DOCUMENT RESUME

ED 287 888	TM 870 649

AUTHOR	Cason, Carolyn L.; And Others
TITLE	Off-Setting Differences in Reviewer Stringency.
PUB DATE	Jul 87
NOTE	22p.; Paper presented at the Sigma Theta Tau International Research Congress (Edinburgh, Scotland, July, 1987). Some tables are marginally legible; for related document, see TM 870 648.
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; Validity
IDENTIFIERS	Performance Rating Theory (Cason and Cason); Sigma Theta Tau International Research Congress

ABSTRACT

A total of 1,071 rating sheets were completed by individual reviewers evaluating abstracts submitted to Sigma Theta Tau International Honor Society for the International Research Congress in Edinburgh, Scotland. Only 972 sheets contained usable data. Reviewers indicated a total of 12 ratings with possible comments for each abstract on machine scannable rating sheets specifically prepared for this use. Application of Cason and Cason's simplified model of their deterministic theory to reviews of abstracts resulted in good fit of the model to the data, and detection of significant inter-reviewer differences in individual reviewer's stringency. The availability of calibrated ratings, i.e. those from which these reviewer differences had been removed, greatly eased the task of the Abstract Selection Committee. The Committee no longer needed to deal with ratings which were confounded with variation in reviewer standards. Abstract selection was based on ratings that more accurately reflected the true quality of the abstract without regard to who happened to have reviewed it. When the Abstract Selection Committee used these calibrated ratings in making selection decisions, there were great improvements in both the reliability (from .579 to .810) and validity (from .485 to .742) of the peer review process. (Author/KSA)

Off-setting Differences in Reviewer Stringency*

Carolyn L. Cason, R.N., Ph.D.

Gerald J. Cason, Ph.D.

University of Arkansas for Medical Sciences
Little Rock, Arkansas, U.S.A.

and

Alice Redland, R.N., Ph.D.

University of Texas
Austin, Texas, U.S.A.

Address correspondence to:

Carolyn L. Cason, R.N., Ph.D.
College of Nursing-529
University of Arkansas for Medical Sciences
4301 West Markham
Little Rock, AR 72205
U.S.A.

Telephone: (501) 661-5163

*Presented at the International Research Congress, Edinburgh Scotland, July 1987.

We wish to thank the members of the Abstract Selection Committee and staff of the Program Office, especially Helen Ferrence and Susan Mihalo, of Sigma Theta Tau International Honor Society, U.S.A., for their cooperation and support. Their cooperation made this study possible.

Members of the Department of Computer Services, University of Arkansas for Medical Sciences, U.S.A., especially Mark Boughter, Tom Hart, and Tom Lewis, provided special programming and data processing support.

Off-setting Differences in Reviewer Stringency

Abstract

Application of Cason and Cason's (1984) simplified model of their deterministic rating theory to reviews of abstracts submitted to Sigma Theta Tau for the International Research Congress, Edinburgh, Scotland, resulted in good fit (R >.89) of the model to the data and detection of significant (p <.0001) inter-reviewer differences in individual reviewer's stringency. The availability of calibrated ratings, i.e., those from which these reviewer differences had been removed, greatly eased the task of the Abstract Selection Committee: they no longer needed to deal with ratings which were confounded with variation in reviewer stringency (as occurs in the observed mean ratings). Abstract selection was based on ratings that more accurately reflected the true quality of the abstract without regard to who happened to have reviewed it. Use by the Abstract Selection Committee of these calibrated ratings in making selection decisions greatly improved both the reliability (from .579 to .810) and validity (from .485 to .742) of the peer review process.

Off-setting Differences in Reviewer Stringency

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

Alice Redland
University of Texas, Austin, TX, USA

Peer review serves as the basis for making many highly important decisions: funding of proposals, publication of manuscripts, papers to be presented at professional meetings such as this. As such it determines, in part, what knowledge will be sought by and shared with the scientific community and by whom. What is selected and the process by which it is selected is very important to both the body of scientific information and the individual researcher's professional career.

The Abstract Selection Committee, Sigma Theta Tau International Honor Society, made the decisions about which abstracts submitted to it for the International Research Congress, Edinburgh, Scotland would be selected for inclusion in the program. The Committee was assisted in its task by multiple reviews completed by volunteer reviewers of each abstract submitted. As most of us who have had our work reviewed by different reviewers know, there seems to be substantial variation in the apparent standards of reviewers. To the degree that such variation exists among Sigma Theta Tau reviewers, the task of the Abstract Selection Committee becomes more difficult and both the reliability and validity of the process become compromised.

Interestingly, with few exceptions, variation in reviewer standards and its impact on the review process have been largely unevaluated and even less attention has been given to adjusting for differences in reviewer standards. Two recent exceptions are Marsh and Ball's 1981 study of variation among reviewers of manuscripts submitted to the Journal of Educational Psychology and our own work on paper proposals submitted for consideration in the program of Division I: Professions Education, American Educational Research Association (Cason, Cason, & Stritter, 1986a and 1986b). Marsh and Ball found no significant reviewer effect but this may well have been because their data contained a rather large number of manuscripts which had been reviewed by only a single reviewer (excluding the journal editor's review) and many (i.e., two thirds of the) reviewers who had reviewed only one or two manuscripts. In our previous study, we found significant and important reviewer effects in the reviews of paper proposals; effects which reduced both the reliability and validity of the selection process. Both of these studies were retrospective, that is they used data on selection decisions already made: reviewer effects were not formally considered in making actual decisions about acceptance or rejection of the manuscript/proposal.

This study of the Sigma Theta Tau abstract review and selection process had two objectives. They were

1. To evaluate the extent to which variation in standards/stringency exists among reviewers.

2. To provide to the Abstract Selection Committee, ratings of abstracts from which the effects of such variation had been removed.

Our performance rating theory and its derivative simplified model served as the framework for the study (Cason & Cason, 1984; Cason & Cason, 1986). It was briefly described in another paper in this forum (Cason & Cason, 1987). Application of the theory and simplified model have as an objective detection, quantification, and mathematical control of variation among reviewer/rater stringencies. Mathematical control of differences in reviewer stringency is intended to augment the more usual methods of controlling error associated with ratings (e.g., rater training, improved inventory reliability, all raters/reviewers rating all subjects/abstracts) and to off-set systematic rater error when such methods are impractical. Application of the model may be likened to a calibrat'on procedure in which knowledge of the reviewers' stringencies is used to adjust or calibrate abstract ratings so as to take into account the different stringencies of the reviewers who reviewed abstracts.

Data Source and Methods

The ratings given by individual reviewers to research abstracts submitted to Sigma Theta Tau International Honor Society for the International Research Congress in Edinburgh, Scotland served as the data for these analyses. Reviewers indicated their ratings of each abstract on machine scannable rating sheets specifically prepared for this use. Figure 1 illustrates this special scan sheet.

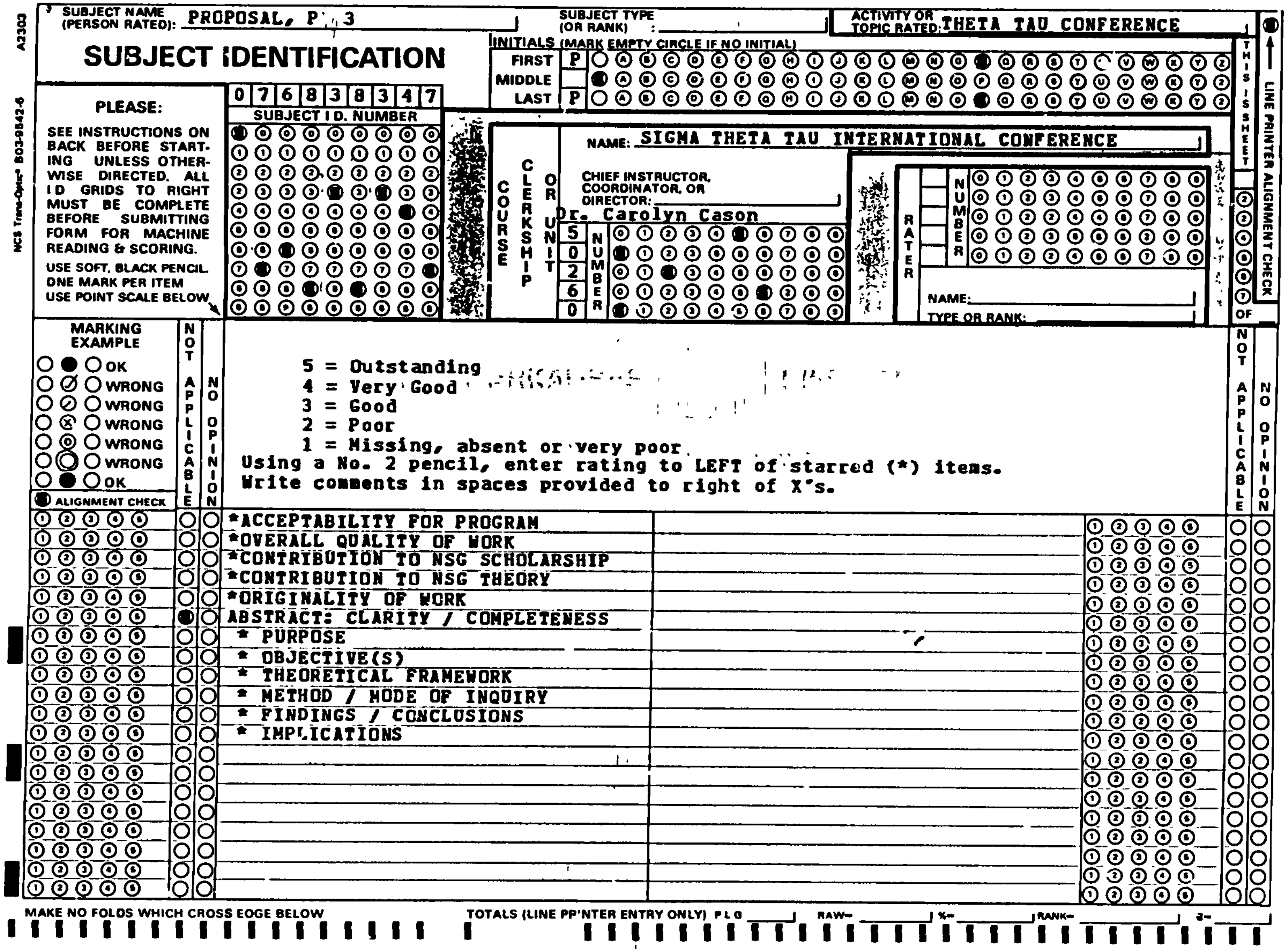
These machine scannable rating sheets were pre-printed (with a computer's line printer) by the Department of Computer Services, University of Arkansas for Medical Sciences (UAMS), U.S.A., with the information contained in Figure 1: identifying information (subject name or in this case abstract number), criteria or inventory of rating items upon which each abstract was rated, and the scale to be used for making the ratings. Two copies of each of 650, i.e., abstract identification numbers P001 through P650, were printed. So that information about the performance of reviewers could be obtained, a list of 75 unique rater identification numbers was provided by the Department of Computer Services. Preparation of such rating sheets and identifying numbers is a routine service provided to faculty who intend to use the UAMS Objective Test Scoring and Performance Rating (OTS-PR) system for clinical performance rating of students enrolled in the various programs on the UAMS campus.

These pre-printed rating sheets and the rater identification numbers were sent to the Program Office, Sigma Theta Tau International Honor Society. As individuals agreed to serve as voluntary reviewers, they were assigned by a staff member one of the rater identification numbers. As an abstract was received in the Program Office it was logged in and given an identification number by the staff of the office. All abstracts (research and congress-related topic) were numbered sequentially as they were received. For those abstracts identified as research abstracts, the staff members obtained the rating sheets with the corresponding proposal identification number (subject name), selected the two individuals who would serve as reviewers, entered the reviewer's identification number on the appropriate rating sheet, and sent the rating sheet and abstract to each of the two reviewers. Reviewers for each abstract were selected randomly by the Program Office staff with the only restriction being that the reviewer not be located in the same institution or general area of the country as the author(s) of the abstract.

Upon receipt of the rating sheet and research abstract, the reviewers recorded their ratings of the abstract on each of six general criteria: acceptability for program, overall quality of work, contribution to nursing scholarship, contribution to nursing theory, originality of work, and clarity and completeness of the abstract. These are shown in Figure 1. The last general criterion contained six specific items to be used in evaluating the abstract's clarity and completeness: purpose, objective(s), theoretical framework, method/mode of inquiry, findings/conclusions,and implications. Thus, reviewers were asked to make a total of 12 ratings on each abstract by filling in the numbered circle to the left of each item on the inventory which represented their rating of the abstract under consideration. Possible responses included: outstanding, very good, good, poor, and missing, absent or very poor; and, not applicable and no opinion. Reviewers who wished to make comments about the abstract could do so in the space provided to the right of the items. Finally, reviewers were asked to sign the sheet and return it to the Program Office.

Completed rating sheets were collected by the Program Office staff and then forwarded to us for processing. A total of 1071 rating sheets were forwarded. Of these, only 972 contained usable data. Rating sheets were not usable primarily for two reasons: reviewers gave no

Figure 1. Pre-printed rating sheet with inventory.



SUBJECT NAME (PERSON RATED): PROPOSAL, P 3

SUBJECT TYPE (OR RANK):

ACTIVITY OR TOPIC RATED: THETA TAU CONFERENCE

← LINE PRINTER ALIGNMENT CHECK

SUBJECT IDENTIFICATION

INITIALS (MARK EMPTY CIRCLE IF NO INITIAL)

FIRST	P	
MIDDLE		
LAST	P	

THIS IS SHEET OF

PLEASE:

SEE INSTRUCTIONS ON BACK BEFORE STARTING UNLESS OTHERWISE DIRECTED. ALL I D GRIDS TO RIGHT MUST BE COMPLETE BEFORE SUBMITTING FORM FOR MACHINE READING & SCORING.

USE SOFT, BLACK PENCIL. ONE MARK PER ITEM USE POINT SCALE BELOW

SUBJECT I D. NUMBER: 0 7 6 8 3 8 3 4 7

COURSE CLERKSHIP OR UNIT

NAME: SIGMA THETA TAU INTERNATIONAL CONFERENCE

CHIEF INSTRUCTOR, COORDINATOR, OR DIRECTOR: Dr. Carolyn Cason

NUMBER: 5 0 2 6 0

RATER NUMBER

NAME:

TYPE OR RANK:

MARKING EXAMPLE

○ ● ○ OK
○ ⊘ ○ WRONG
○ ⊘ ○ WRONG
○ ⊗ ○ WRONG
○ ◎ ○ WRONG
○ ◯ ○ WRONG
○ ● ○ OK

● ALIGNMENT CHECK

NOT APPLICABLE

NO OPINION

5 = Outstanding
4 = Very Good
3 = Good
2 = Poor
1 = Missing, absent or very poor

Using a No. 2 pencil, enter rating to LEFT of starred (*) items.
Write comments in spaces provided to right of X's.

*ACCEPTABILITY FOR PROGRAM
*OVERALL QUALITY OF WORK
*CONTRIBUTION TO NSG SCHOLARSHIP
*CONTRIBUTION TO NSG THEORY
*ORIGINALITY OF WORK
ABSTRACT: CLARITY / COMPLETENESS
* PURPOSE
* OBJECTIVE(S)
* THEORETICAL FRAMEWORK
* METHOD / MODE OF INQUIRY
* FINDINGS / CONCLUSIONS
* IMPLICATIONS

MAKE NO FOLDS WHICH CROSS EDGE BELOW

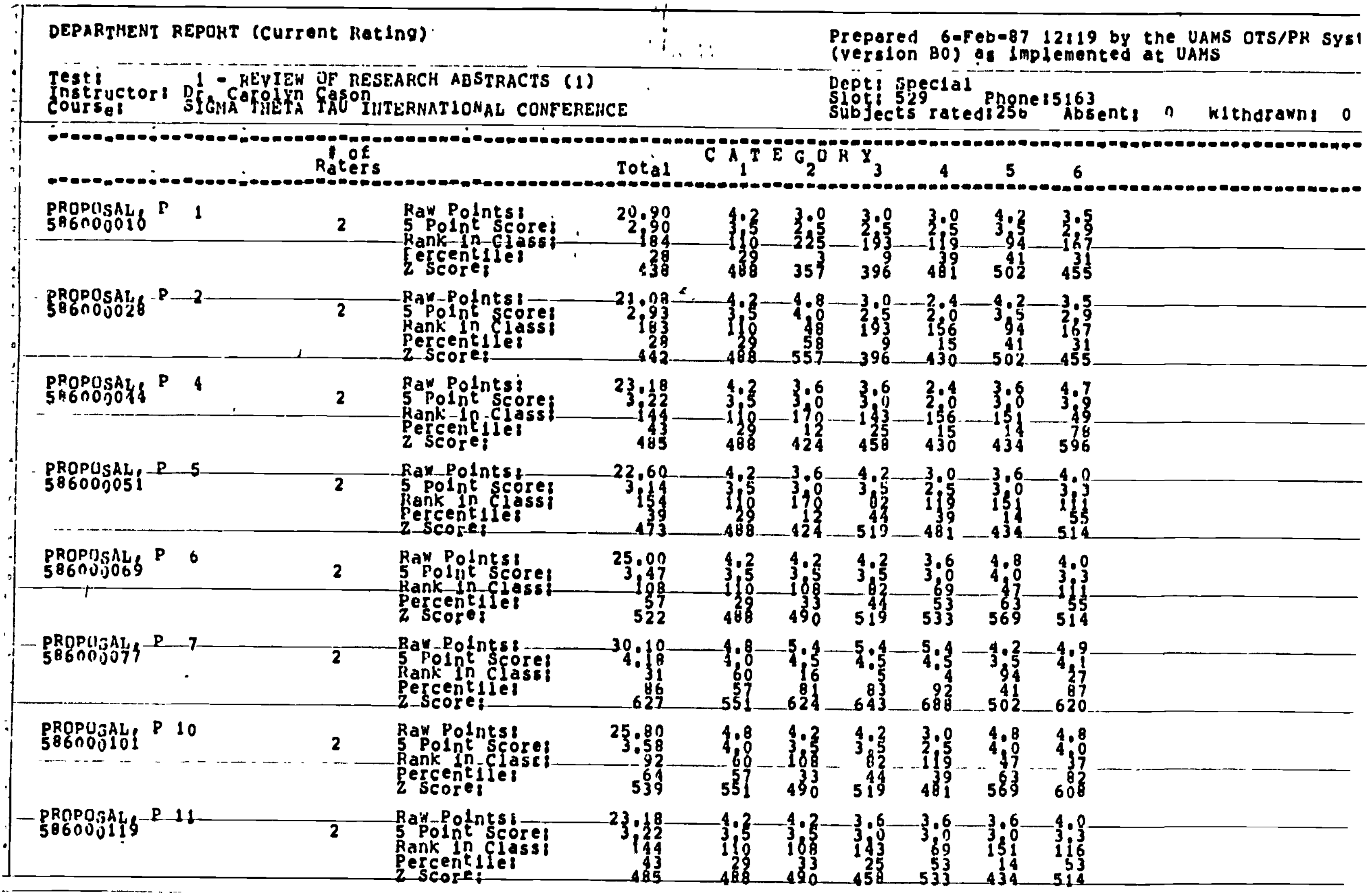
TOTALS (LINE PRINTER ENTRY ONLY) PLG RAW- %- RANK- ā-

NCS Trans-Optic® B03-9542-6 A2303

ratings but returned the sheets with comments indicating that they were personally familiar with the research or its author(s) or they thought that the abstract was not a research abstract but rather should be considered as a congress-related topic. The 972 usable rating sheets contained ratings of 503 abstracts (only one rating sheet was available on 26 abstracts while two rating sheets were available on each of the other 477 abstracts). There were a total of 61 volunteer reviewers. The actual number of research abstracts reviewed by each reviewer varied from 4 to 28. On average, each reviewer reviewed 16 abstracts (S.D.=5). (These reviewers also served as reviewers of abstracts on congress related topics. Their reviews of those abstracts are not reflected in these analyses).

Two sets of analyses were conducted on the rating data provided by the abstract reviewers. The item-level observed ratings contained on the rating sheets were processed through the programs of the OTS-PR system. OTS-PR produced its standard set of reports on subjects (abstracts), raters (reviewers), and the assessment procedure (i.e., rating inventory). The second set of analyses was accomplished by specialized computer programs which provide estimates of formal parameters of the simplified model of our rating theory; that is, computer programs which provide estimates of reviewer stringency and abstract true quality. Parameter estimation is achieved using a highly specialized application of regression analysis (Cason & Cason, 1986). The specialized computer programs also provide estimates of (a) the fit of the model with the data, (b) contribution of reviewer stringency and abstract quality to the observed ratings (i.e., reviewer and abstract effects), and (c) calibrated ratings which represent the rating expected when reviewer effects are removed (i.e., if all reviewers rated all abstracts and the average for each abstract was used). In the item-level quantitative analyses carried out by the OTS-PR programs, scale values were defined as depicted in Figure 1, i.e., 5=outstanding, 4=very good, 3=good, 2=poor, 1=missing, absent, or very poor. Responses of not applicable or no opinion were dealt with as missing data. Estimation of reviewers' stringency and abstracts' true quality were completed using a weighted total percent score as the observed dependent measure (i.e., regression criterion variable) for each reviewer-abstract pair in the observed data (i.e., each rating sheet). This score was obtained by first finding a mean rating for all of those items associated with the general criterion, "abstract clarity and completeness". This score and the rating assigned to each of the other five criteria were summed, divided by the number of criteria (6) and multipled by 100 (i.e., expressed as a percentage). This transformation was made to simplify and facilitate the analyses.

In preparation for analyzing the item-level observed data, the OTS-PR programs were modified so that the system could process ratings on up to 400 unique subjects (abstracts). The actual number of abstracts reviewed was 503. Time did not permit altering these programs again; therefore, in order to complete the initial processing of these ratings, the rating sheets were divided into two subsets: the first subset contained ratings on 256 abstracts; the second, ratings on 247 abstracts. In general, the first subset contained those ratings of abstracts which were forwarded to us first (i.e., about mid-January) while the second set contained all others (i.e., those forwarded between mid-January and January 30). Each set of data was processed separately through the OTS-PR system. In order to apply the simplified model to the ratings of research abstracts, the data had to be divided into 3 subsets. Each of the two OTS-PR system data subsets were too large; the DEC-System 10 mainframe computer, used to complete all of the analyses, and associated memory and disk space could not efficiently execute the programs. The three sets were determined randomly with the restriction that all ratings for an abstract were contained in a single data set. When this requirement was not met, the ratings on an abstract were merged into the smaller data set. (Due to a clerical error, each of 3 abstracts were located in more than one data set.) Each of the three newly created data subsets (AYE, BEE, and CEE) was evaluated to determine that each met the coupling assumption required by the model. Using programs FIXSTT, GVEC4, LMS, and LOCATE, each data set was analyzed separately for rater effects and to obtain estimates of rater stringency and abstract quality.



DEPARTMENT REPORT (Current Rating)

Prepared 6-Feb-87 12:19 by the UAMS OTS/PR Sys
(version B0) as implemented at UAMS

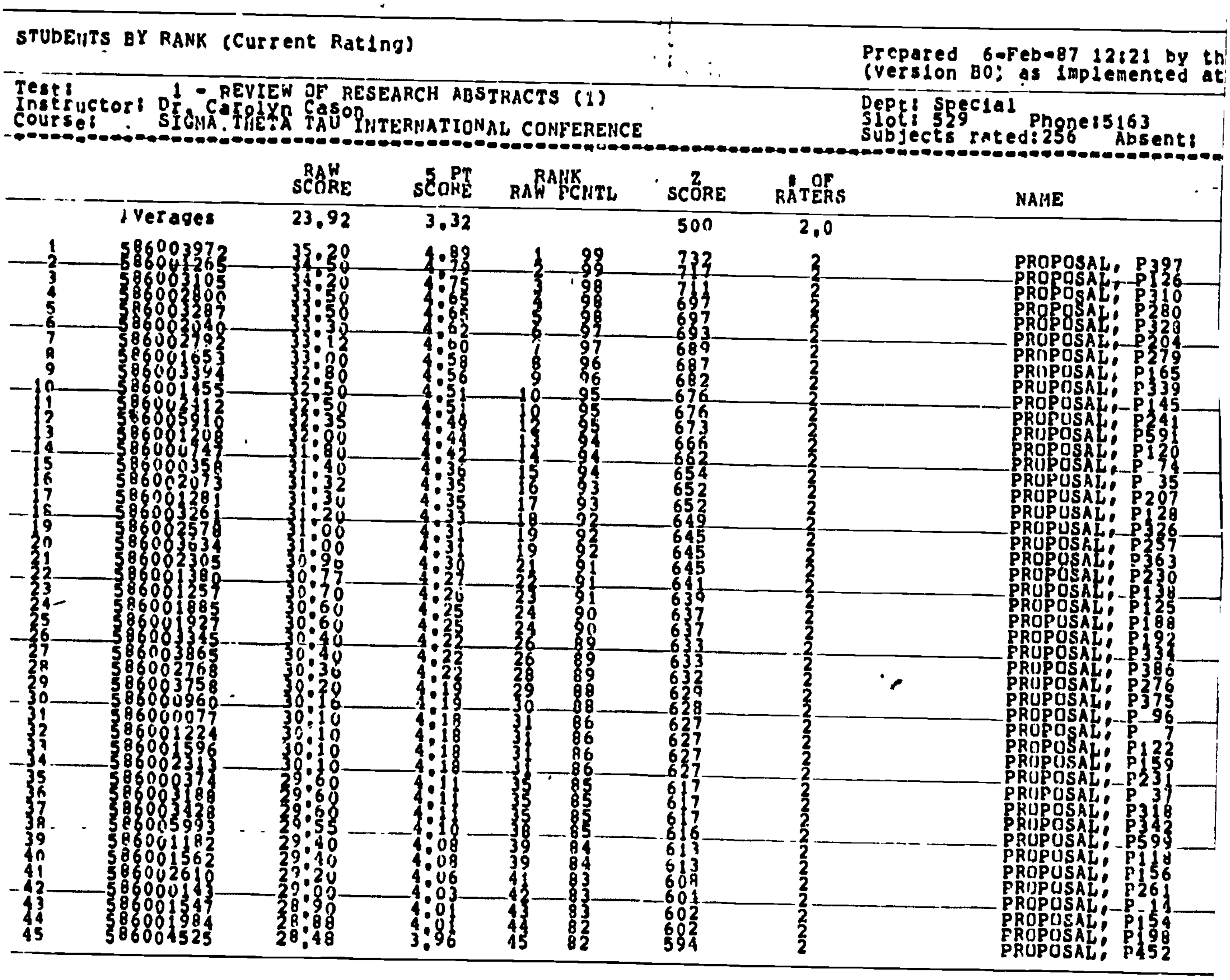
Test: 1 - REVIEW OF RESEARCH ABSTRACTS (1)
Instructor: Dr. Carolyn Cason
Course: SIGMA THETA TAU INTERNATIONAL CONFERENCE

Dept: Special
Slot: 529 Phone:5163
Subjects rated:256 Absent: 0 Withdrawn: 0

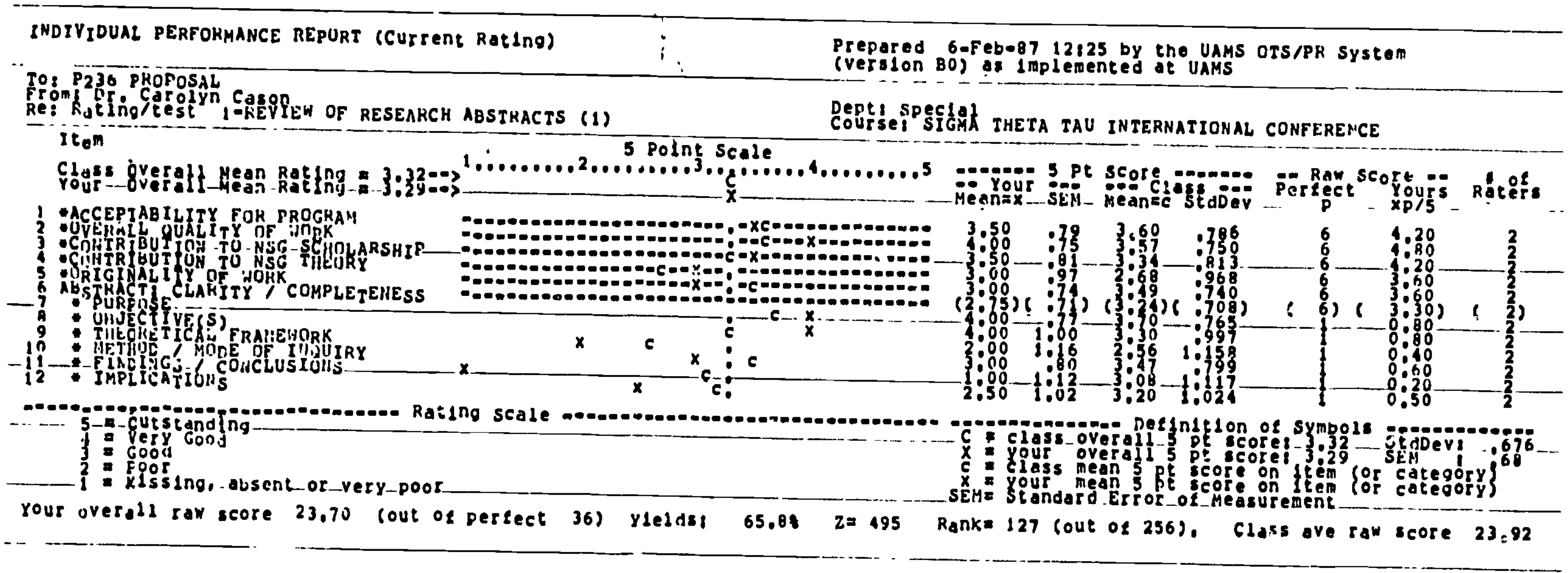
	# of Raters		Total	CATEGORY					
				1	2	3	4	5	6
PROPOSAL, P 1 586000010	2	Raw Points:	20.90	4.2	3.0	3.0	3.0	4.2	3.5
		5 Point Score:	2.90	3.5	2.5	2.5	2.5	3.5	2.9
		Rank in Class:	184	110	225	193	119	94	167
		Percentile:	28	29	3	9	39	41	31
		Z Score:	438	488	357	396	481	502	455
PROPOSAL, P 2 586000028	2	Raw Points:	21.08	4.2	4.8	3.0	2.4	4.2	3.5
		5 Point Score:	2.93	3.5	4.0	2.5	2.0	3.5	2.9
		Rank in Class:	183	110	48	193	156	94	167
		Percentile:	28	29	58	9	15	41	31
		Z Score:	442	488	557	396	430	502	455
PROPOSAL, P 4 586000044	2	Raw Points:	23.18	4.2	3.6	3.6	2.4	3.6	4.7
		5 Point Score:	3.22	3.5	3.0	3.0	2.0	3.0	3.9
		Rank in Class:	144	110	170	143	156	151	49
		Percentile:	43	29	12	25	15	14	78
		Z Score:	485	488	424	458	430	434	596
PROPOSAL, P 5 586000051	2	Raw Points:	22.60	4.2	3.6	4.2	3.0	3.6	4.0
		5 Point Score:	3.14	3.5	3.0	3.5	2.5	3.0	3.3
		Rank in Class:	154	110	170	82	119	151	111
		Percentile:	39	29	12	44	39	14	55
		Z Score:	473	488	424	519	481	434	514
PROPOSAL, P 6 586000069	2	Raw Points:	25.00	4.2	4.2	4.2	3.6	4.8	4.0
		5 Point Score:	3.47	3.5	3.5	3.5	3.0	4.0	3.3
		Rank in Class:	108	110	108	82	69	47	111
		Percentile:	57	29	33	44	53	63	55
		Z Score:	522	488	490	519	533	569	514
PROPOSAL, P 7 586000077	2	Raw Points:	30.10	4.8	5.4	5.4	5.4	4.2	4.9
		5 Point Score:	4.18	4.0	4.5	4.5	4.5	3.5	4.1
		Rank in Class:	31	60	16	5	4	94	27
		Percentile:	86	57	81	83	92	41	87
		Z Score:	627	551	624	643	688	502	620
PROPOSAL, P 10 586000101	2	Raw Points:	25.80	4.8	4.2	4.2	3.0	4.8	4.8
		5 Point Score:	3.58	4.0	3.5	3.5	2.5	4.0	4.0
		Rank in Class:	92	60	108	82	119	47	37
		Percentile:	64	57	33	44	39	63	82
		Z Score:	539	551	490	519	481	569	608
PROPOSAL, P 11 586000119	2	Raw Points:	23.18	4.2	4.2	3.6	3.6	3.6	4.0
		5 Point Score:	3.22	3.5	3.5	3.0	3.0	3.0	3.3
		Rank in Class:	144	110	108	143	69	151	116
		Percentile:	43	29	33	25	53	14	53
		Z Score:	485	488	490	458	533	434	514

Figure 2. Scores in various units of measure on each abstract.

Figure 3. Rank order listing of abstracts.

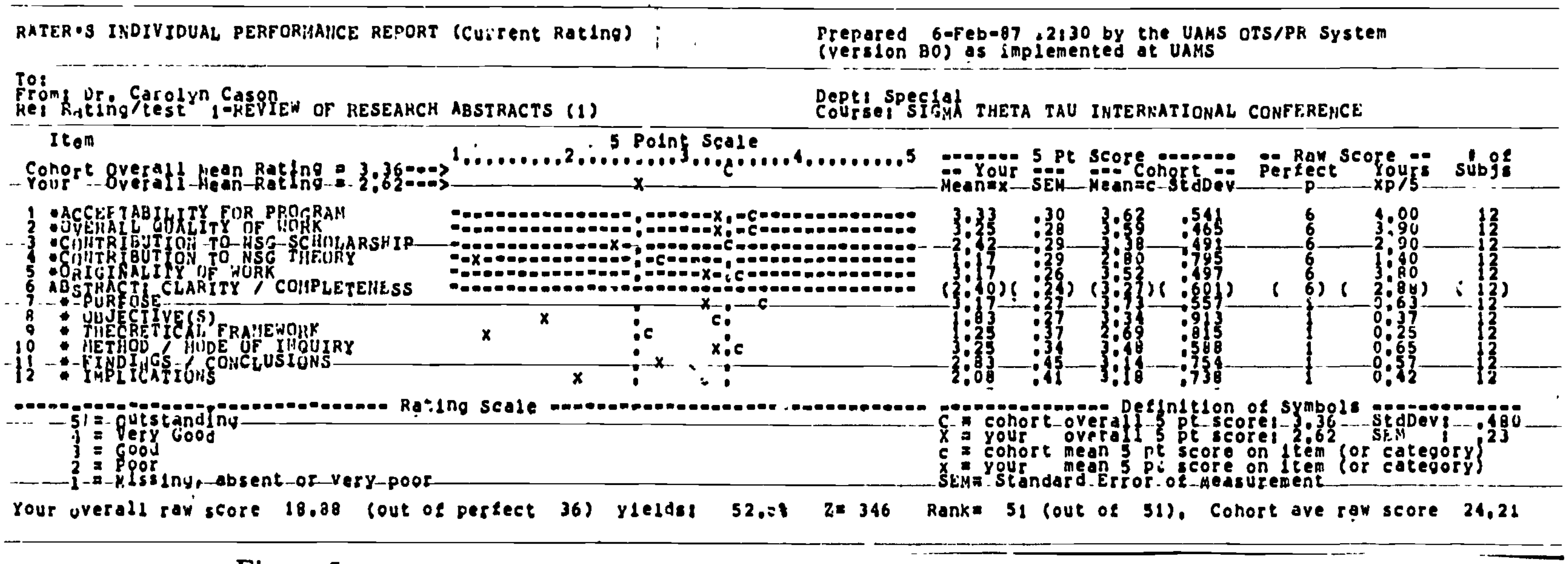


```
STUDENTS BY RANK (Current Rating)
                                                              Prepared  6-Feb-87 12:21 by th
                                                              (version 80; as implemented at
Test:        1 - REVIEW OF RESEARCH ABSTRACTS (1)             Dept: Special
Instructor: Dr. Carolyn Cason                                 Slot: 529      Phone:5163
Course:   .  SIGMA.THETA TAU INTERNATIONAL CONFERENCE         Subjects rated:256   Absent:
---------------------------------------------------------------------------------------------

                     RAW      5 PT       RANK         Z      # OF
                    SCORE     SCORE   RAW PCNTL     SCORE   RATERS             NAME

         Averages   23,92     3,32                   500      2,0

  1    586003972    35.20     4.89     1   99        732       2        PROPOSAL, P397
  2    586001265    34.50     4.79     2   99        717       2        PROPOSAL, P126
  3    586003105    34.20     4.75     3   98        711       2        PROPOSAL, P310
  4    586002800    33.50     4.65     4   98        697       2        PROPOSAL, P280
  5    586003287    33.50     4.65     5   98        697       2        PROPOSAL, P320
  6    586002040    33.30     4.62     6   97        693       2        PROPOSAL, P204
  7    586002792    33.12     4.60     7   97        689       2        PROPOSAL, P279
  8    586001653    33.00     4.58     8   96        687       2        PROPOSAL, P165
  9    586003394    32.80     4.56     9   96        682       2        PROPOSAL, P339
 10    586001455    32.50     4.51    10   95        676       2        PROPOSAL, P145
 11    586002412    32.50     4.51    10   95        676       2        PROPOSAL, P241
 12    586005910    32.35     4.49    12   95        673       2        PROPOSAL, P591
 13    586001208    32.00     4.44    13   94        666       2        PROPOSAL, P120
 14    586000747    31.80     4.42    14   94        662       2        PROPOSAL, P 74
 15    586000358    31.40     4.36    15   94        654       2        PROPOSAL, P 35
 16    586002073    31.32     4.35    16   93        652       2        PROPOSAL, P207
 17    586001281    31.30     4.35    17   93        652       2        PROPOSAL, P128
 18    586003261    31.20     4.33    18   92        649       2        PROPOSAL, P326
 19    586002578    31.00     4.31    19   92        645       2        PROPOSAL, P257
 20    586003634    31.00     4.31    19   92        645       2        PROPOSAL, P363
 21    586002305    30.96     4.30    21   91        645       2        PROPOSAL, P230
 22    586001380    30.77     4.27    22   91        641       2        PROPOSAL, P138
 23    586001257    30.70     4.26    23   91        639       2        PROPOSAL, P125
 24    586001885    30.60     4.25    24   90        637       2        PROPOSAL, P188
 25    586001927    30.60     4.25    24   90        637       2        PROPOSAL, P192
 26    586003345    30.40     4.22    26   89        633       2        PROPOSAL, P334
 27    586003865    30.40     4.22    26   89        633       2        PROPOSAL, P386
 28    586002768    30.36     4.22    28   89        632       2        PROPOSAL, P276
 29    586003758    30.20     4.19    29   88        629       2        PROPOSAL, P375
 30    586000960    30.16     4.19    30   88        628       2        PROPOSAL, P 96
 31    586000077    30.10     4.18    31   86        627       2        PROPOSAL, P  7
 32    586001224    30.10     4.18    31   86        627       2        PROPOSAL, P122
 33    586001596    30.10     4.18    31   86        627       2        PROPOSAL, P159
 34    586002313    30.10     4.18    31   86        627       2        PROPOSAL, P231
 35    586000374    29.60     4.11    35   85        617       2        PROPOSAL, P 37
 36    586003188    29.60     4.11    35   85        617       2        PROPOSAL, P318
 37    586003428    29.60     4.11    35   85        617       2        PROPOSAL, P342
 38    586005993    29.55     4.10    38   85        616       2        PROPOSAL, P599
 39    586001182    29.40     4.08    39   84        613       2        PROPOSAL, P118
 40    586001562    29.40     4.08    39   84        613       2        PROPOSAL, P156
 41    586002610    29.20     4.06    41   83        608       2        PROPOSAL, P261
 42    586000143    29.00     4.03    42   83        601       2        PROPOSAL, P 14
 43    586001547    28.90     4.01    43   83        602       2        PROPOSAL, P154
 44    586001984    28.88     4.01    44   82        602       2        PROPOSAL, P198
 45    586004525    28.48     3.96    45   82        594       2        PROPOSAL, P452
```



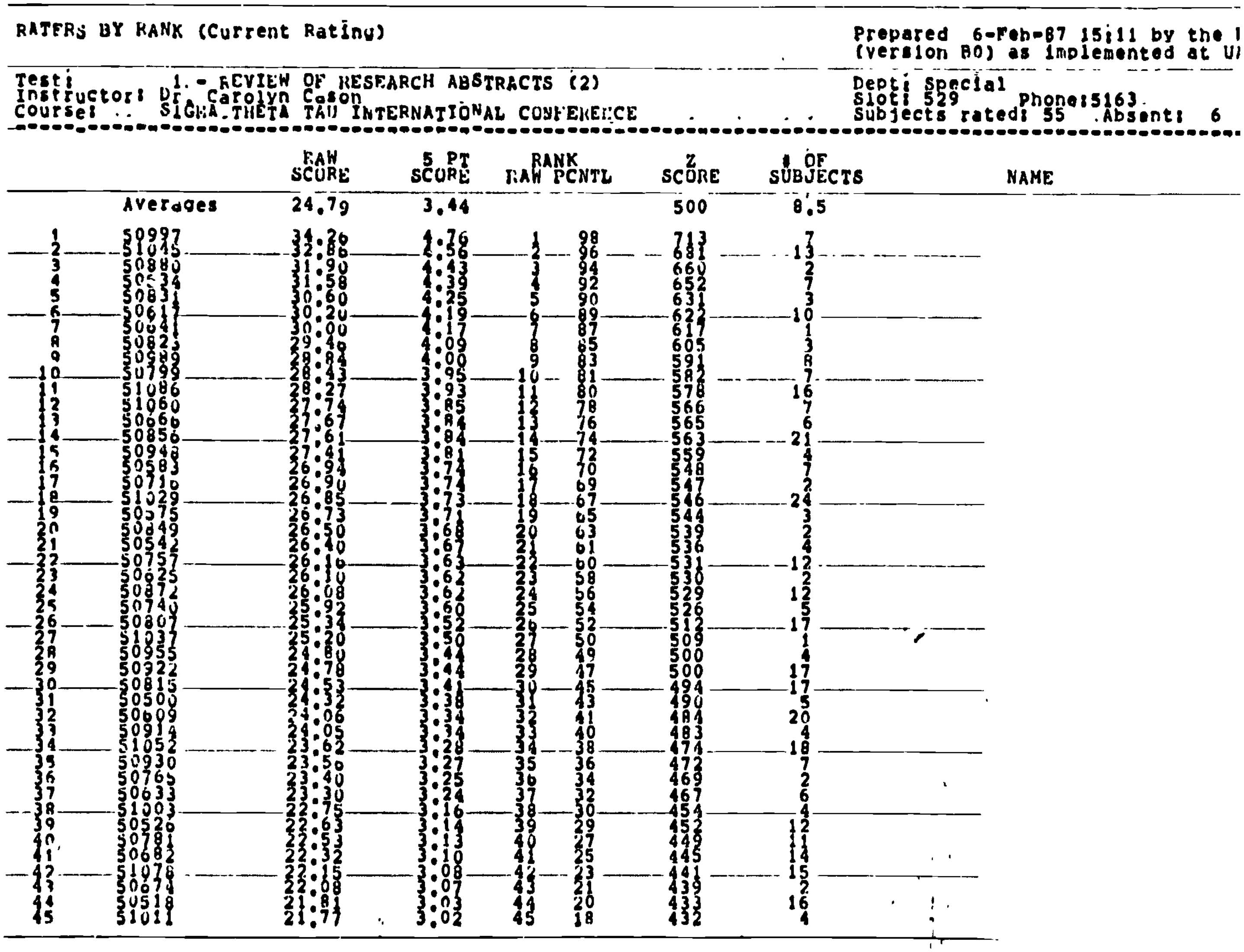
```
INDIVIDUAL PERFORMANCE REPORT (Current Rating)                    Prepared  6-Feb-87 12:25 by the UAMS OTS/PR System
                                                                  (version B0) as implemented at UAMS
To: P236 PROPOSAL
From: Dr. Carolyn Cason                                           Dept: Special
Re: Rating/test  1-REVIEW OF RESEARCH ABSTRACTS (1)               Course: SIGMA THETA TAU INTERNATIONAL CONFERENCE
   Item                                    5 Point Scale
   Class Overall Mean Rating = 3.32-->1.........2..........3..C......4.........5  ------- 5 Pt Score -------  -- Raw Score --   # of
   Your  Overall Mean Rating = 3.29-->                        X                   -- Your ---  --- Class ---  Perfect   Yours   Raters
                                                                                  Mean=x  SEM  Mean=c StdDev     p      xp/5
 1  *ACCEPTABILITY FOR PROGRAM         -----------------------,-XC-------------     3.50   .79   3.60   .786       6      4.20      2
 2  *OVERALL QUALITY OF WORK           -----------------------,--C---X---------     4.00   .75   3.57   .750       6      4.80      2
 3  *CONTRIBUTION TO NSG SCHOLARSHIP   -----------------------C-X--------------     3.50   .81   3.34   .813       6      4.20      2
 4  *CONTRIBUTION TO NSG THEORY        ------------------C--X--,---------------     3.00   .97   2.68   .968       6      3.60      2
 5  *ORIGINALITY OF WORK               --------------------X--,-C--------------     3.00   .74   3.49   .740       6      3.60      2
 6  ABSTRACT: CLARITY / COMPLETENESS   ----------------------------------------    (2.75)( .71) (3.24)( .708)   (  6) (  3.30)  (  2)
 7   * PURPOSE                                                 ,---C- X             4.00   .77   3.70   .765       1      0.80      2
 8   * OBJECTIVE(S)                                            C      X             4.00  1.00   3.30   .997       1      0.80      2
 9   * THEORETICAL FRAMEWORK                    X     C        ,                    2.00  1.16   2.56  1.158       1      0.40      2
10   * METHOD / MODE OF INQUIRY                           X    , C                  3.00   .80   3.47   .799       1      0.60      2
11   * FINDINGS / CONCLUSIONS          X                    C--,                    1.00  1.12   3.08  1.117       1      0.20      2
12   * IMPLICATIONS                                 X       C,                      2.50  1.02   3.20  1.024       1      0.50      2
-------------------------------- Rating Scale -------------------------------- --------------- Definition of Symbols -------------
      5 = Outstanding                                                          C = class overall 5 pt score: 3.32   StdDev: .676
      4 = Very Good                                                            X = your  overall 5 pt score: 3.29   SEM   : .68
      3 = Good                                                                 c = class mean 5 pt score on item (or category)
      2 = Poor                                                                 x = your  mean 5 pt score on item (or category)
      1 = Missing, absent or very poor                                         SEM= Standard Error of Measurement
Your overall raw score  23.70  (out of perfect  36)  yields:  65.8%  Z= 495  Rank= 127 (out of 256).  Class ave raw score  23.92
```

Figure 4. Individual (abstract) performance report.



```
RATER'S INDIVIDUAL PERFORMANCE REPORT (Current Rating)            Prepared  6-Feb-87 12:30 by the UAMS OTS/PR System
                                                                  (version B0) as implemented at UAMS
To:
From: Dr. Carolyn Cason                                           Dept: Special
Re: Rating/test  1-REVIEW OF RESEARCH ABSTRACTS (1)               Course: SIGMA THETA TAU INTERNATIONAL CONFERENCE
   Item                                     5 Point Scale
 Cohort Overall Mean Rating = 3.36--->1.........2..........3...C......4.........5  ------- 5 Pt Score -------  -- Raw Score --   # of
 Your   Overall Mean Rating = 2.62--->                   X                         -- Your ---  --- Cohort --  Perfect   Yours   Subjs
                                                                                   Mean=x  SEM  Mean=c StdDev     p      xp/5
 1  *ACCEPTABILITY FOR PROGRAM         ----------------,------X,-C--------------     3.33   .30   3.62   .541       6      4.00     12
 2  *OVERALL QUALITY OF WORK           ----------------,------X,-C--------------     3.25   .28   3.59   .465       6      3.90     12
 3  *CONTRIBUTION TO NSG SCHOLARSHIP   --------------X-,-------C----------------     2.42   .29   3.38   .491       6      2.90     12
 4  *CONTRIBUTION TO NSG THEORY        -X--------------,-C-----,----------------     1.17   .29   2.80   .795       6      1.40     12
 5  *ORIGINALITY OF WORK               ----------------,-----X-,C---------------     3.17   .26   3.52   .497       6      3.80     12
 6  ABSTRACT: CLARITY / COMPLETENESS   -----------------------------------------    (2.40)( .24) (3.27)( .601)   (  6) (  2.88)  ( 12)
 7   * PURPOSE                                         X-,--C                        3.17   .27   3.73   .557       1      0.63     12
 8   * OBJECTIVE(S)                          X      ,    C,                          1.83   .27   3.34   .913       1      0.37     12
 9   * THEORETICAL FRAMEWORK             X          ,C                               1.25   .37   2.69   .815       1      0.25     12
10   * METHOD / MODE OF INQUIRY                     ,    X,C                         3.25   .34   3.48   .588       1      0.65     12
11   * FINDINGS / CONCLUSIONS                       ,-X-   ,                         2.83   .45   3.14   .754       1      0.57     12
12   * IMPLICATIONS                            X    ,      ,                         2.08   .41   3.18   .738       1      0.42     12
-------------------------------- Rating Scale -------------------------------- --------------- Definition of Symbols -------------
      5 = Outstanding                                                          C = cohort overall 5 pt score: 3.36   StdDev: .480
      4 = Very Good                                                            X = your   overall 5 pt score: 2.62   SEM   : .23
      3 = Good                                                                 c = cohort mean 5 pt score on item (or category)
      2 = Poor                                                                 x = your   mean 5 pt score on item (or category)
      1 = Missing, absent or very poor                                         SEM= Standard Error of Measurement
Your overall raw score  18.88  (out of perfect  36)  yields:  52.?%  Z= 346  Rank=  51 (out of  51).  Cohort ave raw score  24.21
```

Figure 5. Individual (reviewer) performance report.



```
RATERS BY RANK (Current Rating)                                    Prepared  6-Feb-87 15:11 by the I
                                                                   (version B0) as implemented at U/
Test:          1. - REVIEW OF RESEARCH ABSTRACTS (2)               Dept: Special
Instructor: Dr. Carolyn Cason                                      Slot: 529      Phone:5163
Course: ..  SIGMA THETA TAU INTERNATIONAL CONFERENCE   .   .  . .  Subjects rated: 55   Absent:  6
-----------------------------------------------------------------------------------------------------

                          RAW        5 PT       RANK          Z       # OF
                         SCORE      SCORE    RAW PCNTL      SCORE   SUBJECTS          NAME
-----------------------------------------------------------------------------------------------------
            Averages     24.79      3.44                    500       8.5

     1      50997        34.26      4.76      1    98       713       7
     2      51045        32.86      4.56      2    96       681      13
     3      50880        31.90      4.43      3    94       660       2
     4      50534        31.58      4.39      4    92       652       7
     5      50831        30.60      4.25      5    90       631       3
     6      50617        30.20      4.19      6    89       622      10
     7      50641        30.00      4.17      7    87       617       1
     8      50823        29.46      4.09      8    85       605       3
     9      50989        28.84      4.00      9    83       591       8
    10      50799        28.43      3.95     10    81       582       7
    11      51086        28.27      3.93     11    80       578      16
    12      51060        27.74      3.85     12    78       566       7
    13      50666        27.67      3.84     13    76       565       6
    14      50856        27.61      3.84     14    74       563      21
    15      50948        27.41      3.81     15    72       559       4
    16      50583        26.94      3.74     16    70       548       7
    17      50716        26.90      3.74     17    69       547       2
    18      51029        26.85      3.73     18    67       546      24
    19      50575        26.73      3.71     19    65       544       3
    20      50849        26.50      3.68     20    63       539       2
    21      50542        26.40      3.67     21    61       536       4
    22      50757        26.16      3.63     22    60       531      12
    23      50625        26.10      3.62     23    58       530       2
    24      50872        26.08      3.62     24    56       529      12
    25      50740        25.92      3.60     25    54       526       5
    26      50807        25.34      3.52     26    52       512      17
    27      51037        25.20      3.50     27    50       509       1
    28      50955        24.80      3.44     28    49       500       4
    29      50922        24.78      3.44     29    47       500      17
    30      50815        24.53      3.41     30    45       494      17
    31      50500        24.32      3.38     31    43       490       5
    32      50609        24.06      3.34     32    41       484      20
    33      50914        24.05      3.34     33    40       483       4
    34      51052        23.62      3.28     34    38       474      18
    35      50930        23.56      3.27     35    36       472       7
    36      50765        23.40      3.25     36    34       469       2
    37      50633        23.30      3.24     37    32       467       6
    38      51003        22.75      3.16     38    30       454       4
    39      50526        22.63      3.14     39    29       452      12
    40      50781        22.53      3.13     40    27       449      11
    41      50682        22.32      3.10     41    25       445      14
    42      51078        22.15      3.08     42    23       441      15
    43      50674        22.08      3.07     43    21       439       2
    44      50518        21.81      3.03     44    20       433      16
    45      51011        21.77      3.02     45    18       432       4
-----------------------------------------------------------------------------------------------------
```

Figure 6. Rank order listing of reviewers.

Results

The observed data analyses completed using the OTS-PR system yielded a variety of reports including (a) information about the inventory, (b) information about the abstracts, and, (c) information about the raters/reviewers. Reports on abstracts included: total and category (criterion) scores for each abstract (in various units of measure), illustrated in Figure 2; a rank order listing of abstracts including total scores and number of raters, illustrated in Figure 3; and, individual abstract performance reports which as illustrated in Figure 4 both graphically and numerically depict the subject's performance relative to group performance. Reports on reviewers included: the number of subjects/abstracts rated; the average ratings given across these subjects (total and category (criterion) scores in various units); individual rater reports which as illustrated in Figure 5 depict both graphically and numerically the rater's performance relative to group performance; and, a rank order listing of the raters in terms of these ratings, illustrated in Figure 6.

In all of these reports generated by the OTS-PR system, average ratings are computed as a simple arithmetic mean of observed scores; the computational procedure most commonly used to obtain a summary score. These scores and this computational approach assume that the standards of the reviewers/raters are highly similar or at least that differences in reviewers will be "balanced out" where an average across reviewers is used; an assumption that most individuals who have had their performance or products evaluated by different persons have sometimes found to be unwarranted. That the standards of the reviewers who provided ratings of research abstracts were different is suggested in at least two of the reports which OTS-PR produces: the rater performance report (Figure 5) and the rater rank order report (Figure 6). The rater/reviewer performance report is intended to give the rater information about the standards he or she used relative to other raters who rated subjects/abstracts. It was developed as a means of providing feedback to individual raters much as the individual performance report provides feedback to the subject about his performance. As can be seen from Figure 5, the average ratings given by this reviewer to the proposals he/she rated departed from what the average of the (average) ratings given by all reviewers. If the average quality of the abstracts rated by each reviewer were the same (as would be expected because of random assignment of abstracts to reviewers), then, any substantial variation in average ratings given by different reviewers would reflect differences in standards (i.e., stringency).

The rater rank order report (Figure 6) also suggests that the standards used by the reviewers who reviewed research abstracts differed. This report provides information only about the relative mean observed ratings of the reviewers, i.e., those at the top of the figure having assigned higher ratings to their abstracts than those raters at the bottom. The presence of such differences in the mean observed ratings and implied differences in standards/stringencies of reviewers makes the determination of the true quality of abstracts more difficult: how much of the total or average score is a function of true abstract quality and how much it is a reflection of who happened to review the abstract (and their standards/stringency relative to the pool of potential reviewers) remains obscure.

As can be seen in Table 1, application of our simplified model to the data yielded quite good fit ($R > .89$). There were significant differences ($p < .0001$) in rater standards in each of the data sets. For details on the way in which these effects were tested, see the description of the statistical models provided in Cason and Cason (1984). Table 1 also shows the relative contribution of reviewer stringency, proposal quality, and random error to the variance in observed ratings in each of the three data sets and for all data sets considered together. Components of variance in Table 1 were estimated as a sum of the products of the respective standardized weights (Beta_j) and correlations (r_{jy}) between predictor variables (one binary vector per reviewer and one per abstract) and the criterion in the regression equation.

(Equation 1)

$$\text{Proportion of Variance} = (\text{Beta}_i * r_{iy})$$

where i = 1 to n abstracts; or, 1 to k reviewers.

The summation of products is across the set of either reviewers or abstracts (Hays, 1963). Across all three data sets, differences in reviewer stringency accounted for very nearly as much variance (.404) in the observed raw ratings as did true abstract quality (.407).

Table 1
Fit of Cason and Cason's Model to Research Abstracts
Submitted for the Edinburgh Conference

	All	Data Subsets		
	Data	AYE	BEE	CEE
Multiple R	.899	.901	.905	.894
Components of Variance				
Reviewers	.404	.401	.397	.416
Abstracts	.407	.410	.423	.383
Error	.192	.189	.201	.181
Number of				
Reviewers	61	50	48	49
Abstracts	503	176	162	168
Observations	972	344	297	331

Note: All Rs are significant at $p < .00001$. Rater effects were significant at $p < .0001$.

The origin of the stringency/ability scale was set in each analysis by assigning an arbitrarily chosen reviewer a stringency of 500. This produced apparent differences in both mean reviewer stringency and mean abstract quality for each data set. These differences are shown in Table 2 and are labeled as "preliminary" means. Because abstracts were randomly assigned to the three groups, it was reasonable to assume that mean abstract quality was equal across groups. There being far more abstracts than reviewers the sampling error of the mean for abstract quality was smaller (as shown in Table 2). Therefore, mean abstract quality was better suited as a basis for calibrating the results of the analyses on the three subsets of data.

Table 2
Means and Standard Errors of Model Parameters
in Each of the Three Data Sets

	AYE	BEE	CEE
Reviewer Stringencies			
Preliminary Mean	518	482	468
Calibrated Mean	518	508	503
Standard Error	9	12	14
Abstract Quality			
Preliminary Mean	570.14	543.89	535.30
Calibrated Mean	570.14	570.14	570.14
Standard Error	4.20	5.90	4.50

Calibrated values from the separate analyses were obtained by adding a constant to each of the stringency and ability parameters obtained in data set BEE (26.25) and CEE (34.85) such that the mean abstract quality for each group equaled the mean abstract quality found in group AYE. This may be understood as moving the origin (zero point) of the scale for groups BEE and CEE 26 and 35 points respectively while not altering the distances between the reviewer stringencies or abstract qualities within each group. The effect on the means is shown in Table 2 labeled "calibrated". Note that there are small differences between reviewer means, much smaller than before calibration. However, after calibration the differences in mean stringencies is well within expected sampling fluctuations, as would be expected because abstracts were randomly assigned to reviewers.

In principle the abstract quality parameter values could be directly used for the selection of abstracts. They do represent the best estimate of each abstract's quality independent of the standards (stringency) of the particular reviewers that rated given abstracts. However, they are on an unfamiliar scale and not easily interpretable in terms of the definitions on the original rating inventory. Therefore a calibrated (adjusted) rating was computed for each abstract that was in percent units on the original rating scale. The calibrated rating was computed as the mean of the expected ratings an abstract would have received had all the reviewers (in all the sub-groups) rated all the abstracts. For a given abstract, its calibrated abstract quality parameter value and the stringency parameter values of all raters in all groups were used to obtained expected ratings; then, the mean of these was taken as the calibrated (adjusted) rating for that abstract. Program NULOCS was used to generate the calibrated parameters and then the calibrated ratings. Table 3 shows that this approach achieved highly similar means and standard errors in calibrated ratings across the three sets as would be expected from the assumed equal mean quality of abstracts arising from random assignment.

Table 3
Mean, Standard Error, and Standard Deviation of Calibrated Ratings

	AYE	BEE	CEE
Mean	66.40	65.50	66.00
Standard Error	1.17	1.38	1.17
Standard Deviation	15.50	17.51	15.20

According to Hays (1963, p. 424), the intra-class correlation (r_{ic}) is a function of the variance attributable to an effect (σ_a) as a proportion of total variance.

(Equation 2)

$$r_{ic} = \sigma_a / (\sigma_a + \sigma_e)$$

The proportion of variance attributable to proposal quality in Table 1 can thus be interpreted as the intra-class correlation of reviewers with respect to their observed ratings of abstracts. As Hays points out, this is equivalent to the reliability of a single reviewer's observed rating. Alternatively, this value may be interpreted as the expected correlation between the ratings given by randomly chosen pairs of reviewers. The reliability of a mean of several reviewers' ratings, as is available in these data (where number of reviewers = k), is given by the Spearman-Brown expansion formula:

(Equation 3)

$$r_k = (k * r)/(1 + ((k - 1) * r))$$

where r = the reliability of a unit length measure, in this case a single reviewer; and,
= number of reviewers.

Table 4 shows the impact of calibrating ratings on the reliability of both a single reviewer and aggregate ratings calculated from 2 reviewers. The reliabilities for the single reviewer calibrated ratings were obtained by including only the sum of the random error and abstract variances in the denominator in Equation 2. The reliabilities of the observed ratings must include the variance associated with reviewers in addition to that associated with proposals and error (Ebel, 1951). As can be seen from Table 4, the reliability of calibrated ratings from a single reviewer (.68) is substantially higher than the observed rating from a single reviewer (.41). A small percentage of the abstracts were indeed reviewed by only a single reviewer and for those the single reviewer reliabilities are most accurate. However, as the vast majority of abstracts were reviewed by two reviewers, in general the two reviewer reliability better represents the overall reliability of the review process. The two reviewer reliability for observed ratings is so low (i.e., r < .60) that a great deal of error would arise were observed ratings used as the basis for selection of papers for the program. While imperfect, the two reviewer reliability for calibrated ratings (.81) indicates that these ratings provide a good basis for selecting papers for the program.

Table 4
Reliability of Ratings
Intra-Class Correlations

	Single Reviewer		Aggregate of Reviewers	
	k = 1		k = 2	
	Observed	Calibrated	Observed	Calibrated
All Data	.407	.680	.579	.810
Subset				
AYE	.410	.685	.582	.813
BEE	.423	.678	.595	.808
CEE	.383	.679	.554	.809

Although consistency among reviewers, represented as an intra-class correlation (Ebel, 1951; Stanley, 1961), is frequently interpreted as a measure of reliability, it may also be interpreted as a measure of validity. Stanley (1961) observed that each reviewer may be considered a different method of measuring a given construct (e.g., abstract quality). Therefore, the single rater reliabilities (intra-class correlations) reported in Table 4 may be equally well interpreted as both single rater reliability coefficients and single rater validity coefficients. However, validity (Equation 4) does not expand as rapidly as does reliability (Equation 3) with increased numbers of independent observations. (Gulliksen, 1950).

(Equation 4)

$$r_{xy,k} = (r_{xy} * (k^{1/2})/(1 + ((k-1) * r_{xx})^{1/2})$$

where $r_{xy,k}$ is the validity based on k independent raters;
r_{xy} is the validity of a single rater;
r_{xx} is the reliability of a single rater; and,
k is the number of independent reviewers/ratings.

Table 5 reports the validity of ratings from a single reviewer and the aggregate of ratings from two reviewers as measures of abstract quality. As discussed above, the validities associated with a single reviewer's observed and calibrated ratings are in this special case equal to the corresponding reliabilities associated with a single reviewer's observed and calibrated ratings reported in Table 4. As with reliability, a non-trivial improvement in convergent construct validity was obtained by calibrated ratings when contrasted with observed ratings.

Given these results, the calibrated ratings offer a more reliable and valid basis for making decisions about disposition of the research abstracts. The work of the Abstract Selection Committee was facilitated by sorting these caliabrated ratings into descending rank order and printing them along with their abstract identification numbers and mean observed ratings. This list as well as Tables 1 and 4 and the reports from OTS-PR were forwarded to members of the Abstract Selection Committee. All decisions about abstracts to be included in the program were made by the Abstract Selection Committee.

Table 5
Validity of Ratings

	Single Reviewer k = 1		Aggregate of Reviewers k = 2	
	Observed	Calibrated	Observed	Calibrated
All Data	.407	.680	.485	.742
Subset				
AYE	.410	.685	.488	.746
BEE	.423	.678	.501	.740
CEE	.383	.679	.461	.741

When the 503 research abstracts were examined by the Abstract Selection Committee, they found that of these only 500 were unique (3 abstracts had been given two identification numbers and were reveiwed by two sets of two reviewers) and that another nine should be considered as congress related topics. Thus, there were 491 research abstracts considered for inclusion in the program. Of these, the Abstract Selection Committee selected 302 for inclusion in the program (94 as paper presentations and 208 as poster presentations).

One way of depicting the impact of using calibrated rather than mean observed ratings in making the program selection decision is shown in Table 6. Using the simplified decision rule of accepting the top 302 rated research abstracts regardless of any other considerations would have resulted in 64 being accepted under one measure and rejected under the other. If abstract selection had been based only on the judged quality of an abstract (the task completed by each reviewer on each abstract which they reviewed), then use of the calibrated ratings rather than the mean observed ratings could have produced as great as a 21% difference in the specific abstracts selected for the program.

Table 6
Transitions in Selection Outcome Resulting from
Using Calibrated or Mean Observed Ratings

	Outcome Based on Calibrated Rating		
Outcome Based on Observed Rating	Select	Reject	Total
Select	238	64	302
Reject	64	125	189
Total	302	189	491

However, reviewers provide only one level of evaluative information. They are asked to judge only the abstract under consideration in terms of its quality. They make these judgements independent of such other considerations as comprehensiveness or representativeness of the final program. These other considerations in abstract selection are evident when one examines the correlations between disposition and mean observed ratings (r = .47; N = 491) and calibrated ratings (r = .69; N = 491). The Abstract Selection Committee used a combination of decision rules in making decisions about the inclusion or exclusion of abstracts in the program including (a) selecting the top rated abstracts, (b) selecting only a single abstract from an author with multiple submissions, (c) selecting

Table 7
Research Abstracts Selected or Not for Inclusion in the Program
Means, Standard Deviations, and Ranges of Observed and Calibrated Ratings

	Observed Ratings	Calibrated Ratings
Selected for Inclusion		
N = 302		
Mean	72.9	74.9
Standard Deviation	11.9	10.0
Minimum-Maximum	44.1-98.3	54.0-99.7
Not Selected for Inclusion		
N = 189		
Mean	59.3	52.9
Standard Deviation	13.4	12.9
Minimum-Maximum	28.3-95.0	7.1-84.9

abstracts so as to achieve representativeness of participating countries. The use of such decision rules in making decisions about the program is also reflected in the means and standard deviations shown in Table 7. Those abstracts selected for inclusion in the program had mean observed and mean calibrated ratings well above that of those abstracts not included. But, there was overlap in the range of ratings of abstracts accepted and not. The range of the calibrated ratings of abstracts selected for inclusion was from 54 to 99.7 while for those not selected it was 7.1 to 84.9. However, of those not selected for inclusion all with calibrated ratings of greater than 61.8 were authored by individuals who had multiple submissions and had had one of the other research abstracts selected for inclusion. Excluding the ratings of other abstracts of authors having one abstract selected, there was only a small overlap in the maximum calibrated rating of an abstract that was rejected (61.8) and the minimum of one selected (54.0). This small overlap results from the other factors, e.g., balance relative to countries, that were considered.

Discussion and Conclusions

Application of our simplified model to these data revealed that about half (.40) of the total (.81) variance accounted for was attributable to differences among reviewers and that those differences in reviewer stringency were statistically significant. These reviewer effects were stronger than those observed in previous research (Cason, Cason, & Stritter, 1986a) where reviewer effects accounted for only .117, .189, and .144 of the variance. The proportion of the variance attributable to abstract quality in this study (.407) was highly similar to that found in the earlier study (.459, .393, and .415).

Table 8
Single Reviewer Reliability and Validity
for Mean Observed and Calibrated Ratings

	Reliability & Validity	
	Observed Rating	Calibrated Rating
Sigma Theta Tau	.407	.680
Cason et al (1986)		
AERA 1983	.459	.520
AERA 1985	.393	.485
AERA 1986	.415	.485
Marsh & Ball (1981)	.340	.350

Calibration of ratings, i.e., removing the large and significant reviewer effects, yielded ratings of research abstracts more reflective of their true quality. These results are also consistent with those reported previously. As shown in Table 8, the single reviewer reliabilities and validities for observed ratings are about the same as those found by Marsh and Ball (1981) and Cason, Cason, and Stritter (1986a). Calibration of ratings improved the single reviewer reliabilities and validities in all cases. In the Sigma Theta Tau data, the improvement in reliability and validity was noticably greater than in either of the other data sets. This was a direct result of a larger component of variance being associated with systematic rater bias, i.e., reviewer stringencies, in the Sigma Theta Tau data.

In each of these studies abstracts/proposals/manuscripts were reviewed by more than a single reviewer. Since multiple reviewers were used in each case, the reliabilities and validities of the peer review process are those reported for the aggregate of reviewers. These reliabilities and validities are shown in Table 9. As shown in Table 9, the mean observed ratings from the Sigma Theta Tau data had the lowest reliability and the calibrated ratings the second highest of the three studies. Thus application of the model to the Sigma Theta Tau data obtained the largest improvement in reliability of the overall process even though only two reviewers were used as compared with four reviewers in the Cason et al study. The pattern of results for validity is similar. The validity of the observed mean ratings for the Sigma Theta Tau data was lowest. However, calibration of Sigma Theta Tau ratings yielded much greater improvements than those found in the other studies even though only two reviewers reviewed each abstract. Had mean observed ratings been used for selection, the Sigma Theta Tau review process would have been the weakest (i.e., least reliable and valid) of these studies. Using the calibrated ratings probably resulted in the Sigma Theta Tau review process being the most valid among these studies.

Table 9
Reliability and Validity of the Peer Review Process

	Number	Reliability		Validity	
	of	Observed	Calibrated	Observed	Calibrated
Study	Reviewers	Rating	Rating	Rating	Rating
Sigma Theta Tau	2	.579	.810	.485	.742
Cason et al (1986)					
AERA 1983	4	.768	.813	.595	.650
AERA 1985	4	.722	.790	.532	.619
AERA 1986	4	.739	.790	.554	.619
Marsh & Ball (1981)	3*	.670	.683	.509	.522

*Includes journal editor.

The availability of calibrated ratings to the Committee greatly eased the Committee's task: they no longer had to deal with ratings which were confounded by variation in reviewer stringency (as occurs in the mean observed ratings) Abstract selection could be based on ratings that more accurately reflected the quality of the abstract without regard to who happened to have reviewed it. Use by the Abstract Selection Committee of these calibrated ratings in making selection decisions greatly enhanced both the reliability and validity of the peer review process.

References

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, 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., & Cason, C.L. (1987, July). Practical and theoretical requirements for controlling rater stringency in peer review. Presented at the International Research Congress sponspored by Sigma Theta Tau International Honor Society, the Royal College of London and the University of Edinburgh, Edinburgh, Scotland.

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

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Marsh, W.H., & Ball, S. (1981). Interjudgemental reliability of reviewers for the Journal of Educational Psychology. Journal of Educational Psychology, 73(6), 203-219.

Hays, W.L. (1963). Statistics. New York: Holt, Rinehardt, & Winston.

Stanley, J.C. (1961). Analysis of unreplicated three-way classifications with applications to rater bias and trait independence. Psychometrika, 26(2), 203-219.