DOCUMENT RESUME

ED 412 236 TM 027 466

AUTHOR O'Neill, Thomas R.; Lunz, Mary E.

TITLE A Method To Compare Rater Severity across Several Administrations.

PUB DATE 1997-03-00

NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Ability; *Comparative Analysis; Equated Scores; *Interrater Reliability; Item Response Theory; *Scoring; *Test Results; Testing

IDENTIFIERS *Rasch Model; *Rater Stringency Error

ABSTRACT

This paper illustrates a method to study rater severity across exam administrations. A multi-facet Rasch model defined the ratings as being dominated by four facets: examinee ability, rater severity, project difficulty, and task difficulty. Ten years of data from administrations of a histotechnology performance assessment were pooled and analyzed to establish a scale. Next, the 17 individual administrations were anchored to that scale and reanalyzed. The severity of the nine raters who graded most often were listed and plotted by administration. These plots show the consistency of each rater's level of severity. Results show that: (1) raters have an individual level of severity; (2) raters can usually maintain that level; and (3) some raters maintain their level of severity more consistently than others. The implications for equating performance assessments prospectively are discussed. (Contains 2 tables, 10 figures, and 5 references.) (Author/SLD)

A Method to Compare Rater Severity Across Several Administrations

Thomas R. O'Neill
and
Mary E. Lunz

American Society of Clinical Pathologists
2100 W. Harrison Street
Chicago, Illinois 60612

Paper presented at the Annual Meeting of the American Educational Research Association, Chicago. March 1997.

ABSTRACT

The purpose of this paper is to illustrate a method to study rater severity across exam administrations. A multi facet Rasch model defined the ratings as being dominated by four facets: examinee ability, rater severity, project difficulty, and task difficulty. Ten years of data were pooled and analyzed to establish a scale. Next, the 17 individual administrations were anchored to that scale and re-analyzed. The severity of the nine raters who graded most often were listed and plotted (± 2 SEs) by administration. These plots show the consistency of each rater's level of severity. The results show that (1) raters have an individual level of severity, (2) raters can usually maintain that level, and 3) some raters can more consistently maintain their level of severity than others. The implications for equating performance assessments prospectively are discussed.

Keywords: rater, severity, Rasch, IRT, performance assessment

Division D

A Method to Study Rater Severity Across Several Administrations

Thomas O'Neill and Mary Lunz

Introduction

Performance assessments are often thought to have greater validity than multiple choice tests because the actual performance of the task is rated. However, the reproducibility of examination results derived from performance assessments is sometimes questioned because the performances must be graded by raters who often have different individual standards of excellence. Therefore, any given rating will be influenced not only by the examinee's ability and the item's difficulty, but also by a third facet, rater severity. In order for examination results to be meaningful, differences in raters must be accounted for, so that all results are expressed from the same frame of reference. The extension of the Rasch (1960/1980) model to the Many Facet Rasch Model (MFRM, Linacre, 1989) has made accounting for rater severity possible, by placing rater severity in the same frame of reference as item difficulty and examinee ability. The MFRM estimates each rater's severity, each project's difficulty, and/or other such facets, and removes their influence before computing an examinee's ability. In this way, similar examination results are expected for the same examinee even when different raters and projects are used.

Recent MFRM research supports that raters are able to maintain a consistent degree of severity even when rating examinees of very different ability levels (Lunz, Stahl, & Wright, 1996; O'Neill & Lunz, 1996). When the goal is to carry forward the same scale, a linking strategy must be employed so that the severity of new raters is expressed in the same frame of reference as that of the original raters. Using common raters to link together two test administrations, requires that the raters maintain their same degree of severity in the second administration as in the first. For this reason, studies regarding the stability of rater severity across administrations are important. To this end, Lunz, Stahl, and Wright (1996) compared the severity of eleven raters across two test administrations that were six months apart. They found that, generally, raters do not change their severity across administrations.

This paper describes a retrospective method that can be applied to data spanning many administrations for the purpose of assessing rater stability over time. The retrospective multi-administration method is demonstrated on data from a histotechnology performance assessment that spans a ten year period. How to prospectively use the retrospective information is also discussed.

Administration-to-Administration Equating

As part of the equating process, rater stability is verified from administration to administration. This is done by comparing the severity of several common "anchor" raters on the current administration with their degree of severity from the prior administration and then checking that their severity on the current administration places them in the same relative position as in the past. If their relative positions hold, it is reasonable to conclude that their severity has not changed. In cases, where only one or two of the anchor raters have changed positions, it is reasonable to conclude that those one or two raters have changed their degree of severity and should be treated as new raters, but the rest of the anchor raters can be used to link the new raters to the established scale. Yet, if several raters change places and the number of anchor raters is few, it becomes more complicated to determine which of the anchor raters changed their severity and which remained the same. To prevent this from happening, psychometricians try to employ as many stable pre-calibrated raters as possible, so that any anomalous raters will stand out more clearly.

Multi-Administration Analysis

The retrospective multi-administration method begins by taking advantage of the larger number of ratings available by pooling the ratings from all administrations and computing ability measures for all examinees and computing calibrations for all of the projects (and other facets that represent agents of measurement). Because the number of ratings per project is greater for the pooled analysis than for any individual analysis, the pooled analysis produces the most precise calibrations for projects. Because examinees don't overlap administrations (examinees that overlap failed the first time and hopefully improved before retaking the test), the pooled analysis and individual administration analyses provided the same number of observations for each examinee. While the number of observations are the same, the examinee ability estimates from the pooled analysis will contain less error because they are based on the more precise project calibrations.

Next, each administration is re-analyzed individually, but with each examinee's ability and each project's difficulty set to its value from the pooled analysis. In this way, rater severity becomes the object of measurement and the test's other facets (examinee ability and project difficulty) are held constant. Next, the individual rater's severity estimates are collected for each administration, and then plotted (± 2SEs) to illustrate the intra-rater cross-administration changes.

Methods

The Rasch Model

The Rasch (1960/1980) model is a logistic latent trait model of probabilities which analyzes items and persons independently, and then expresses both the item difficulties and the person abilities on a single continuum. The Many Facet Rasch Model (MFRM) extends the Rasch model to account for other differences in context, such as particular items, projects, raters, tasks, session, etc., so that the results generalize beyond the specific occasion in which the data were collected. In this way, the actual examinee ability level is expressed so that the particular items or raters are of no importance.

In this study, the focus is the degree of rater severity across administrations. Because MFRM accounts for differences in the particular examinees or projects rated, it is possible to focus on changes in rater severity across administrations.

Examinees

The examinees were candidates for histotechnician certification. Each of the 4,683 examinees submitted a work-sample for evaluation. The subjects in this study represent the examinees from 17 different administrations that span ten years. The number of examinees per administration ranged from 168 to 385 with 275 examinees being the average. Failing examinees were permitted to submit another examination, but were treated as independent cases because hopefully the examinees had taken steps to improve their ability before retesting.

Raters

The raters were experts in their field. While only 11 to 20 raters were needed to grade any single administration, a total of 57 different raters were used over the ten years. Most rater's graded in more than one administration and on average, raters graded in 6 administrations. Prior to each grading session, the raters attended a three hour orientation session to re-familiarize them with the scoring criteria and the particular projects under consideration. The performance of the nine raters who graded in 10 or more administrations is reported in detail in this paper. Other raters showed comparable patterns, but graded in only one to nine administrations.

Instruments

The examination requires examinees to submit 15 projects (histology slides) made according to prespecified requirements for type of tissue and stain. The projects varied across administrations, but there was sufficient overlap to equate the different versions of the exam. Each project is rated on three different tasks, thus each exam is evaluated on the basis of 45 ratings. The three tasks and the rating scales for the three tasks remained the same during the period of this study. Ratings on the examination (see Figure 1) were modelled as being governed by four facets: (1) examinee ability, (2) project difficulty, (3) task difficulty, and (4) rater severity.

Procedures

FACETS (Linacre, 1994), a MFRM computer program, was used to calibrate candidates, raters, projects, and tasks. The initial pooled data analysis included all 17 administrations which established a benchmark scale. Because the number of observations per task and per project was greater for the pooled analysis than for any individual administration, the pooled analysis produced the most precise calibrations for projects and tasks. While the pooled analysis and individual administration analyses provided the same number of observations per examinee, the pooled analysis examinee ability estimates are more precise because project and task difficulty calibrations are estimated more precisely.

Next, the 17 administrations were then re-analyzed individually, but each examinee's ability, each project's difficulty, and each task's difficulty was set to its value from the pooled analysis. In this way, only rater severity calibrations were permitted to vary because the test's other facets (examinee ability, project difficulty, and task difficulty) were held constant. Thus, the examinees, projects, and tasks become agents used to assess the raters, so the raters are defined as the object of measurement. The rater severity calibrations from each administration were collected and summarized.

Analysis

First, the separation reliability for each facet from the pooled analysis was computed. Next, descriptive statistics for each exam administration were computed from the individual analyses. In order to get a clear picture of rater severity over time, only the raters who graded in ten or more administrations had their severity (± 2SEs) plotted across administrations. The raters who graded in fewer than 10 administrations (N=47) were calibrated, but their severities were not plotted for this paper.

Results

Pooled Analysis

The pooled analysis demonstrated that the test adequately discriminated among examinees (separation reliability = .81). The slide-projects were significantly different in difficulty (separation reliability = .99), and the severity estimates of raters were significantly different from each other (separation reliability = .97). The tasks were also very different in difficulty from each other (separation reliability > .995). The errors of measurement were very small due to the large number of observations used in the pooled analysis.

Individual Administration Analysis

Table 1 shows the mean project difficulties, rater severities, and examinee ability estimates for the pooled analysis and each administration. While there was some variability across administrations, the mean project difficulty and rater severity remained reasonably comparable, indicating that overall the administrations were of similar difficulty. Because the same tasks were used across all administrations, the task difficulty was identical. Examinee ability estimates also showed some variability across administrations, but overall the examinee pool is comparable.

Of the 57 raters, nine graded in at least ten administrations during this ten year period. The severity of these nine raters is listed in Table 2 and was plotted (± 2SEs) across administrations (Figures 2 through 10). Upon inspection of these plots, three things can be clearly seen. First, it is clear that different raters have different levels of severity. Second, raters are usually able to maintain a self-consistent level of severity across administrations. Third, some raters are more consistent than others over time.

A comparison of severity estimates for Rater 5 (Figure 2) and Rater 46 (Figure 3) illustrate that different raters maintain different personal levels of severity. Rater 5 has an overall severity of -1.90 (SE= ±0.02) and Rater 46 has a severity of -1.17 (SE= ±0.02). A side-by-side comparison of the their plotted severities makes this difference even more obvious. The raw score impact of this difference in rater severities can be substantial. For example, an examinee who receives 57 out of a possible 75 points from Rater 5 would likely receive only 47 points from Rater 46.

Usually raters can maintain a self-consistent level of severity across administrations. To illustrate, these nine raters, as a group were consistent 67% of the time, that is a rater's severity was within two standard errors of the rater's overall degree of severity. A higher percentage can be expected if frequently inconsistent raters (i.e. Judge 62) are screened out in advance.

Another finding is that some raters are more consistent than others. This is illustrated by comparing Rater 46 (Figure 3) with Rater 62 (Figure 4). Rater 46 exhibits a very consistent degree of severity across a span of eight years while Rater 62 exhibits a noticeably less consistent degree of severity across a similar span of time. While severity for Rater 62 (Figure 4) showed some variation across administrations, the degree of severity *within* administrations was relatively uniform (Infit MS=1.1, Outfit MS=1.1). Another, less striking, detail of the data is that for five of the nine raters (Figures 2, 5, 6, 7, and 8), the first severity estimate is not in line with the other severity estimates. This is probably the result of a new rater learning how they will apply the rating scale. It seems that most raters settle on a uniform degree of severity that they can apply consistently after one or two administrations.

Discussion

Generally, raters have their own unique internal standard which they apply fairly consistently. The results of this study confirm that raters' perceptions of excellence are not interchangeable, but are usually self-consistent. However, some raters maintain a standard more consistently than others and even very consistent raters can vary occasionally. It can never be known in advance exactly how severe a particular rater will be on any given occasion. Yet, a rater's past performance often suggests how they will rate in the future. This information can be helpful to psychometricians who are organizing or equating performance assessments across administrations.

The method for analyzing rater severity proposed in this paper is not a replacement for an on-going equating procedure, but it can aid developers of more established exams, those with historical data, in making decisions about raters. For example, a psychometrician may select a few raters to participate in several consecutive administrations for the purpose of maintaining the same frame of reference for rater severity. Common raters should be selected on the basis of their documented ability to maintain a uniform level of severity. Armed with historical information, psychometricians can seek out stable raters like Rater 46 (Figure 3) for this purpose. Others, like Rater 62 (Figure 4) can still be used across administrations because their degree of severity is consistent within administrations, but knowing their across-administration degree of severity has more variance, the psychometrician would not want to use them as a link back to the initial scale. They should be thought of as new raters each time they grade.

Additionally, viewing rater severity in this manner can generate hypotheses regarding how individual raters behave over time. For example, Figure 8 suggests that Rater 9 is becoming slightly more lenient with experience. A similar tendency could be suggested for Rater 23 (Figure 9) based on the last three administrations, but the data is less persuasive. If the psychometrician thinks that there has been a shift in severity and that the new level severity is likely to be stable, the psychometrician may want to consider updating the rater calibration bank with the new severity calibration. Another hypothesis suggested by this data set is that Rater 6 (Figure 6) initially had occasional problems maintaining a uniform degree of severity, but had internalized a standard by administration 7 and thereafter was very stable. Perhaps, for anticipating the future performance of Rater 6, administrations 1 through 6 should not be considered.

The primary importance of this paper is the methodology used to investigate rater severity. In practice, one would only use this methodology to analyze several administrations because it is rather labor intensive. An advantage of this method is that the plotted calibrations with their error bands (±2 SEs) provide a useful description of rater behavior over time. This picture permits the psychometrician to verify that things are going well or to identify problem areas.

The most obvious information noticeable from these charts is which raters are consistent and which are erratic across administrations. As stated earlier, this information can be used to select anchor raters, but it can also be used after the data has been collected. Suppose that out of 14 raters, only four raters had a known degree of severity. Further, suppose that two of these four anchor raters were more lenient by approximately the same amount on the current administration than in earlier administrations. How would the psychometrician know if the two raters who really became more lenient were more lenient? It would seem equally plausible that the two raters that really remained the same had become more severe. A potential answer is to review the historical performance of the four raters. It seems probable that the historically more stable raters would be less likely to be the ones who changed.

To prevent the above scenario, enough common raters should be employed so that if a small percentage of raters change in severity, it will be easy to identify which raters changed. Reviewing the historical data can allow the psychometrician to make a good guess that given the available pool of anchor raters (with known severity and cross-administration stability) (1) which raters should be selected, (2) how many of the raters are expected to change severity during this administration, and (3) how many raters will be needed to clearly identify those that have changed severity.

These findings are also important because they address a primary concern about the reliability of performance examinations. To achieve any reproducibility in pass/fail or placement decisions, differences among raters must be accounted for both within and across administrations. These data confirm again that raters do have their own unique perceptions of excellence and sometimes that perception changes over time.

References

Linacre, J. M. (1994). *FACETS, a computer program for analysis of examinations with multiple facets*. Chicago, IL: MESA Press.

Linacre, J. M. (1989). *Many-Facet Rasch Measurement*. Chicago, IL: MESA Press.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1996). The invariance of rater severity calibrations. In G. Engelhard & M. Wilson (Eds.), *Objective Measurement: Theory into practice, volume 3* (pp. 99-112). Norwood, New Jersey: Ablex Publishing.

O'Neill, T. R. & Lunz, M. E. (1996). *Examining the invariance of rater and project calibrations using a multi-facet Rasch model.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Rasch, G. (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago, IL: MESA Press.

Table 1. Individual Administration Summary Statistics for Raters, Projects, & Examinees

Administration	Rater Severity			Project Difficulty			Examinee Ability		
	Mean	SD	N	Mean	SD	N	Mean	SD	N
1	-1.41	(.20)	13	-.07	(.37)	15	.27	(.72)	226
2		missing			missing			missing	
3	-1.58	(.23)	14	-.11	(.37)	15	-.03	(.68)	210
4	-1.51	(.26)	18	-.12	(.32)	15	-.17	(.67)	300
5	-1.65	(.43)	18	-.07	(.37)	15	-.09	(.72)	217
6	-1.45	(.25)	17	.01	(.31)	15	.06	(.64)	324
7	-1.51	(.18)	15	.02	(.40)	15	-.03	(.59)	260
8	-1.53	(.20)	19	-.04	(.33)	15	.00	(.73)	321
9		missing			missing			missing	
10	-1.60	(.21)	18	-.04	(.33)	15	.04	(.66)	385
11	-1.52	(.29)	16	-.10	(.34)	15	.04	(.69)	271
12	-1.60	(.39)	18	-.10	(.30)	15	-.03	(.65)	381
13	-1.39	(.40)	12	-.02	(.36)	15	-.02	(.57)	265
14	-1.57	(.29)	18	-.07	(.34)	15	.09	(.71)	364
15	-1.62	(.27)	19	.02	(.38)	15	-.12	(.62)	192
16	-1.54	(.29)	18	.02	(.38)	15	-.10	(.57)	339
17	-1.57	(.38)	11	-.04	(.36)	15	-.08	(.59)	168
18	-1.61	(.33)	16	-.04	(.36)	15	.10	(.67)	281
19	-1.58	(.32)	12	-.07	(.31)	15	.10	(.67)	179
Pooled	-1.55	(.35)	57	0.00	(.35)	53	0.00	(.66)	4,683

Presented in logits.

Table 2. Raters' Severity Across Administrations (for 9 selected raters)

Rater	Administration 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
4	-1.01	missing	-1.73	-1.53	-1.91	-1.60	-1.18	-1.50	missing	-1.72	-1.54	-1.71	-	-1.57	-1.61	-1.68	-	-1.58	-1.73
5	-1.42	missing	-	-2.03	-1.63	-1.82	-	-1.99	missing	-1.99	-	-1.88	-1.74	-1.99	-1.91	-1.91	-	-2.13	-2.33
6	-1.07	missing	-1.55	-1.53	-1.84	-1.39	-1.60	-1.69	missing	-1.65	-	-	-1.61	-	-	-	-1.69	-	-1.56
7	-	missing	-	-	-2.05	-1.56	-1.59	-1.30	missing	-1.74	-1.39	-	-1.74	-1.57	-1.37	-1.58	-1.61	-1.54	-
9	-1.45	missing	-1.65	-1.71	-1.70	-1.48	-1.55	-1.99	missing	-	-1.76	-1.65	-	-1.75	-1.96	-1.87	-	-1.92	-
19	-1.48	missing	-	-1.39	-1.40	-1.04	-1.54	-1.42	missing	-1.50	-1.20	-	-1.41	-1.31	-1.47	-1.63	-1.50	-1.50	-
23	-	missing	-	-1.30	-.95	-1.09	-1.34	-1.40	missing	-1.30	-	-1.29	-	-1.58	-1.51	-1.51	-	-	-
46	-	missing	-	-1.09	-1.02														

Figure 1

$$\log\left(\frac{P_{nptjx}}{P_{nptjx-1}}\right) - B_n - D_p - C_t - S_j - F_x$$

where:

P_{nptjx} = Probability of examinee receiving score x by rater j on project p on task t.
$P_{nptjx-1}$ = Probability of examinee receiving score $x-1$ by rater j on project p on task t.

and

B*n*	= Ability of examinee *n*	(Facet 1)
D*p*	= Difficulty of project *p*	(Facet 2)
C*t*	= Difficulty of task *t*	(Facet 3)
S*j*	= Severity of rater *j*	(Facet 4)
F*x*	= Difficulty of achieving score *x* relative to score *x-1* on the rating scale.	

Figure 2

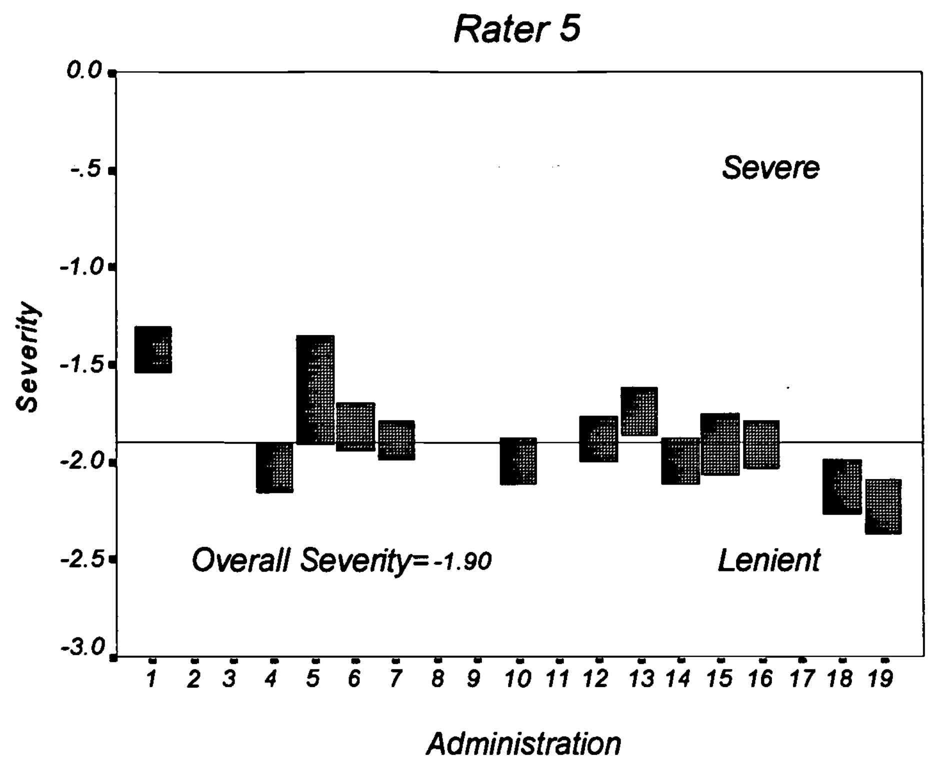


Figure 3

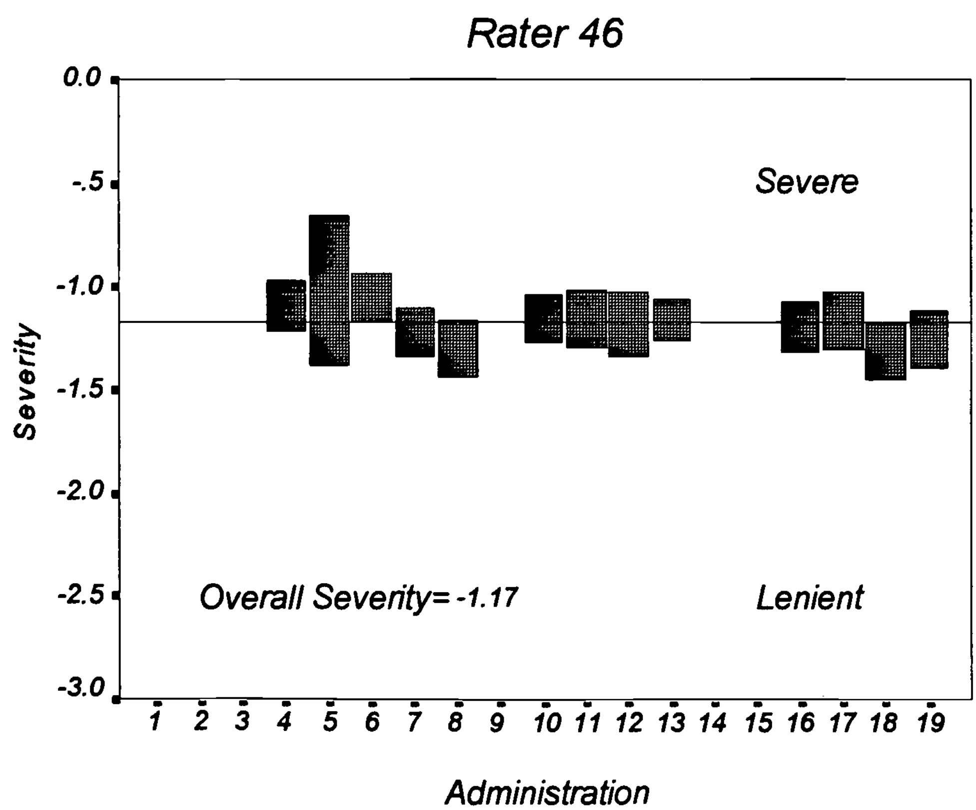


Figure 4

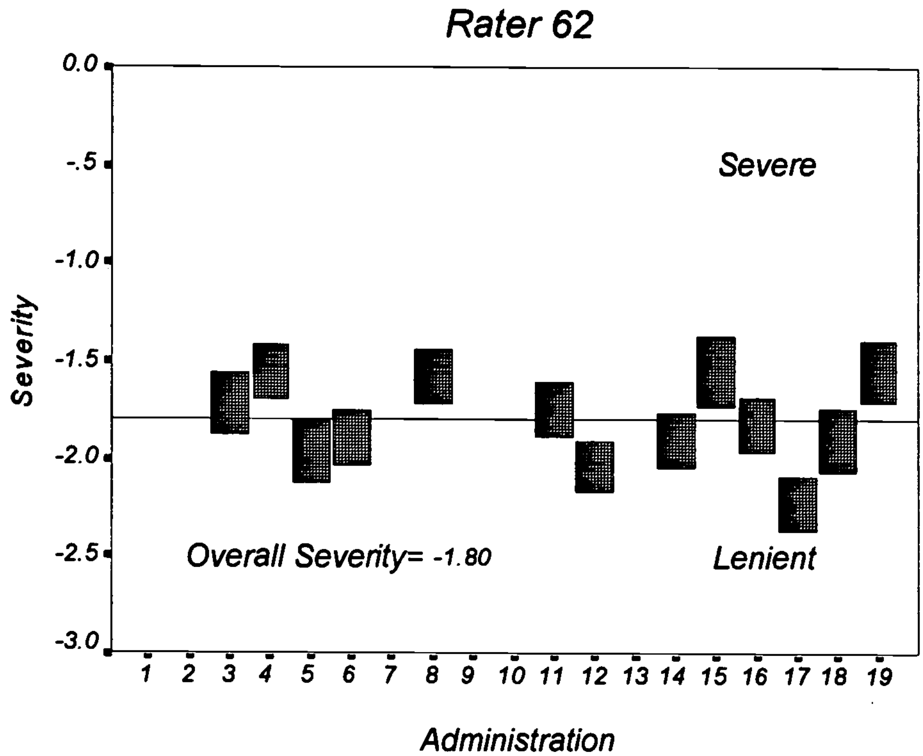


Figure 5

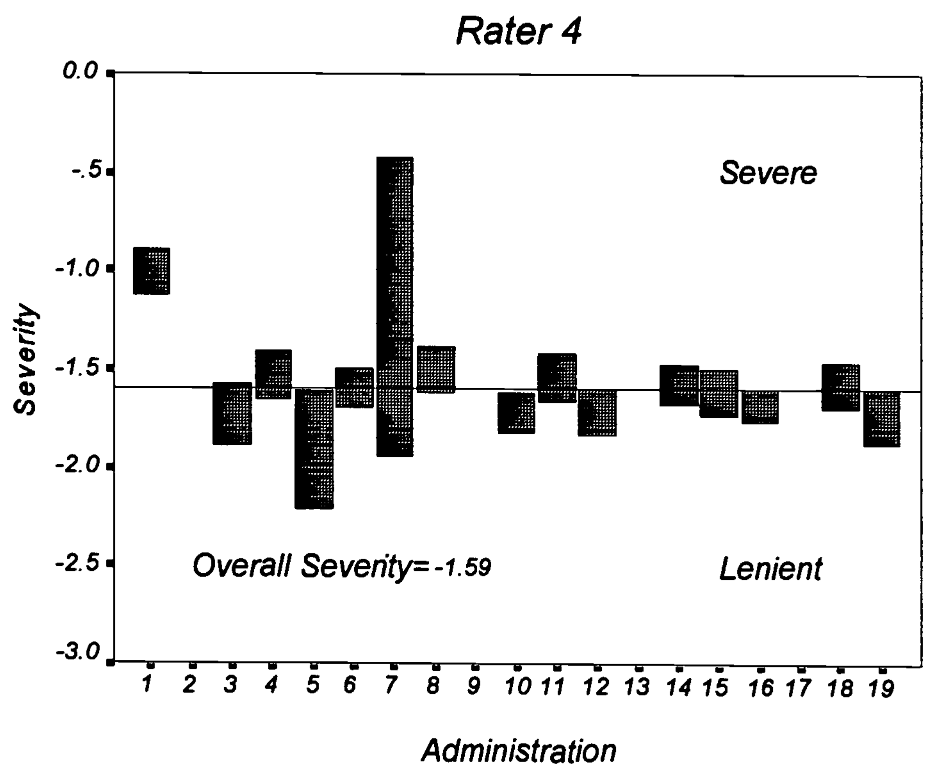


Figure 6

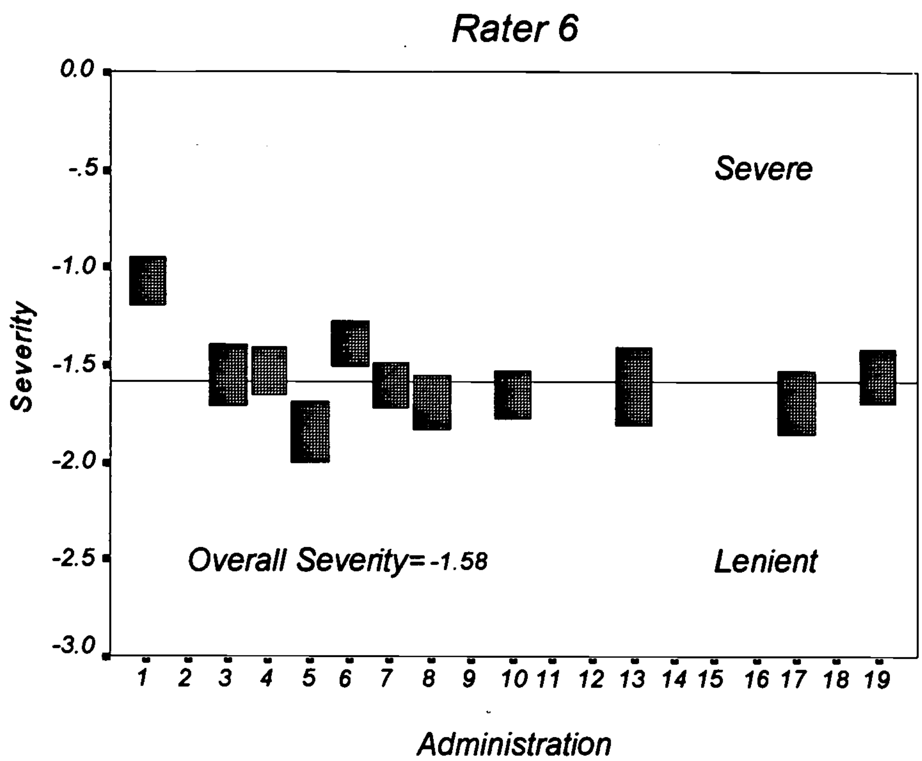


Figure 7

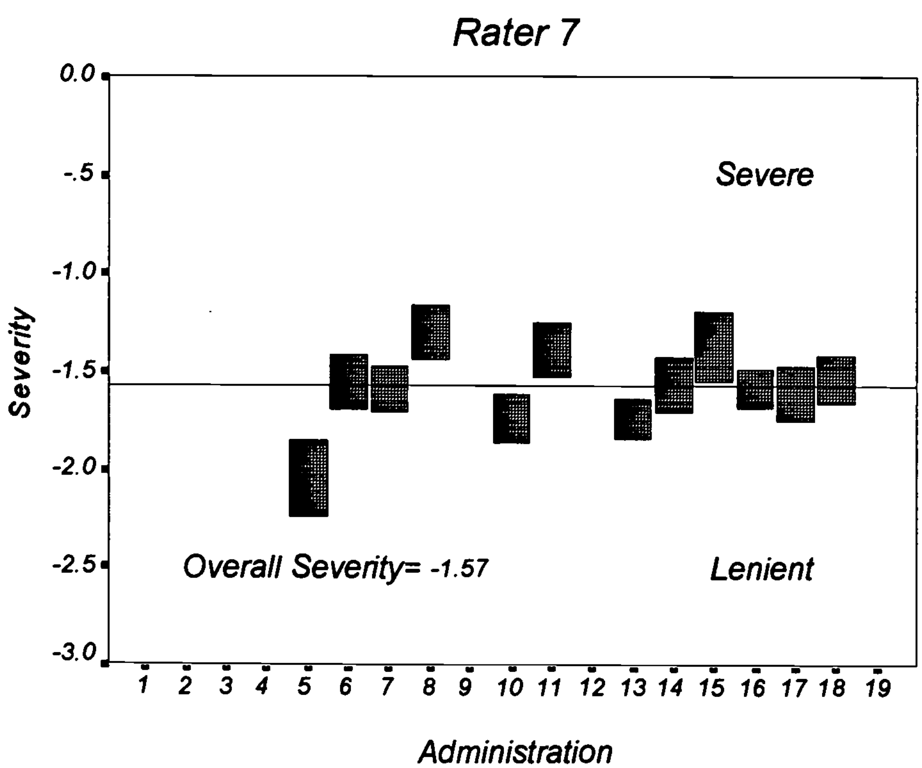


Figure 8

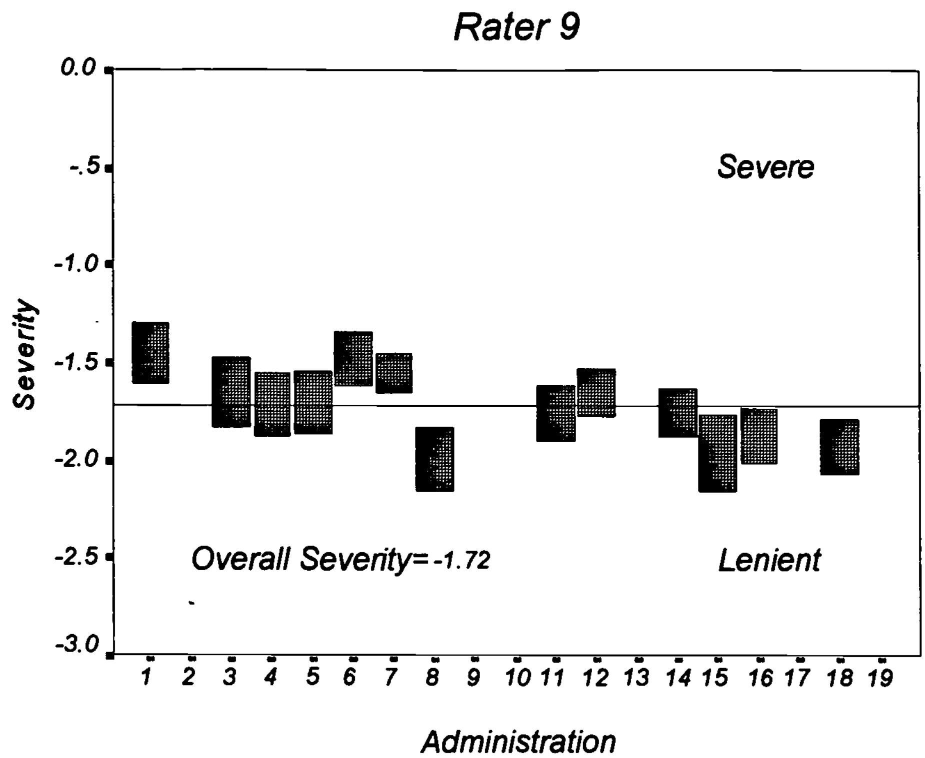


Figure 9

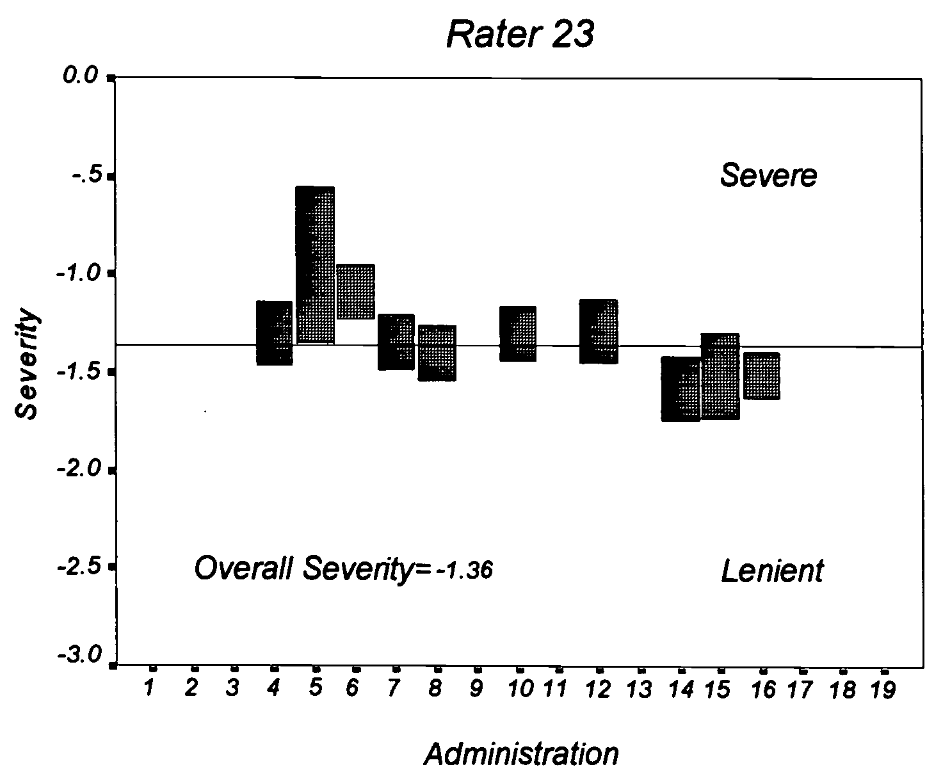
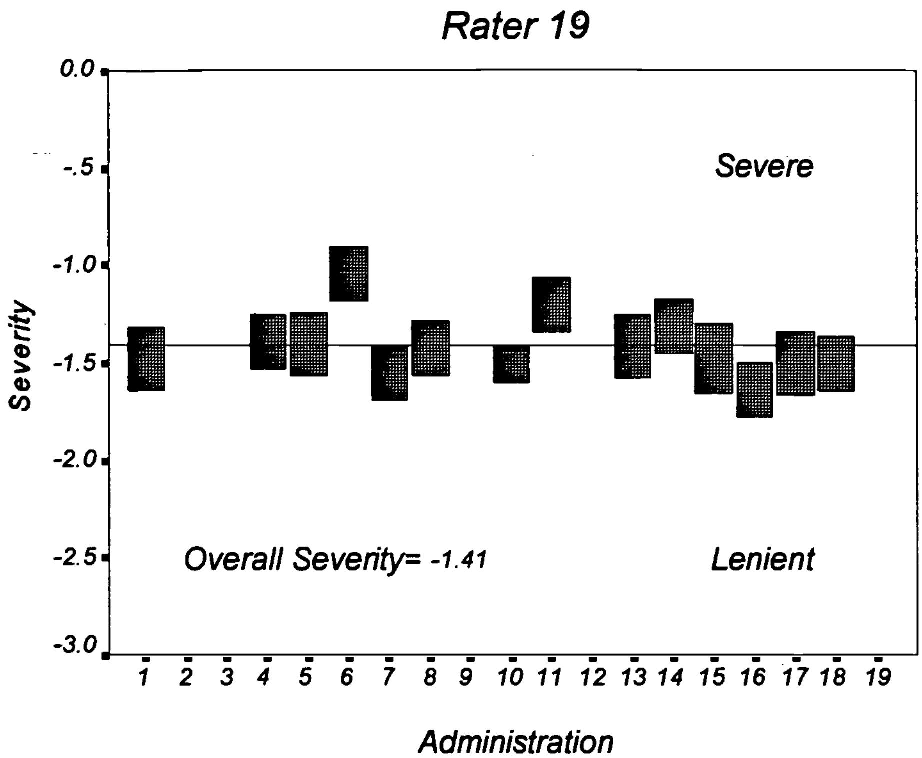


Figure 10


Rater 19
0.0
-.5
-1.0
-1.5
-2.0
-2.5
-3.0
Severity
Severe
Overall Severity= -1.41
Lenient
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Administration

TM027466

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A Method to Compare Rater Severity Across Several Administrations	
Author(s): Thomas R. O'Neill and Mary E. Lunz	
Corporate Source: American Society of Clinical Pathologists	Publication Date: March 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

[X] **Check here For Level 1 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all Level 2 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

[ ] **Check here For Level 2 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature: [signature]	Printed Name/Position/Title: Thomas R. O'Neill, Manager Evaluation and Research	
Organization/Address: American Society of Clinical Pathologists 2100 W. Harrison Street Chicago, IL 60612	Telephone: 312-738-1336	FAX: 312-738-5808
	E-Mail Address: TOMO@OM.ASCP.COM	Date: 5-28-97

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064

800 464-3742 (Go4-ERIC)

April 25, 1997

Dear AERA Presenter,

Hopefully, the convention was a productive and rewarding event. We feel you have a responsibility to make your paper readily available. If you haven't done so already, please submit copies of your papers for consideration for inclusion in the ERIC database. If you have submitted your paper, you can track its progress at http://ericae2.educ.cua.edu.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are soliciting all the AERA Conference papers and will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form on the back of this letter and stet **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can mail your paper to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 1997/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/E

ERIC® Clearinghouse on Assessment and Evaluation