DOCUMENT RESUME

ED 287 887 TM 870 648

AUTHOR Cason, Gerald J.; Cason, Carolyn L.
TITLE Practical and Theoretical Requirements for Controlling Rater Stringency in Peer Review.
PUB DATE Jul 87
NOTE 15p.; Paper presented at the Sigma Theta Tau International Research Congress (Edinburgh, Scotland, July, 1987). Some tables contain small print. For related document, see TM 870 649.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Abstracts; *Computer Oriented Programs; *Conference Papers; Data Processing; Evaluation Criteria; *Evaluation Methods; *Interrater Reliability; Measurement Techniques; Models; *Peer Evaluation; Selection; Standards
IDENTIFIERS *Performance Rating Theory (Cason and Cason); Sigma Theta Tau International Research Congress

ABSTRACT

This study describes a computer based, performance rating information processing system, performance rating theory, and programs for the application of the theory to obtain ratings free from the effects of reviewer stringency in reviewing abstracts of conference papers. Originally, the Performance Rating (PR) System was used to evaluate the clinical performance of nursing students, medical students, and residents, as well as faculty teaching performance. It was also used for processing reviews of proposals for Sigma Theta Tau's International Research Congress in 1987. In general, the capabilities of the PR System which make it useful for processing the peer review of abstracts include: (1) the ease and speed with which a new inventory may be implemented; (2) the low cost and speed of data collection; (3) processing and reporting arising from the use of optically scanned rating sheets; and (4) the appropriateness of its reports. Improvements in the peer review process can be obtained through the use of specialized data management and analysis systems. As these systems become more generally available, there may be a concomitant improvement in the reliability and validity of the formal, technical peer review process. (KSA)

Practical and Theoretical Requirements for Controlling Rater Stringency in Peer Review*

Gerald J. Cason, Ph.D. and Carolyn L. Cason, Ph.D

University of Arkansas for Medical Sciences
Little Rock, Arkansas, USA
72205

Address correspondence to:

Gerald J. Cason, Ph.D.
UAMS-OED-595
4301 West Markham Street
Little Rock, AR 72205
USA

Telephone: (501) 661-5720

* Presented at the International Research Congress held in Edinburgh, Scotland, July 1987. The Congress was sponsored by Sigma Theta Tau, International Honor Society of Nursing, Indianapolis, Indiana, USA; the Department of Nursing Studies, University of Edinburgh, Scotland, UK; and, the Royal College of Nursing, London, England, UK.

Members of the Department of Computer Services, UAMS, especially Mark Boughter, Tom Hart and Tom Lewis, were very helpful during this project.

Practical and Theoretical Requirements for Controlling Rater Stringency in Peer Review

Abstract

Peer review is a process usually conducted under conditions of extreme scarcity of resources: very little money beyond minimal postage costs is available; the reviews are done by volunteers with little time to spare; little clerical or technical support staff is available; and, for professional meetings, the schedule requires that the process be completed in little time. These constraints usually prevent the practical application of reviewer training and the use of manual data entry and general purpose data-base or statistical programs to detect and off-set the effects of differences in the stringency (i.e., standards) of reviewers. This paper describes a computer based, performance rating information processing system, performance rating theory and programs for the application of the theory to obtain ratings free from the effects of reviewer stringency. In spite of the otherwise usual lack of resources, the prior existence of these systems, originally developed for the assessment of the clinical performance of students in health professions educational programs, provided the practical capability for controlling reviewer stringency in the peer review process for an international research conference. (Results of this application are given in a separate paper.) Improvements in the peer review process can be obtained through the use of appropriate specialized data management and analysis systems. As systems similar to those described here become more generally available, we may expect a concomitant improvement in the reliability and validity of formal, technical peer review processes.

Practical and Theoretical Requirements for Controlling Rater Stringency in Peer Review

Gerald J. Cason, Ph.D. and Carolyn L. Cason, Ph.D
University of Arkansas for Medical Sciences
Little Rock, Arkansas, USA 72205

Where the peer review process involves a large number of reviewers and proposals or manuscripts to be processed in a relatively brief period, for example a large international scientific conference, the control of rater bias requires both practical, logistical capabilities as well as a theoretical understanding of the rating process. The capabilities brought to bear in processing the reviews of abstracts submitted from the Americas, Pacific Oceana, India and Asia to Sigma Theta Tau (USA) for this conference were originally developed to address needs and problems in the assessment of the clinical performance of students in health professions training programs. The purpose of this paper is to briefly describe our performance rating information processing system, our performance rating theory, and their application capabilities. The prior existence of these capabilities permitted: (a) setting up the data collection procedures, collecting the data, and processing the data on a very brief schedule, but with no special staffing and almost no budget beyond postage costs; and, (b) once the data were in machine (i.e., magnetic) form, obtaining measures of abstract quality that were independent of the standards of the specific reviewers who happened to review particular abstracts. Here our performance rating system and our rating theory are briefly described. Their actual application to the reviews of abstracts for this conference is given in another paper in this symposium (C. Cason, Cason & Redland, 1987).

Performance Rating System

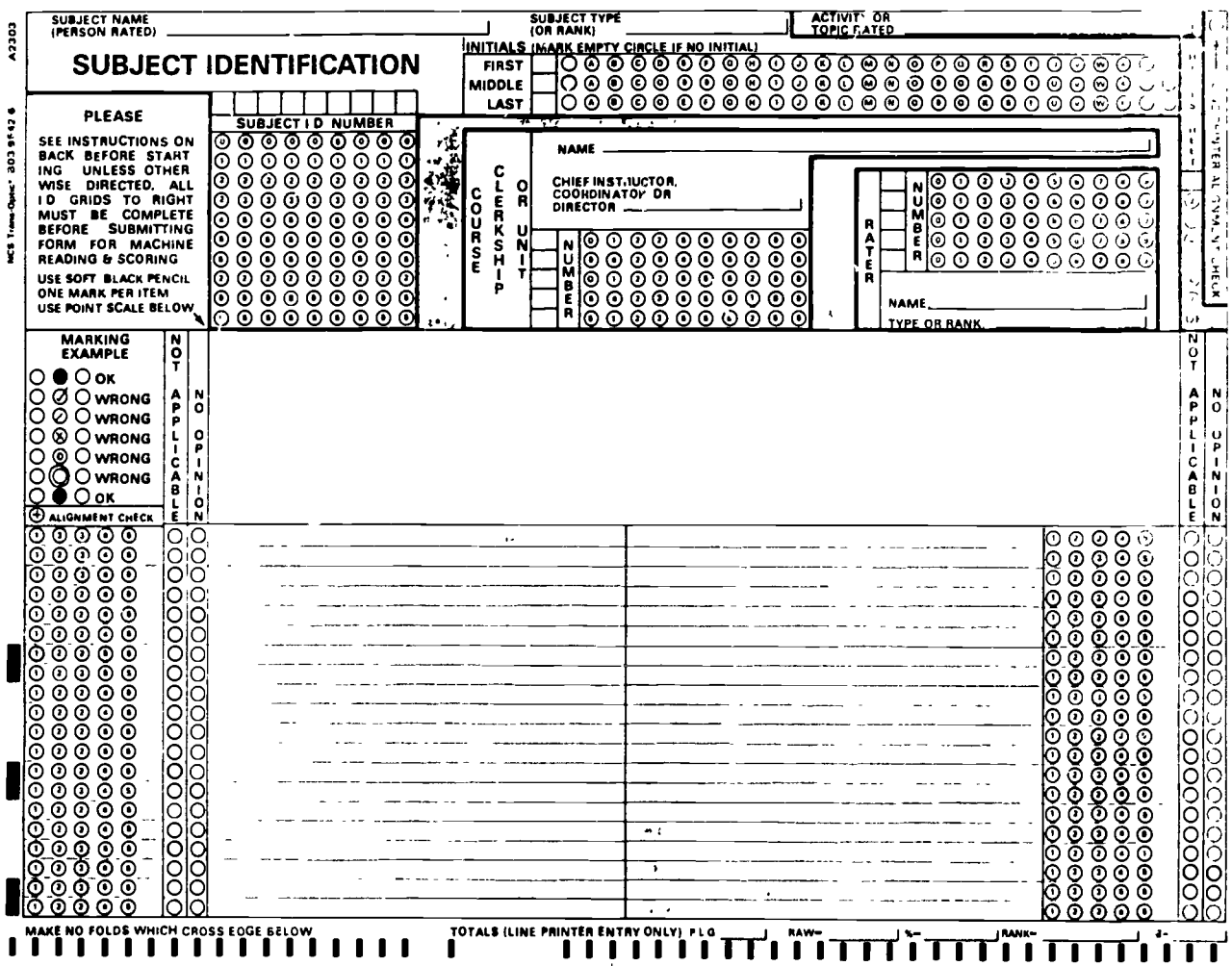
We developed the Performance Rating (PR) enhancement to the UAMS Objective Test Scoring (OTS) system to assist the clinical teacher to assess the clinical performance of students in much the same way that the OTS portion supports the classroom teacher. The system has proven useful in a wide range of applications, including evaluating clinical performance of nursing students, medical students and residents; the teaching performance of faculty; and, in one previous study, processing reviews of proposals for meetings of an international scientific organization (Cason, Cason & Stritter, 1986a; 1986b).

Because of the PR system's original purpose many of its capabilities are irrelevant to the present case of peer review data. For example, the PR system provides records keeping across multiple assessments (e.g., examinations and clinical performance rating occasions); allows the use of multiple, different rating inventories within a course, each inventory being tailored to the performance evaluation needs of specific clinical settings; and, easily allows the integration of scores from many different assessment methods: essays, multiple-choice questions, performance ratings. These capabilities have been previously described in detail with extensive examples of specific clinical performance rating applications (Cason, Schoultz, Cason, Glenn, Jones, Golden, Lang, & Doyle, 1986) and shall not be elaborated here. Only those features of the PR system that are relevant to the present topic are addressed in detail. In general, the capabilities of the PR system which make it useful for processing the peer reviews of abstracts submitted to a conference such as this include: the ease and speed with which a new inventory may be implemented; the low cost and speed of data collection, processing and reporting arising from the use of optically scanned rating sheets; and, the appropriateness of its reports.

The OTS-PR system's 70 modular FORTRAN programs run on a mainframe (Digital Equipment Corporation VAX-8530) computer. But, to ease accessibility to the faculty, the

user interacts with a full-time Assessment Scoring Coordinator (ASC), who in turn interacts with the OTS-PR system programs, scanner, and so forth. Annually, the ASC routinely supports an average of about 100 courses and other applications of OTS-PR. This service is available year round to faculty users without charge for internal projects and approved external projects, such as the support of analysis of abstract reviews for this conference. Thus, producing the abstract review rating sheets, processing them through the scanner, and obtaining OTS-PR reports was accomplished without special staffing or funding (other than postage costs) for the project. Further, these routine services provided by the ASC were delivered on the normal delivery schedule for any user and thus were well within the time requirements for this application. It normally takes the ASC one day to set up a new user's area in the computer. For a rating inventory not previously used in the system, three days are needed to deliver the scannable rating sheets (in total quantities of up to about 3000 sheets). When completed rating sheets (i.e., containing rating data) are returned by the user to the ASC, processing is completed and standard reports are normally available the following working day. The only unusual requirements associated with the processing of the peer review of research abstracts involved some programming changes to accommodated 400 subjects (i.e., abstracts) rather than the normal 200 maximum used in processing student performance data.

Figure 1. Blank General Purpose Rating Sheet



For this application we provided the ASC with the rating inventory (i.e., list of rating criteria and scale definitions) and a list of "names" of subjects to be rated (i.e., P1, P2, etc., for proposal 1, 2 etc.). This information was entered in the computer by the ASC. OTS-PR programs used this information, subject identification numbers, blank rating forms (illustrated in Figure 1), and the computer's line printer to produce as many copies of the sheet per abstract as we needed for distribution to the reviewers.

Figure 2. Rating Sheet with Generic Clinical Inventory

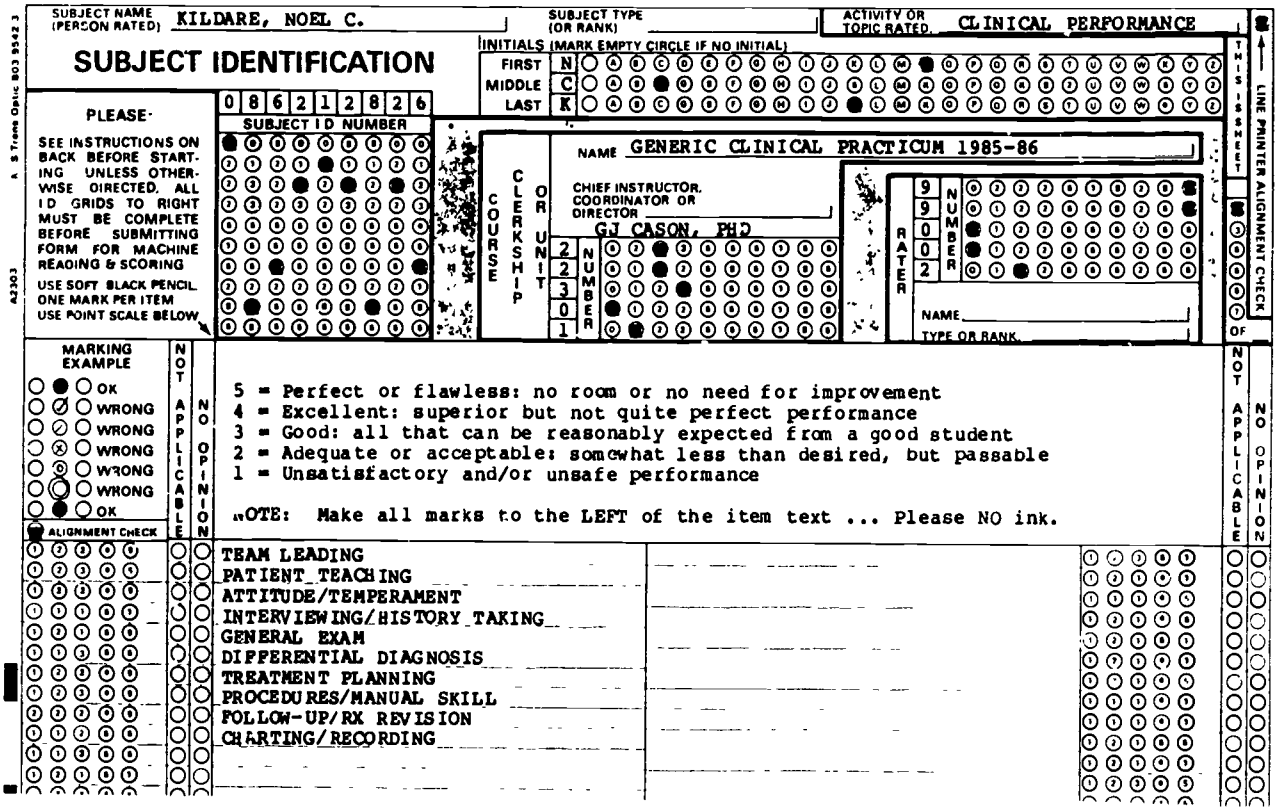


Figure 2 provides an illustration of the rating sheet with a hypothetical clinical performance rating inventory printed on it. Field tests with prototype rating data processing programs (circa 1978-79) demonstrated that it is essential to keep to a bare minimum the quantity of data manually entered on the rating sheet, especially that data entered by the rater. Otherwise, errors increase and excessive time for recording information reduces the acceptability and usefulness of the system. For this reason, the computer's line printer overprints both identifying information and the inventory's text on the machine scannable sheets. This includes "slugging" identification data in the scannable data grids. Using the line printer to print both the inventory and subject identifying data on the rating sheet provides the user maximum flexibility, ease of editing and revision, while also minimizing the quantity of information that must be manually entered on the sheets.

As can be seen in Figures 1 and 2, the sheet provides room for up to 40 one-line criteria of 35 characters each. The system permits sub-scales and multi-line rating items. The rater records his or her judgment by marking a circle: numbered 1 through 5 (with 5 always best), or labeled "no opinion", or "not applicable." Space for written comments is provided on the back of the sheet. The subject identification number printed on the rating sheet is not confidential. A different, confidential one is used in reports and records in the OTS-PR system.

For this application one of the options of the system was used and required unique rater ID numbers to be manually entered on the sheets. This permitted the automatic generation of reports on the performance of each abstract reviewer. Rater ID numbers were entered on the sheet by a member of the Sigma Theta Tau central office clerical staff.

The reports had by a specific user are determined by that user's selection from a menu of 17 available reports (some of which are listed in Table 1). Examples of only the three most relevant to the current topic, peer review for an international conference, are illustrated here.

Table 1: Partial List of OTS-PR Reports

For Current Assessment (Test or Rating)
- Assessment Instrument Analysis
  - Analysis Summary
  - Item Analysis
- Subject (e.g., student, abstract) Performance
  - Department (detailed scores for archive)
  - Subjects by Rank Order
  - Histogram of Subject Scores
  - Posting (subjects' names excluded)
  - Individual Subject Performance
- Individual Rater (e.g., reviewer) Performance

Across Multiple Assessments (Tests or Ratings)
- For Coordinator's Use
  - Alpha-ordered, all subjects' scores to date
  - Rank ordered, subjects' cumulative totals
- Subject's Individual Cumulative Scores
- Posting (subjects' names excluded)
  - ID-ordered, all scores to date
  - ID-ordered weighted totals

The Individual Performance Report (IPR) illustrated in Figure 3 provides information on a single subject (e.g., student, abstract) regarding a single rating occasion. The IPR gives item, subscale and total average ratings in both graphic and tabular form. In the graphic part of the report, the "x" profile makes it easy to rapidly determine, by visual inspection, this subject's relative strengths and weaknesses. The "c" profile provides a comparison with average ratings obtained by all members of the group on whom data were included when the report was prepared. When unique rater ID numbers are used, the report on an individual rater is similar in structure and appearance to the IPR. However, the individual rater report gives the average rating that the rater assigned to subjects he or she rated compared with the average of rater averages for each criterion.

The Students by Rank Order report illustrated in Figure 4 provides information on the whole group of subjects rated on a particular occasion. It gives a listing of all subjects (e.g., abstracts) from highest-scoring to lowest-scoring with total score reported in several units of measurement, e.g., percentage, rank, and Z-score (standard scores with mean = 500; s.d. = 100). Also given is the total number of raters (or rating sheets) that the total for each subject was averaged across. As Figure 4 shows, the number of raters per subject need not be constant. A similar Raters by Rank Order Report is also available. Raters are ranked by the average total rating they each assigned to subjects.

The Rating Analysis Summary Report is illustrated in Figure 5. It provides information on the performance of the assessment inventory and procedure. Information is given at the category (sub-scale) and total score level. If two or more raters rated each subject (student, abstract), then meaningful inter-rater reliabilities are reported. (When only one rater rates

Figure 3. Individual Performance Report

INDIVIDUAL PERFORMANCE REPORT (Current Rating)

Prepared 29-Jul-86 14:53 by the UAMS OIS/PR System (version B0) as implemented at UAMS

To: NOEL C. KILDARE
From: GJ CASON, PHD
Re: Rating/test 3-INPATIENT RATING - FACULTY

Dept: EDUCATIONAL DEVELOPMENT
Course: GENERIC CLINICAL PRACTICUM (SYNTHETIC DATA) 1985-86

Class Overall Mean Rating = 3.74-->
Your Overall Mean Rating = 3.92-->

Item	5 Point Scale 1.....2.....3.....4.....5	5 Pt Score: Your Mean=x	SEM	Class Mean=c	StdDev	Raw Score: Perfect p	Yours xp/5	N of Raters
1 TEAM LEADING	.-.x.-	4.00	.32	3.97	.428	5	4.00	4
2 PATIENT TEACHING	.x.-	3.75	.36	3.80	.513	5	3.75	4
3 ATTITUDE/TEMPERMENT	.-.c-.--x--	4.75	.32	4.08	.455	5	4.75	4
4 INTERVIEWING/HISTORY TAKING	.-c--x-.-	4.50	.38	3.92	.540	5	4.50	4
5 GENERAL EXAM	x--c-.-	3.00	.38	3.67	.552	5	3.00	2
6 DIFFERENTIAL DIAGNOSIS	x-c.-.-	3.33	.37	3.63	.481	5	3.33	3
7 TREATMENT PLANNING	xc.-.-	3.50	.30	3.61	.435	5	3.50	4
8 PROCEDURES/MANUAL SKILL	cx-.-	3.67	.41	3.64	.486	5	3.67	3
9 FOLLOW-UP/RX REVISION	xc.-.-	3.50	.39	3.64	.555	5	3.50	2
10 CHARTING/RECORDING	x--c.-.-	3.00	.39	3.57	.494	5	3.00	2

Rating Scale
5 = Perfect or flawless: no room or no need for improvement
4 = Excellent: superior but not quite perfect performance
3 = Good: all that can be reasonably expected from a good student
2 = Adequate or acceptable: somewhat less than desired, but passable
1 = Unsatisfactory and/or unsafe performance

Definition of Symbols
C = class overall 5 pt score: 3.74 StdDev: .398
X = your overall 5 pt score: 3.92 SEM : .24
c = class mean 5 pt score on item (or category)
x = your mean 5 pt score on item (or category)
SEM= Standard Error of Measurement

Your overall raw score 39.17 (out of perfect 50) yields: 78.3% Z= 544 Rank= 18 (out of 60). Class ave raw score 37.41

Figure 4. Students by Rank Order Report

STUDENTS BY RANK (Current Rating)

Prepared 30-Jul-86 14:27 by the UAMS OIS/PR System (version B0) as implemented at UAMS

Test: 3 - INPATIENT RATING - FACULTY
Instructor: GJ CASON, PHD
Course: GENERIC CLINICAL PRACTICUM (SYNTHETIC DATA) 1985-86

Dept: EDUCATIONAL DEVELOPMENT
Slot: 595 Phone:661-5720
Subjects rated: 60 Absent: 68 Withdrawn: 8

		RAW SCORE	5 PT SCORE	RANK RAW	RANK PCNTL	Z SCORE	N OF RATERS	NAME
	Averages	37.41	3.74			500	3.8	(Fictitious names and data)
1	286001297	49.33	4.93	1	98	799	3	ZEIMANOVICH, MARINA K.
2	286007980	46.22	4.62	2	96	721	3	BEAKER, JULIO M.
18	286002197	39.17	3.92	18	70	544	4	KILDARE, NOEL C.
60	286001645	28.67	2.87	60	0	281	3	ARNOLD, FRANKLIN D.

Figure 5. Rating Analysis Summary Report

RATING ANALYSIS SUMMARY (Current Rating)

Prepared 29-Jul-86 14:40 by the UAMS OIS/PR System (version B0) as implemented at UAMS

Test: 3 - INPATIENT RATING - FACULTY
Instructor: GJ CASON, PHD
Course: GENERIC CLINICAL PRACTICUM (SYNTHETIC DATA) 1985-86

Dept: EDUCATIONAL DEVELOPMENT
Slot: 595 Phone:661-5720
Subjects rated: 60 Absent: 68 Withdrawn: 8

Category	N Rated Items & Total Points (1)		Average Score Raw	5 Pt	1 Rater Reliab.	Mean N Raters & Reliability (2)		Standard Dev Raw	5 Pt	Std Error of Measure Raw	5 Pt	Z (3)
1 TEAM LEADING	1	5	4.0	3.97	0.17	3.7	0.43	0.4	.428	0.32	.32	75
2 PATIENT TEACHING	1	5	3.8	3.80	0.21	3.7	0.49	0.5	.513	0.37	.37	71
3 ATTITUDE/TEMPERMENT	1	5	4.1	4.08	0.20	3.7	0.48	0.5	.455	0.33	.33	72
4 INTERVIEWING/HISTORY TAKING	1	5	3.9	3.92	0.21	3.5	0.48	0.5	.540	0.39	.39	72
5 GENERAL EXAM	1	5	3.7	3.67	0.36	2.7	0.60	0.6	.552	0.35	.35	63
6 DIFFERENTIAL DIAGNOSIS	1	5	3.6	3.63	0.18	3.2	0.41	0.5	.481	0.37	.37	77
7 TREATMENT PLANNING	1	5	3.6	3.61	0.21	3.6	0.49	0.4	.435	0.31	.31	72
8 PROCEDURES/MANUAL SKILL	1	5	3.6	3.64	0.12	2.1	0.22	0.5	.486	0.43	.43	88
9 FOLLOW-UP/RX REVISION	1	5	3.6	3.64	0.35	2.8	0.60	0.6	.555	0.35	.35	63
10 CHARTING/RECORDING	1	5	3.6	3.57	0.24	3.0	0.49	0.5	.494	0.35	.35	72
Overall	10	50	37.4	3.74	0.30	3.8	0.61	4.0	.398	2.47	.25	62

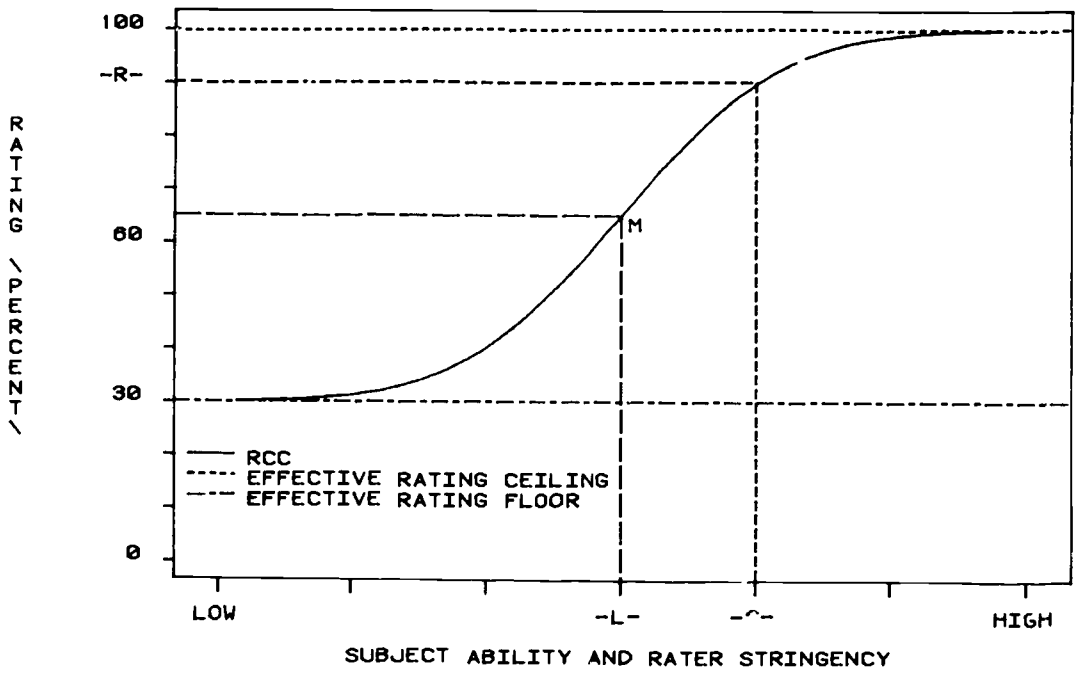
each subject a meaningless zero is reported.) Because the number of raters rating each subject may vary, the reliability of the average (geometric mean) number of raters rating the average subject is reported. The information provided in the rating analysis summary report can assist the user to improve both the reliability and validity of the rating inventory through practical trial, judicious editing, and revision. At UAMS the senior author of this paper and other members of the staff of the Office of Educational Development provide assistance to faculty users of OTS-PR to help them make best use of the information given in the rating analysis report.

One of the uses of the rater reports is the determination of the presence of differences in raters' standards, i.e., their stringency in evaluating the subjects of the assessment. If subjects are randomly assigned to raters, and each rater rates a sufficient number of subjects, there should be only a small variation in the observed mean rating given by each rater. Except in those settings where extensive and repeated rater training occurs, ordinarily there are practically important differences in the standards of different raters. While OTS-PR currently provides reports from which this may be inferred, it provides no way to test this inference, nor to off-set such differences should they exist. The development and application of our performance rating theory addresses this problem.

General, Qualitative Rating Theory

Our performance rating theory (Cason & Cason, 1981; 1984; 1986) evolved in response to our concern about the reliability and validity of ratings-based measures of complex human performance (and products), originally the patient care activities of health care professions students, especially where the usual methods of controlling systematic rater error (i.e., rater training, improved inventories, more raters per student, all raters rating all students) were frequently impractical. Essentially the same concerns arise for the same reasons in the typical process used to evaluate scientific products: proposals, abstracts, manuscripts. Our theory and its derivative simplified model of performance rating were developed to provide a mathematical basis for controlling systemic rater error or bias arising from differences in the standards and other characteristics of the raters (e.g., teachers, reviewers) who happen to judge an individual subject (e.g., student, abstract).

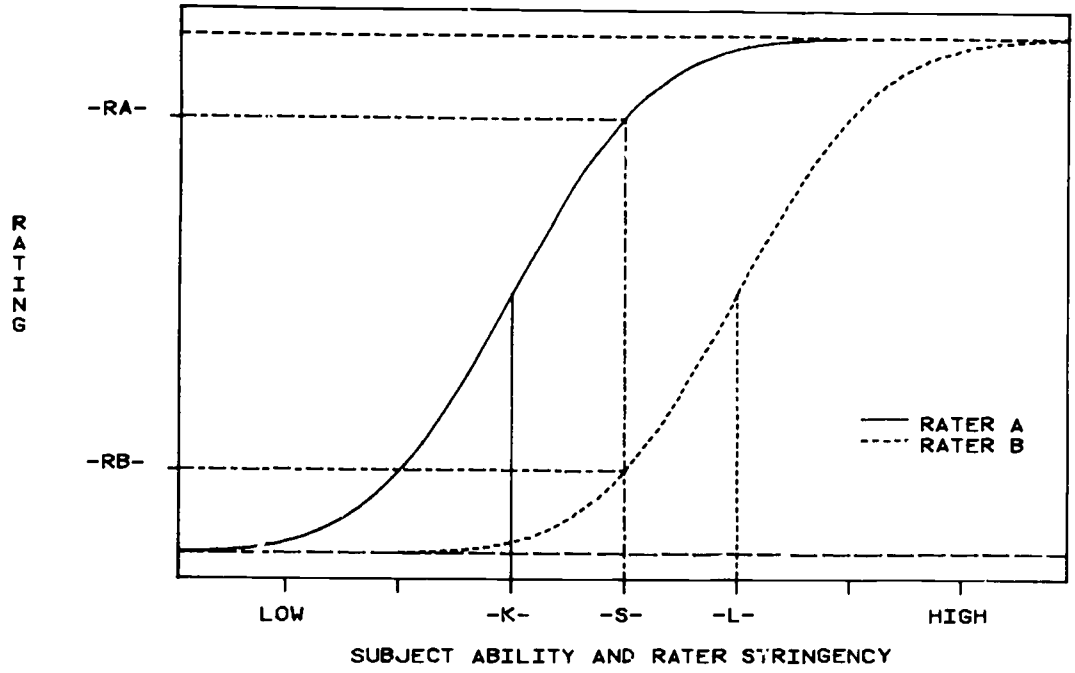
Figure 6. General Rater Characteristic Curve (RCC)



Our performance rating theory (Cason & Cason, 1984) posits that the rating received by a subject (e.g., student, abstract) is a function of the subject's true ability (i.e., competence, quality) and the rater's (e.g., teacher's, reviewer's) characteristics. The **rater's characteristics** include : (a) **resolving power**, i.e., the capacity to assign different ratings to different amounts of subject ability, (b) **sensitivity**, i.e., the maximum value of the rater's resolving power, (c) **stringency**, i.e., the general tendency to require more or less ability for a given rating when other characteristics are equal across raters, and (d) **effective rating floor and ceiling**, i.e., the minimum and maximum ratings a rater will actually assign in spite of the ostensible range, e.g., that printed on the rating inventory or scale. It is these characteristics (and random error) that determine the relationship between subject ability and assigned rating; this relationship is illustrated by the **rater characteristic curve** (RCC) in Figure 6. The RCC arises from the net, joint effects of the rater's knowledge, understanding, and beliefs about the task to be performed, its difficulty, constraints imposed by the setting or situation, the rating procedure (including the inventory used), and related factors.

Rater resolving power and the notion of a rater's pivotal rating standard are the two most primitive concepts underlying the theory. Resolving power is reflected in the slope (i.e., steepness) of the rater characteristic curve (RCC) at a given point. The assumed nature of the change in rater resolving power, as subject ability varies from very low (i.e., a great distance below the rater's pivotal standard) to very high (i.e., a great distance above the rater's pivotal standard), implies an s-shaped curve. Resolving power is not treated as a formal parameter of the theory. The **four formal rater parameters** of the theory, implied by the rater characteristics given above, are associated with mathematical parameters of the RCC. The first, **sensitivity**, is measured by the slope of the RCC at the point which evaluates to a rating half way between the rater's effective rating floor and ceiling. The projection of this point on the stringency scale is the rater's pivotal standard or **rater reference point** (RPP). The **effective rating floor and ceiling**, i.e., the asymptotic minimum and maximum ratings that the rater will actually assign (and not necessarily 0% and 100%, respectively) may

Figure 7. RCC's for Simplified Model: Raters of Stringencies K and L Give Subject of Ability S Ratings RA and RB, Respectively.



be viewed as the third and fourth formal parameters of the theory. Thus, the rating obtained by the subject is a function of the **subject's ability point** (SAP), i.e., the subject's true ability, and the location and shape of the RCC as defined by its general equation and values of its four parameters.

In our empirical work we use a simplified model of our theory to ease the problems of estimating the values of the formal subject and rater parameters. Our simplified model (Cason & Cason, 1984) accounts for all systematic variation in performance ratings exclusively by variation in one rater parameter (i.e., stringency) and one subject parameter (i.e., ability or quality). As illustrated in Figure 7, this simplified model assumes that (a) all raters have equal sensitivity (i.e., the slopes of the RCC's are equal) and (b) all raters have effective rating floors and ceilings of 0% and 100%, respectively. The model is applicable where there is sufficient over-lap in who rates whom, i.e., where there is sufficient coupling of the data. This coupling is frequently present in the structure of data found in health professions clinical education settings, i.e., where each student is rated by several but not all raters and each rater rates several but not all students. This structure is, of course, frequently found in the reviews of scientific proposals and abstracts.

The necessary coupling between data points (i.e., ratings) may be understood by analogy to acquaintanceship relationships. If a rating is assumed to be a metaphorical handshake and this is taken to mean the rater and subject are acquainted, then the simplified model applies to sets of data in which there is a path of mutual acquaintance leading from any subject or rater to every other subject and rater.

Application of the model in previous studies of clinical performance rating permitted off-setting variations in rater stringency and calculation of adjusted ratings having improved reliability and validity in each of 17 independent data sets obtained from two schools, with different amounts of rater training, different rating inventories (one behaviorally anchored, one not), and each inventory having different levels of trait specificity. C. Cason, Cason, and Littlefield (1983) further demonstrated that the model was equally applicable to each of two commonly cited (e.g., Dielman, Hull, & Davis, 1980) dimensions of clinical performance (i.e., cognitive-technical versus affective-interpersonal skills). The application of the model was demonstrated to be equally useful on the ratings on paper proposals considered for presentation at three meetings of an international scientific organization (Cason, Cason & Stritter, 1986a; 1986b).

Simplified, Quantitative Rating Model

In our simplified model the **expected subject rating** (ESR), expressed as a percent of the maximum possible 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). This relationship is modified by an arbitrary **scaling factor** (SF = 100).

$$z = (SAP - RRP)/SF \qquad [1]$$

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

$$ESR = p(z) * 100 \qquad [2]$$

This is a deterministic not a probabilistic relationship. The model predicts a point value for the expected rating, not a probability distribution.

In previous research the **observed subject rating** (OSR) was defined as equal to ESR plus (random) error:

$$OSR = ESR + error \qquad [3]$$

Converting from percent to proportion (dividing by 100) and substituting the definition of ESR from Equation 2 gives:

$$OSR = p(z) + error \qquad [4]$$

Because OSR is (now) a proportion, it must fall between 0 and 1. From Equation 4, it follows that the sum of the error and area below z also must fall between 0 and 1. That is, the sum of the expressions on the right side of Equation 4 may be treated as a proportion. Without asserting anything different about the psychological location of the random error in the rating process, the model may be expressed as:

$$OSR = p(z + error) \qquad [5]$$

Taking the **inverse normal probability** (ZIN, i.e., obtaining the z associated with a given proportion) of both sides of Equation 5 gives:

$$ZIN(OSR) = z + error \qquad [6]$$

That is, Equation 6 shows that the inverse z-transform of the observed ratings is composed of the difference (z) between subject ability (SAP) and rater stringency (RRP) plus random error. Equation 6 permits the application of regression analysis to estimate these parameter values rather than the less well known procedure used in earlier studies to solve Equation 3. While Equation 3 was convenient for earlier studies, Equation 6 will provide equivalent (to within a linear transformation) estimates of subject ability and rater stringency using a more generally familiar method.

Regression Based Estimates of Parameters

Estimation of the model parameters (i.e., RRP's and SAP's) is accomplished in two phases. First, the observed ratings are transformed to proportions then to z's (using an inverse normal probability function). These z's are used as the criterion values (Y vector) in a regression model of the general form of Equation 7. The z's in the criterion vector may be thought of as distances on the underlying stringency-ability scale (but containing error as in Equation 6) between RRP's and SAP's which are implied by the original observed ratings.

$$Y = cU + b_1R^1 + b_2R^2 + \ldots + b_nR^n + b_{n+1}S^{n+1} + b_{n+2}S^{n+2} + \ldots + b_{n+k}S^{n+k} + E \qquad [7]$$

where:

Y is the criterion vector;

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

R^i (i=1 through n; n=number of raters) 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; k=number of subjects) is a vector containing a 1 if the observationin Y is associated with subject j-n, zero otherwise; and,

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

A special purpose computer program, GENVEC (Cason, 1987) is used to generate the above model from input (files from OTS-PR) which specifies a rater identification number and subject identification number for each observed rating. Program LMS (Linear Model Solver; Cason, 1986), based on Ward and Jennings' (1973) program MODEL, gives an appropriate regression analysis of the model generated by GENVEC. (The regression program must allow redundant vectors in the model.) LMS yields least-squares, raw regression weights (b's not beta's) for Equation 7. As shown in Equation 8, pairs of b's and the unit vector weight provide an estimate of the inferred "error free" distance between a rater-subject pair on the stringency and ability scale:

$$RXTOS(I) = BOFS(I) - BOFRX + CONST \quad [8]$$

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 regresion constant (i.e., unit vector weight).

The second phase of the process of estimating the parameter values is to convert the regression weights into theoretical distances (using Equation 8) and then into locations (i.e., RRP's and SAP's) on the stringency and ability scale (using Equation 9). A rater is arbitrarily chosen (e.g., the rater with most ratings and lowest ID number) to anchor the scale and that rater's RRP is set equal to 500. Once this is done, the location of all subjects is determined (with respect to the arbitrary anchor point) by the linear equation:

$$SLOC(I) = ANCVAL + RXTOS(I) \quad [9]$$

where:
SLOC(I) is the location (SAP) of subject I; and,
ANCVAL is the arbitrary value (e.g., 500) used to anchor the scale.

As soon as all subjects are located, an analogous set of equations is solved to obtain the remaining rater locations (i.e., rater stringencies or RRP's). Program LOCATE (Cason, 1985) is used to solve Equations 8 and 9 and the analogous equations for raters to obtain estimates of all model parameter values, i.e., RRP's and SAP's. LOCATE also calculates a calibrated or adjusted rating for each subject. The adjusted rating is calculated (using Equations 1 and 2) as the mean of the ratings expected from a subject's SAP and the RRP's of all the raters. Thus, the adjusted or calibrated rating is what the subject would have received (disregarding random measurement error) had all the raters rated the person's performance or product.

Summary and Conclusions

Our work over the last ten years in clinical performance evaluation yielded the theoretical understanding of the rating process, the computer programs and the administrative support unit that permitted practical application of this understanding to the specialized needs of the abstract review process for this international research conference. This suggests that two of our major goals for our performance rating system, i.e., flexibility and practicality, have been substantially achieved. It may also suggest one set of reasons why the peer review process has as yet been so little affected by the advent of computers. Computer programs which are sufficiently appropriate to the needs of peer review processes have not been generally available. In most peer review settings, the use of general purpose statistical or data base programs and manual data entry to accomplish only compilation and reporting of observed ratings would require impractically large expenditures of money or staff time. Even if

available, the specialized programs for finding calibrated ratings (from which the effects of differences in rater standards have been removed) may not be practically applied unless an economical mechanism is present for converting the abstract reviews into quantitative, magnetic form. Most problems in clinical performance evaluation and peer review are not responsive to quick, easy, simple or cheap solutions. But, improvements can be made through development and use of appropriate data management and analysis systems. Even though such systems are very expensive to develop, the costs can be justified if the resulting systems have sufficient flexibility to be broadly applicable to peer review and or performance evaluation. As systems similar to OTS-PR become more commonly available, we may expect a concomitant improvement in the reliability and validity of formal, technical peer review processes.

References

Cason, C.L., Cason, G.J., & Littlefield, J.H. (1983, April). Variation in intra-rater stringency in cognitive-technical and affective-interpersonal clinical performance domains. Presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Cason, C.L., Cason, G.J., & Stritter, F.T. (1986). Reviewer standards in Division I program selection. Resources in Education, 21(10), 145. (Abstract) ERIC Document ED-270-458

Cason, C.L., Cason, G.J., & Stritter, F.T. (1986). Reviewer stringency and proposal quality in the selection of the 1986 AERA Division I program. Professions Education Research Notes, 8(1), 7-8.

Cason, C.L., Cason, G.J., & Redland, A. (1987, July). Off-setting differences in reviewer stringency. Presented at the International Research Congress sponsored by Sigma Theta Tau International Honor Society, the Royal College of London, and the University of Edinburgh, Edinburgh, Scotland.

Cason, G.J. (1985). LOCATE (Version 2): From regression weights and constant, computes first distances between rater-subject pairs, then locations of raters (RRPs) and subjects (SAPs) according to Cason and Cason's simplified model [Computer program]. Little Rock, AR: University of Arkansas for Medical Sciences.

Cason, G.J. (1986). LMS (Version 3): A FORTRAN Linear Model Solver based on Ward and Jennings (1973) program MODEL [Computer program]. Little Rock, AR: University of Arkansas for Medical Sciences.

Cason, G.J. (1987). GENVEC (Version 4): A FORTRAN program to structure input for program LMS [Computer program]. Little Rock, AR: University of Arkansas for Medical Sciences.

Cason, G.J., & Cason, C.L. (1981, April). Some promising early results from a rudimentary latent-trait theory of performance rating. Paper presented at the annual meeting of the American Educational Research Association, Los Angelos.

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. (1986). A regression solution to Cason and Cason's model of clinical performance rating: Easier, cheaper, faster. Resources in Education, 21(2), 147. (Abstract) ERIC Document ED-262-079

Cason, G.J., Schoultz, T.W., Cason, C.L., Glenn, R.E., Jones, J.G., Golden, K.A., Lang, N.P., & Doyle, K.L. (1986). A flexible system for processing clinical performance ratings: Illustrative applications in a residency and four clerkships. In M. B. Anderson (Organizer), Innovations in medical education exhibits: 97th annual meeting of the AAMC (p. 35). Washington, D.C.: Association of American Medical Colleges, Group on Medical Education. (Abstract) Resources in Education, June 1987. (Abstract) ERIC Document ED-278-692.

Dielman, T.E., Hull, A.L., & Davis, W.K. (1980). Psychometric properties of clinical performance ratings. Evaluation & the Health Professions, 3(1), 103-117.

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