_2, and $w_3 \ldots w_k$ are least squares weights that minimize the squared errors in E. U is a vector of "1"s and X_1 and $X_2 \ldots X_k$ are the K predictor vectors. The dependent variable, Y, is the set of judgments or ratings of the situations characterized by the predictor data. The regression approach outlined here is a variation of a technique called "policy capturing," (Christal, 1968a, b; Christal & Bottenberg, 1969; Ward, 1979). The combination of the regression weights applied to each variable is taken as defining the rater's "policy" with regard to Y, the dependent variable.

The general hypotheses to be tested are: "Does the policy used by one rater differ from that used by another?" and "If the two policies differ, do they differ by a constant amount?"

The models for interrater comparison are presented first followed by their application to sample BNBAS data.

METHOD

Model Development

The following regression equations were designed to test the judgments of the two raters, Nurse 1 and Nurse 2, on the four BNBAS dimension scores. Each nurse's equation would take the general form,

$$Y_{\text{nurse}} = \text{function of (Dimension 1, Dimension 2, Dimension 3, and Dimension 4)} + E \qquad (2)$$

where Y is a nurse's judgment of an infant's performance. A similar regression equation is established for Nurse 2.

Model 1. Model 1, which incorporates both nurses' equations into a single model, takes into account the possibility that Nurse 1 makes ratings of infants that yield an equation (weights a_0, a_1, a_2, a_3, a_4) that differs from he corresponding equation (weights b_0, b_1, b_2, b_3, b_4) of urse 2. The equation is:

$$Y = a_0P1 + a_1(P1*D1) + a_2(P1*D2) + a_3(P1*D3) + a_4(P1*D4) + b_0P2 + b_1(P2*D1) + b_2(P2*D2) + b_3(P2*D3) + b_4(P2*D4) + E1 \quad (3)$$

where Y is the vector of future infant performance ratings from both nurses, D1 to D4 are the four BNBAS dimension scores, P1 is "1" for Nurse 1 and 0 otherwise, P2 is "1" for Nurse 2 and 0 otherwise, and E1 is the error in Model 1. In other words, the nurses are assumed to have based their predictions on two completely different policies. The least squares solution for Equation 3 will yield two sets of weights that might be different. Dimension 1 for Nurse 1 (P1*D1) has one weight (a_1) assigned to it, dimension 1 for Nurse 2 (P2*D1) may have another weight (b_1) assigned to it, and so on. Furthermore P1 is assigned one weight (a_0) and P2 may have another weight (b_0).

Model 2. To test the hypothesis that the two nurses' predictions differed by a constant, restrictions are imposed on Model 1 to obtain Model 2, Equation 4. To illustrate this point, we would act as if the hypothesis is: when two nurses are presented with 10 babies and asked to make predictions independently on those 10 babies, the predictions will differ by a constant amount. The restrictions implied by the hypotheses of constant differences are:

$$a_1 = b_1 = c_1,\ a_2 = b_2 = c_2,\ a_3 = b_3 = c_3, \text{ and } a_4 = b_4 = c_4$$

Substituting these restrictions in Model 1 gives Model 2.

$$Y = a_0 P1 + b_0 P2 + c_1 D1 + c_2 D2 + c_3 D3 + c_4 D4 + E2. \qquad (4)$$

Observe that this model has the same weights (c_1, c_2, c_3, c_4) for the two nurses, but that the nurses' judgments will differ by the constant value $a_0 - b_0$.

Model 3. Model 3 assumes that the policies used by Nurses 1 and 2 are identical. The restriction on Model 2 implied by this hypothesis is $a_0 = b_0 = c_0$. Substituting this restriction in Model 2 gives Model 3,

$$Y = c_0 U + c_1 D1 + c_2 D2 + c_3 D3 + c_4 D4 + E3, \qquad (5)$$

where $U = P1 + P2$, the Unit Vector containing a "1" in every element. Observe that this model has given up all information that distinguishes the two nurses.

Testing the Hypotheses.

After Models 1, 2 and 3 (equations 3, 4, and 5) have been developed, the questions of policy differences can be answered by comparing the R^2's (squared multiple correlations) from the equations. The question, "If the two policies differ, do they differ by a constant amount?" can be answered by determining if R_1^2 is significantly larger than R_2^2. This comparison, Equation 6, is made by calculating

$$F = \frac{(R_1^2 - R_2^2) / (n_1 - n_2)}{(1 - R_1^2) / (N - n_1)} \qquad (6)$$

which is distributed as _F_ with degrees of freedom $(df_1) = (n_1$

$- n_2)$ and $(df_2) = (N - n_1)$; n_1 (=10) is the number of coefficients in Model 1, n_2 (=6) is the number of coefficients in Model 2 and N (=45) is the total number of ratings by both nurses. If the F-test is not significant we accept the restricted Model 2, that is the hypothesis that the differences between the two nurse's policies are constant is not rejected. The difference will be $(a_0 - b_0)$. In this case Model 2 would be adopted.

The next step in the analysis depends on the result of the comparison between Model 1 and Model 2. If we reject the constant difference hypothesis we conclude that the policies differ, and, therefore, Model 1 is appropriate.

If we accept the constant difference hypothesis Model 2 is assumed, and to test that the policies are identical we compare Model 2 with Model 3 as in equation (7),

$$F = \frac{(R_2^2 - R_3^2) \;/\; (n_2 - n_3)}{(1 - R_2^2) \;/\; (N - n_2)} \qquad (7)$$

where R_2^2 is compared to R_3^2. If R_2^2 is significantly larger than R^2, the null hypothesis $(a_0 = b_0 = c_0)$ is rejected and it can be concluded that the nurses differ in their ratings and the difference is constant. If the difference in the two R^2 is not significant, it is concluded that there are no differences between the nurses' judgments when expressed in terms of the four BNDAS dimensions.

Model Application.

Subjects and Procedure. Subjects were 45 infants who were seen at term as part of a larger study of metabolic derangements, neurophysiological functioning and behavior. Informed consent was obtained from parents and physicians prior to testing. Brazelton assessments for 25 of these infants were conducted by one rater, Nurse 1, and the remaining 20 by a second rater, Nurse 2. The same assistant recorded the scores during the BNBAS tests done by both nurses. After each test was completed, the test information was combined to form the four dimension scores (Als et al., 1977). Subsequent to the determination of the four dimension scores the nurses made a Judged Future Performance, JFP, for each infant. This JFP was scored as 0, 1, or 2, to correspond with predictions of below average, average, or above average future performance. No other explicit criteria were suggested.

Results The scores for the four dimensions resulting from the test of the two nurses are in Table 1.

Table 1. Judged Future Performance (JFP), and Brazelton Neonatal Assessment Scale Dimension Scores From Two Nurses

Nurse 1						Nurse 2					
Case	JFP	D1	D2	D3	D4	Case	JFP	D1	D2	D3	D4
1	2	1	2	1	1	26	2	1	2	1	1
2	1	2	1	2	1	27	1	2	1	2	1
3	1	1	2	2	1	28	1	2	2	2	1
4	1	2	2	2	1	29	1	3	2	2	1
5	1	2	2	3	1	30	2	1	2	2	1
6	1	1	1	2	1	31	1	2	2	3	0
7	2	1	2	1	1	32	1	3	2	2	1
8	1	2	2	2	1	33	1	3	2	2	1
9	1	2	2	3	1	34	1	3	2	2	1
10	1	2	1	2	1	35	1	2	2	2	1
11	1	1	2	2	0	36	2	1	1	2	1
12	1	2	1	2	1	37	1	2	2	2	1
13	0	3	1	2	0	38	2	2	1	2	1
14	1	1	3	3	1	39	1	3	2	2	1
15	1	3	2	2	0	40	1	3	2	2	1
16	1	3	2	2	1	41	1	2	2	2	1
17	1	3	1	2	1	42	1	1	2	2	1
18	1	2	1	2	1	43	2	1	1	2	1
19	0	3	3	2	1	44	2	2	2	2	1
20	2	1	2	1	1	45	2	2	2	2	1
21	1	2	2	1	1						
22	1	3	2	3	1						
23	1	3	2	2	1						
24	2	1	2	1	1						
25	1	1	1	1	1						

The four dimension scores, nurse identification and ratings of future performance were then entered into the models previously described. The results were $R_1^2 = 0.931$; $R_2^2 = 0.926$; and $R_3^2 = 0.912$. The R^2 values were entered into the $\underline{F}$ -test formulas with the appropriate degrees of freedom. First, Model 1 was compared with Model 2 using Equation 6.

$$\text{Test 1: } \underline{F}_{(4,35)} = \frac{(0.931-0.926) \;/\; (10-6)}{(1 - 0.931) \;/\; (45-10)} = .597 \qquad (8)$$

Test 1 (Model 1 compared with Model 2) was not significant. In light of this result, Model 2 was assumed where $a_1 = b_1 = c_1$, $a_2 = b_2 = c_2$, $a_3 = b_3 = c_3$ and $a_4 = b_4 = c_4$. Since Test 1 indicated that nurses' judgments differed by a constant amount, Model 2 was compared to Model 3 in Test 2, equation (9), using equation (7) above.

$$\text{Test 2: } \underline{F}_{(1,39)} = \frac{(0.926-0.912) \;/\; (6-5)}{(1 - 0.926) \;/\; (45-6)} = 7.26 \qquad (9)$$

The $\underline{F}$ of 7.26 was significant at $\underline{p} < .05$; therefore, the null hypothesis, that $a_0 = b_0 = c_0$, was rejected. While the expected nurses' ratings of future performance differed by a constant amount, the constant difference was not zero. The estimate of the actual difference was $a_0 - b_0 = 1.93 - 2.24 = -.31$ (see Table 2).

Table 2. Regression Results for Model 2 (Equation 4)

Predictor	Coefficient	F-Value* (df=1,39)	Prob- ability
P1-Nurse One	1.93		
P2-Nurse Two	2.24		
D1-Interactive Processes	-.32	16.75	.0002
D2-Motoric Processes	-.02	.04	.8471
D3-Organizational Processes	-.22	3.45	.0708
D4-Physiological Reaction to Stress	.26	1.63	.2098

*F-Values result from the (1,39 degrees of freedom) test that the corresponding coefficient is equal to zero

Since the differences between ratings were constant (Test 1), we can conclude that the relationships between the four BNBAS scores and the judgments of Nurse 1 did not differ from the relationships of Nurse 2. But Test 2 indicated that even though differences were constant there was a significant difference between the level of ratings of the two nurses. Nurse 2 tended to give higher ratings (.31) than Nurse 1.

Since the nurses did not actually rate the same infants it cannot be determined whether these results reflect actual differences in the nurse's policies or differences in the two sets of infants. In this example the relationship of the four BNBAS scores for the judgments was the same for the two nurses; therefore, it was of interest to examine each of the four coefficients c_1, c_2, c_3, c_4. Inspection of the Model 2 regression equation in Table 2 reveals that the two nurses based their judgments primarily on dimension 1 (Interactive Processes). This conclusion is based on the small probability (p = .0002) associated with the hypothesis that babies who have the same scores on Dimension 2, 3, and 4, but different Dimension 1 ratings will have the same expected JFP ratings. The probability of .07 associated with the test on Dimension 3 indicates that Organizational Processes also may contribute to the judgment process.

DISCUSSION

Evaluation of judgments of behavior based on observations is a situation that occurs frequently. It is important not only to know on what bases and how consistently the observer is making judgments, but also whether judgments of different observers or raters have similar bases. Techniques which address these questions are demonstrated in Test 1 and Test 2, multiple regression models which have been described as policy capturing. This approach describes the set of variables or observations that best characterize a judgment.

One possible application of judgment analysis or policy capturing would be training programs where the goals are to evaluate and increase degree of intra- and inter-rater reliability. If the policy or combination of independent variables (observations), does not account for a significant proportion of the variance in the dependent variable, it can be inferred that the judgment of the observer is, to a large degree based on information other than that contained in the predetermined set of observations. In other words, the person is utilizing information not summarized in the behaviors represented by the values of the independent variables in the equation. For example, if the observer is instructed to make an assessment of an infant's future performance based on the results of the BNBAS, and the BNBAS values do not support or predict the JFP, it may be that knowledge of the child's home environment or some other unknown factor was entering into this judgment. In this situation, it may be necessary to retrain the observer to eliminate other than specified information or it may be more desirable to reconsider the factors in the equation. If two raters (judges or observers) differ in their rating criteria, the criteria of the rater whose judgments best approximate actual future performance can be adopted as the standard for others. These same considerations could be pertinent to questions of conflict resolution, both in refining the dependent variable (Nost & Starr, 1983) and as a way of describing how decisions are arrived at in problem-solving or negotiation settings (Fisher, 1983).

CONCLUSION

This technique can be a valuable aid for detecting implicit weightings of unknown variables which result in unexplained variance in judgments, and for standardizing judgments, that is, insuring that they are based on the same criteria.

REFERENCES

Als, H., Tronick, E., Lester, B.M., and Brazelton, T.B. The Brazelton Neonatal Behavioral Assessment Scale (BNBAS), *Journal of Abnormal Child Psychology*, 1977, *5*, 215-230.

Brazelton, T.B. *Neonatal Behavioral Assessment Scale*, Clinics in Developmental Medicine, No. 50, London: S.I.M.P. with Heinemann Medical, Philadelphia. Lippincott, 1973.

Christal, R.E. JAN: A technique for analyzing group judgment, *The Journal of Experimental Education*, 1968a, *36*, 24-27.

Christal, R.E. Selecting a Harem-and other applications of the policy capturing model, *The Journal of Experimental Education*, 1968b, *36*, 35-41.

Christal, R.E., and Bottenberg, R.A. Grouping criteria - A method which retains maximum predictive efficiency, *The Journal of Experimental Education*, 1968, *36*, 28-34.

Fisher, R.J. Third party consultation as a method of intergroup conflict resolution, *Journal of Conflict Resolution*, 1983, *27*, 301-334.

Most, B.A., and Starr, H. Conceptualizing "war" consequences for theory and research, *Journal of Conflict Resolution*, 1983, *27*, 137-159.

Ward, J.H., Jr. Creating mathematical models of judgment processes: from policy capturing to policy specifying, *The Journal of Experimental Education*, 1979, *48*, 60-84.