DOCUMENT RESUME

ED 262 079 TM 850 560

AUTHOR Cason, Gerald J.; Cason, Carolyn L.
TITLE A Regression Solution to Cason and Cason's Model of Clinical Performance Rating: Easier, Cheaper, Faster.
PUB DATE 85
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Clinical Experience; Computers; Efficiency; *Estimation (Mathematics); Goodness of Fit; Higher Education; *Interrater Reliability; Mathematical Models; *Medical Students; *Regression (Statistics); Statistical Studies; *Student Evaluation; Validity

ABSTRACT

A more familiar and efficient method for estimating the parameters of Cason and Cason's model was examined. Using a two-step analysis based on linear regression, rather than the direct search interative procedure, gave about equally good results while providing a 33 to 1 computer processing time advantage, across 14 cohorts of junior medical students rated by residents and attending physicians in a medicine clerkship. First, regression analysis of z-transformed ratings produced regression weights (b's) associated with each rater and subject. Then, the unit vector and person b's were converted into inter-person distances, then person locations on the underlying stringency and ability scale. Essentially equally good fit with the data was achieved by the new method in 12 of the 14 data sets. In the other two, fit was still quite good. Correlation between parameter values estimated by the two methods was very high in groups where equal fit was achieved. In the other two, moderately high correlations were observed. The new method provided equally good improvement in reliability and convergent validity of adjusted scores. The improved economy and ease of application of the new method further expanded the advantages of using the Casons' model to statistically control rater bias. (Author/GDC)

A Regression Solution to Cason and Cason's Model of Clinical Performance Rating: Easier, Cheaper, Faster

Gerald J. Cason and Carolyn L. Cason
University of Arkansas for Medical Sciences

Using a two step analysis based on linear regression, rather than the less familiar direct search iterative procedure used in previous research, gave about equally good results while providing a 33:1 computer processing time advantage to the new method across 14 cohorts of Junior medical students rated by residents and attending physicians in a Medicine Clerkship. In the first step, regression analysis of z-transformed ratings produced regression weights (b's) associated with each rater and subject. In the second step, the unit vector and person b's were converted into theoretical inter-person distances, then person locations (RRPs and SAPs) on the underlying stringency and ability scale. Essentially equally good fit with the data was achieved by the new, faster method in 12 of the 14 data sets ($.82<R<.95$; $p<.001$). Even in the 2 cohorts in which the new faster method did not find quite as good fitting parameter values, the fit was still quite good ($.79<R<.80$; $p<.005$). Correlation between parameter values estimated by the two methods was very high in groups where equal fit with the data was achieved($.88<r<1.00$). Even in the other two, moderately high correlations were observed between parameter values estimated by the two approaches ($.84<r<.91$). In spite of some differences in fit with data, the new method provided equally good improvement in reliablity and convergent validity of adjusted scores. The increased economy and ease of application of the new method further expanded the advantages of using the Casons' model to statistically control rater bias as either an adjunct to or in the absence of adequate direct control methods such as rater training to improve the reliability and validity of clinical performance measures.

Presented at the Annual Meeting of the American Educational Research Association, Chicago, April, 1985.

Reprints available from:

Gerald J. Cason, Ph.D.
Office of Educational Development - 595
University of Arkansas for Medical Sciences
4301 West Markham
Little Rock, Arkansas 72205

A Regression Solution to Cason and Cason's Model of Clinical Performance Rating: Easier, Cheaper, Faster

Gerald J. Cason and Carolyn L. Cason
University of Arkansas for Medical Sciences

This paper presents a more familiar and economical method, based on linear regression, for estimating the parameters of Cason and Cason's (1984) model. This improved parameter estimation approach will promote research on the Casons' theory of performance rating and ease its application to the problem of separating rater bias from true performance in practical performance assessment.

Theory

Cason and Cason's simplified model of their performance rating theory accounts for all systematic variation in performance ratings exclusively by variation in rater stringency and subject (e.g., student) ability. Cason and Cason (1984; G. Cason et al, 1983; C. Cason et al, 1983) have presented evidence that the model fits clinical performance data in a common type of health profession education setting, i.e., where each student is rated by several but not all raters and each rater rates several but not all students. Where sufficient overlap in who rated whom occurred, variation in rater stringency could be statistically controlled. This led to improved reliability and validity of the adjusted performance ratings. These results were obtained at two different schools, with different rating inventories, different amounts of rater training, and at different levels of trait specificity. However, all prior research was based on estimating rater stringencies and subject abilities using program MERLIN in conjunction with program STEPIT (Chandler, 1965). MERLIN used STEPIT to find the best values for subject ability points (SAPs) and rater reference points (RRPs) in the sense of producing the least-squares fit with the observed data. STEPIT finds local minima of continuous real functions by cyclic relaxation and parabolic interpolation (a variation of direct search). STEPIT was developed for quantum chemistry research and was both relatively slow in this application (therefore expensive) and unfamiliar to most educational researchers. It did have the initial advantage of being easy to apply to the Casons' model.

In the Casons' model the expected subject rating (ESR), measured as a percent of the maximum rating, is a function of the difference, z, between the rater's stringency (i.e., value associated with the Rater Reference Point or RRP) and the subject's ability (i.e., value associated with the Subject Ability Point or SAP). In previous research this relationship was modified by an arbitrary scaling factor (SF=100).

$$z = (SAP - RRP)/SF$$

The theoretically postulated curvilinear relationship between z and the expected subject rating (ESR) has been stipulated as the unit-normal ogive. Thus, the ESR (in percent) for a given z is equal to the proportion of area under the normal curve below z; that is, p(z) times 100:

Presented at the Annual Meeting of the American Educational Research Association, Chicago, April 1985.

Reprints: G.J. Cason, UAMS-OED-595, 4301 W. Markham, Little Rock, Arkansas 72205. Phone (501) 661-5720.

$$ESR = p(z) \times 100$$

Method

Estimation of the model parameters (RRPs and SAPs) was accomplished in two steps. First, the observed ratings were transformed to z's using an inverse normal probability function. These z's were used as the criterion values (Y vector) in a regression model of the form:

$$Y = cU + b_1R^1 + b_2R^2 + \ldots + b_nR^n + b_{n+1}S^{n+1} + b_{n+2}S^{n+2} + \ldots + b_{n+k}S^{n+k} + E$$

where:

U is a "unit vector" containing a 1 for each observation in Y;

R^i (i=1 through n) is a vector containing a 1 if the observation in Y pertains to a rating given by rater i, zero otherwise;

S^j (j=n+1 through n+k) is a vector containing a 1 if the observation in Y is associated with subject j, zero otherwise; and,

c and b_1 through b_{n+k} are the regression weights that minimize the squares of the values in the error vector (E).

A special-purpose program, GENVEC, was used to generate the above model from input which specified ID numbers for rater-subject pairs and associated observed ratings. Program LMS (linear model solver), based on Ward and Jennings' (1973) program MODEL, provided a regression analysis of the model generated by GENVEC. The regression analysis carried out by LMS provided the regression weights (b's not Beta's) for each vector in the model produced by GENVEC. Note however that the regression program must allow for models with redundant predictor vectors (e.g., program MODEL in Ward & Jennings, 1973). The z's in the criterion vector correspond to observed distances (containing error) between raters and subjects on the underlying stringency-ability scale. Pairs of b's and the unit vector weight give the theoretical, "error-free" distance between a rater-subject pair:

$$RXTOS(I) = BOFS(I) - BOFRX + CONST$$

where:

RXTOS(I) is the distance from a rater (RX) to subject I;

BOFS(I) is the regression weight (b) of subject I;

BOFRX is the regression weight (b) of an arbitrarily chosen rater (RX); and,

CONST is the regression weight (c) for the unit vector (U).

The second step was to convert the regression weights into theoretical distances and then into rater and subject locations on the stringency and ability scale. One rater's RRP (i.e., rater with the most ratings and lowest ID number) was arbitrarily chosen as the anchor location for the scale and this point was assigned an arbitrary value (i.e., 500). Once one rater's position (i.e., stringency) was defined, all subjects could be located with respect to that rater by the linear equation:

$$SLOC(I) = ANCVAL + RXTOS(I)$$

where:

SLOC(I) is the location (SAP) of subject(I); and,

ANCVAL is the arbitrary value used for anchoring one rater's RRP.

As soon as all subjects were located, an analogous set of equations could be directly solved to obtain the remaining rater locations (RRPs). Program LOCATE was used to solve these equations and obtain the theoretical, "error-free" distances, and then the RRPs and SAPs.

Jointly GENVEC, LMS, and LOCATE replaced STEPIT as the method of finding the RRPs and SAPs. All of these programs were run on a DEC-10 System computer. For simplicity in reporting, MERLIN-STEPIT refers to the earlier method and MERLIN-LMS refers to the current regression based method of finding RRPs and SAPs. The new method, MERLIN-LMS, described above was applied to 14 data sets previously analyzed by MERLIN-STEPIT (G. Cason et al, 1983). The data were overall ratings of the clinical performance of Junior year medical students in a Medicine Clerkship (i.e., clinically oriented course) during 1981 and 1982 at a school located in the Southwest U.S.A. See G. Cason, Cason, & Littlefield (1983) for a more complete description of the data source. One rater's data (ID 3027) which was included in the earlier analyses was excluded here because the ratings were out of valid range. Ommission of this rater removed two ratings from data set 1981F.

Results

As can be seen from an examination of Table 1, in 12 out of the 14 data sets, MERLIN-LMS achieved essentially as good fit with the data as did MERLIN-STEPIT. In 2 cohorts, 1982 E and F, MERLIN-LMS did not achieve as good a fit. However even in these cases, MERLIN-LMS's fit with the data was still quite good ($R>.79; p<.005$). The R values reported in Table 1 for MERLIN-STEPIT's fit are not the same as those provided in the earlier report (G. Cason, Cason, & Littlefield, 1983; Table 1). Prior to January 1984, MERLIN contained a program coding error which substituted MSQ for SS in computing the RSQ. This produced spuriously low reported values.

As can be seen from Table 1, in each cohort MERLIN-LMS solved the equations for the parameters in only a small fraction of the computer time required by MERLIN-STEPIT. In every case, MERLIN-LMS solved in under 2 minutes. In no case did MERLIN-STEPIT solve in less than 10 minutes. In total, across all 14 data sets, the original approach required 528 minutes while MERLIN-LMS required only 15. Therefore, MERLIN-LMS had a machine processing time advantage of approximately 33:1 while achieving good to excellent fit with the data.

Table 2 provides standard deviations for the parameters of Cason and Cason's model found by MERLIN-LMS and MERLIN-STEPIT on each of the 14 data sets. In 10 of the 14 data sets, variation in rater standards (i.e., RRPs) was greater than variation in student ability (SAPs). In general, MERLIN-LMS produced estimates of the parameters with slightly higher variability than did MERLIN-STEPIT. A notable exception to this pattern was cohort 1981F. The largest variability of parameters, regardless of solution approach used, was found in this cohort: standard deviations twice those found in other cohorts. There was nothing obvious about the pattern of variability of parameters that provided any basis for speculating on why the fit of MERLIN-LMS's parameters were not quite as good in 1982 E and F as were MERLIN-STEPIT's.

Table 1. Fit of Cason and Cason's Model and Time to Solve for each Cohort's Data Using MERLIN-LMS and MERLIN-STEPIT

Year: 1981

	Cohort						
	A	B	C	D	E	F[4]	G
Multiple R[1]							
MERLIN-LMS	.93	.89	.90	.88	.87	.95	.90
MERLIN-STEPIT[2]	.93	.90	.90	.89	.88	.96	.90
Minutes to Solve							
MERLIN-LMS	.75	.59	1.18	1.21	.84	.95	1.07
MERLIN-STEPIT[3]	20.00	17.00	49.00	38.00	10.00	35.00	23.00
Number of raters	37	28	46	38	29	29	31
Number of students	21	20	29	24	17	18	20
Total ratings	108	76	126	102	73	72	91

Year: 1982

	Cohort						
	A	B	C	D	E	F	G
Multiple R[1]							
MERLIN-LMS	.85	.88	.89	.90	.79	.80	.82
MERLIN-STEPIT[2]	.86	.89	.90	.91	.89	.89	.87
Minutes to solve							
MERLIN-LMS	1.36	1.93	.77	.95	1.20	1.12	1.12
MERLIN-STEPIT[3]	53.00	68.00	28.00	53.00	49.00	49.00	36.00
Number of raters	42	42	34	39	43	39	40
Number of students	24	24	16	21	27	25	27
Total Ratings	129	129	80	114	145	131	133

[1] All R's significant at $p < .005$.
[2] Values in earlier report (G. Cason, Cason, & Littlefield, 1983; Table 1) were spuriously low due to coding error in MERLIN which substituted MSQ for SS in computing RSQ.
[3] Times for previous analyses available only to nearest minute.
[4] Both ratings of rater 3027 included in previous analyses, excluded in the present study because they were out of valid range.

Table 2. Standard Deviations of Parameters as Found by MERLIN-LMS and MERLIN-STEPIT[1]

Year: 1981

	Cohort A	B	C	D	E	F	G
Parameter							
RRP							
MERLIN-LMS	34.48	36.66	37.60	32.43	28.83	63.69	29.20
MERLIN-STEPIT	34.02	33.68	33.37	30.83	28.29	92.28	28.25
SAP							
MERLIN-LMS	28.85	41.75	28.55	23.10	21.09	36.17	23.33
MERLIN-STEPIT	28.93	38.15	26.72	21.28	21.45	49.74	22.93
All							
MERLIN-LMS	36.60	43.21	41.33	34.44	30.67	65.43	38.57
MERLIN-STEPIT	36.24	40.01	38.14	32.82	30.49	86.86	37.78

Year: 1982

	Cohort A	B	C	D	E	F	G
RRP							
MERLIN-LMS	44.03	34.95	29.27	39.34	41.72	50.59	39.15
MERLIN-STEPIT	39.41	33.42	27.89	38.04	31.53	31.32	24.37
SAP							
MERLIN-LMS	39.96	35.44	38.34	27.88	28.63	52.69	40.96
MERLIN-STEPIT	33.97	33.49	36.99	26.06	20.01	37.33	26.91
All							
MERLIN-LMS	47.08	40.12	38.39	41.17	41.80	62.54	44.46
MERLIN-STEPIT	42.27	38.48	37.09	39.66	33.06	47.24	31.19

[1]Values for MERLIN-STEPIT obtained from analyses in previous study but not included in original report of that study (Cason, Cason, & Littlefield, 1983).

Table 3. Correlation between Parameters as Found by MERLIN-LMS and MERLIN-STEPIT

Year: 1981

Parameter	A	B	C	D	E	F	G
				Cohort			
RRP	1.00	.99	.99	.98	.97	.99	1.00
SAP	1.00	.99	.97	.96	.96	.99	.99
ALL	1.00	.99	.99	.98	.98	.99	1.00

Year: 1982

	A	B	C	D	E	F	G
				Cohort			
RRP	.97	.99	.98	.98	.87	.84	.90
SAP	.96	.99	.97	.97	.91	.87	.88
ALL	.97	.99	.99	.99	.90	.90	.91

Table 4. Single Rater Reliabilities and Validities of Observed and Adjusted Ratings

Year: 1981

Reliability; Validity	A	B	C	D	E	F	G
				Cohort			
Observed Ratings	.30	.31	.24	.22	.22	.16	.23
Adjusted Ratings							
MERLIN-LMS	.87	.61	.80	.87	.84	1.00	.98
MERLIN-STEPIT[1]	.87	.61	.80	.87	.84	1.00	.96
Ratings per student[2]	5.13	3.79	4.33	4.24	4.28	3.97	4.53

Year: 1982

Reliability; Validity	A	B	C	D	E	F	G
				Cohort			
Observed Ratings	.08	.24	.40	.18	.10	.35	.33
Adjusted Ratings							
MERLIN-LMS	.37	.65	.76	.89	.58	.76	.88
MERLIN-STEPIT[1]	.37	.65	.76	.89	.58	.76	.88
Ratings per student[2]	5.37	5.37	4.99	5.42	5.37	5.23	4.92

[1]Values reported previously (Cason, Cason, & Littlefield, 1983; Table 3).
[2]The geometric mean number of ratings per student (k) is used in the Spearman-Brown expansion formula to determine the reliability of the average of k independent ratings.

Table 3 provides correlations between the parameters estimated by the two methods for each of the 14 cohorts. For RRPs, these ranged from .84 to 1.00 with a mean of .96; for SAPs from .87 to 1.00 with a mean of .95; and, for all parameters considered together .90 to 1.00 with a mean of .97. In most of the cohorts the parameter estimates were essentially the same. As expected in the two cohorts in which the Cason and Cason model fit the data least well, 1982 E and F, the correspondence between parameters estimated by the two methods was lowest. However, even in these two cases the correlations between the parameter estimates were moderately high.

Single rater reliabilities and validities for observed ratings and adjusted ratings for MERLIN-LMS and MERLIN-STEPIT are given in Table 4. The two procedures (LMS and STEPIT) resulted in the same estimated reliabilities, even in those cohorts where the fit achieved by LMS was somewhat less than that of STEPIT. These reliabilites are intra-class correlations computed by a variation of the analysis of variance procedure recommended by Ebel (1951). Details of our procedure were provided in a previous report (Cason, Cason, & Littlefield, 1983). Single rater reliabilities may be understood as either the reliability associated with one rater's rating of a student; or, the average inter-rater correlation expected from randomly chosen pairs of raters. Because students were rated by multiple raters, the overall reliability of ratings in a group is estimated by using the geometric mean number of ratings per student (k) in the Spearman-Brown expansion formula. As we have discussed elsewhere (Cason & Cason, 1984), the single rater reliabilities may be equally well interpreted as convergent validity indexes. As the conventional validity expansion formula shows, k's effect on improving validity is much less pronounced than on improving reliability.

Importance

Unlike separately z-transforming each rater's ratings (Ebel, 1951) or handicapping as recommended by Littlefield et al (1984), neither of which includes a test of the assumptions used to justify these methods of removing rater bias from ratings, the Casons' model provides direct means for testing its general assumptions (i.e., fit with data) and contrasting it with the most common alternative model. The regression based method of solving for the parameters of the Casons' model make it more accessible and economical to conduct research on their theory and to apply the technology in practical settings to achieve more nearly reliable and valid performance measures. The increased economy of this method over the earlier one further expands the cost advantage of statistical control of rater bias when compared with direct control methods such as rater training. The greater familiarity of regression methods makes this approach easier to understand and use by a majority of educational researchers. The greater speed, ease of use, and thus economy of this approach are achieved at practically no cost in accuracy of solution. The regression approach provides adjusted scores whose reliability and validity are improved to the same (large) degree that the earlier, more cumbersome approach attained.

References

Cason, C.L., Cason, G.J., & Littlefield, J.H. (1983). Variation of intra-rater stringency in cognitive-technical and affective-interpersonal clinical performance domains. Presented at AERA, Montreal.

Cason, G.J., & Cason, C.L. (1984). A deterministic theory of clinical performance rating: Promising early results. Evaluation & the Health Professions, 7(2), 221-247.

Cason, G.J., Cason, C.L., & Littlefield, J.H. (1983). Controlling rater stringency error in clinical performance rating: Further validation of a performance rating theory. Presented at AERA, Montreal. Resources in Education, 18(8), 176. (ERIC ED-228-314)

Chandler, J.P. (1965). STEPIT: A FORTRAN II subroutine for finding local minima of real functions. Bloomington: Quantum Chemistry Program Exchange, Indiana University.

Ebel, R.E. (1951). Estimation of reliability of ratings. Psychometrika, 16: 407-424.

Littlefield, J.H., Ellis, R., Cohen, P., & Herbert, R. (1984). Leniency and score distribution differences among clinical raters. In Research in Medical Education 1984: Proceedings of the 23rd Annual Conference. Washington, D.C.: Association of American Medical Colleges, 199-204.

Ward, J., & Jennings, E. (1973). Introduction to linear models. Englewood Cliffs, NJ: Prentice-Hall.