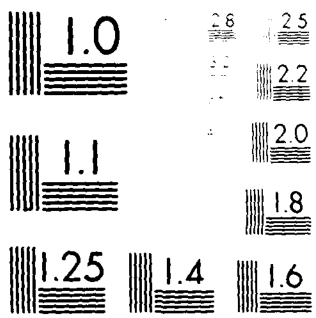
AD-A111 750 ARMY INST OF DENTAL RESEARCH WASHINGTON DC F/G 12/1
TESTING THE STURDINESS OF CLINICAL CRITERIA WITH DIFFERENT METH--ETC(U)
JAN 82 L LORTON, D D PETERS, M WONG

UNCLASSIFIED NL

1 OF 1
AD A
111 750

END
DATE
FILMED
04-82
DTIC


1.0
2.8
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO. AD-A111750

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

Testing the Sturdiness of Clinical Criteria With Different Methods of Simple Statistical Analysis

5. TYPE OF REPORT & PERIOD COVERED

Submission of Paper January 1982

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

L. Lorton; D. D. Peters; and M. Wong

8. CONTRACT OR GRANT NUMBER(s)

N/A

9. PERFORMING ORGANIZATION NAME AND ADDRESS

U.S. Army Institute of Dental Research
Walter Reed Army Medical Center
Washington, DC 20012

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

N/A

11. CONTROLLING OFFICE NAME AND ADDRESS

U. S. Army Medical Research & Development Command
HQDA-IS
Fort Detrick, MD 21701

12. REPORT DATE

29 January 1982

13. NUMBER OF PAGES

27

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)

UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Simple statistical analysis

MAR 5 1982

A

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A set of 96 samples composed of 6 subsets was evaluated by photographic and direct observation. Using a set of subjective criteria for excellence, the samples were rated by 3 different evaluators in both ordinal and categorical scales. The various data sets resulting were subjected to analysis by Krashal-Wallis and STP procedures (for the ordinal data) and by Ridit analysis (for the categorical data). It was found that the ordinal data produced consistent results both by photographic and direct evaluation. The Ridit analysis creates an informative index of excellence which allows a quality level value to be given to an irregular distribution. It is a —

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ADA111750

DTIC FILE COPY

practical method for producing data conducive to graphic display. Certain pitfalls in Ridit analysis make more complex analysis problematic.

Testing the sturdiness of clinical criteria
with different methods of simple statistical analysis

L. Lorton, D.D.S., M.S.D.*
D. D. Peters, D.D.S., M.S.**
M. Wong, D.D.S., M.S.

*Research Dental Officer, US Army Institute . tal Research
Walter Reed Army Medical Center, Washington, D. 20012

**Director, Endodontics Residency, US Army Institute of Dental Research
Walter Reed Army Medical Center, Washington, DC 20012

***Chief, Endodontics, 10th Medical Detachment, Yong San, Korea

The opinions expressed herein are those of the authors and are not to be construed as those of the Department of the Army or the Department of Defense.

82 03 08 060

ABSTRACT

A set of 96 samples composed of 6 subsets was evaluated by photographic and direct observation. Using a set of subjective criteria for excellence, the samples were rated by three different evaluators in both ordinal and categorical scales. The various data sets resulting were subjected to analysis by Krashal-Wallis and STP procedures (for the ordinal data) and by Ridit analysis (for the categorical data). It was found that the ordinal data produced consistent results both by photographic and direct evaluation. The Ridit analysis creates an informative index of excellence which allows a quality level value to be given to a irregular distribution. It is a practical method for producing data conducive to graphic display. Certain pitfalls in Ridit analysis make more complex analysis problematic.

In the areas of clinical or applied research where the data gathered is of a categorical or ordinal nature, there exist major barriers to successful analysis. A significant obstacle is the inability to establish fast criteria for evaluation which are resistant to operator imprecision, inter-operator differences, or imprecision in repeated evaluations.

Much dental research can be best evaluated by a good-better-best rating scheme which falls between strictly categorical data types (e.g., red, green, blue) and ranked or ordinal data. It is a constant source of difficulty to investigators to develop simple criteria which may be easily and universally applied to situations of clinical evaluation. These clinical judgments are often the subjective balancing of many subtle, and ofttimes undefined, variables.

Goldman et al.[1] refers to the low level of agreement amongst examiners (<50%) in evaluating the subjective success or failure of endodontic treatment diagnosed using radiographs. These same evaluators agreed with their own subsequent evaluations at a much higher rate (72-88%). The lack of sharply bounded standards for various levels of success complicate evaluation of clinical studies in all aspects of dental research.

Assuming that the "soft" criteria presently used for clinical researchers are the best and most appropriate available, are there certain schemes of data evaluation and analysis which are uncomplicated in application, presentation, and understanding, yet are resistant to the destructive variables previously mentioned?

The intent of this study was to examine the ability of using "soft" evaluative criteria, both categorical and rank-ordering rating schemes, and different statistical analysis to separate experimental treatment groups. At issue was the resistance of the methods to variations in evaluator and time; i. e., would the methods yield the same scores for the individual categories regardless of who did the evaluating, or when or how it was done?

METHOD

In a study involving the ability of various endodontic filling techniques to reproduce the surface of a root canal,[2] 96 specimens produced by 6 endodontic filling techniques were evaluated by the 3 investigators. At the first evaluation session using a 25X binocular microscope, each investigator categorized the surface of the filling into 1 of 4 types, poor to excellent, using the simple criteria given in Table 1. At this time, a 4 x 5 Polaroid photo was made of the surface.

During the next week, each separate evaluator placed the photos by rank from 1-96 with #1 being the photo considered having the best replication of surface detail, and #96 the photo showing the worse replication of surface detail.

One month later, the photos were evaluated and categorized into one of the four types as done with the microscope originally. At a subsequent session, the 96 photos were again ranked from 1-96. At each of the evaluation sessions, the technique group to which the filling belonged was unknown to the evaluators and the categorizations and ranking were decided independently by each evaluator. This resulted in two sets of data, ordinal and categorical.

The ordinal data was evaluated using the Kruskal-Wallis Analysis of Variance by Ranks[3] and a Simultaneous Test Procedure based on the Mann-Whitney U Test.[4] The categorical data (Tables 2a,b), based on the classification of the observations into the four categories, was analyzed by Ridit analysis.[5]

Ridit, or "reference to an identified distribution" (a specific data set, in this case the entire 96 observations by each operator) is a statistical technique that allows any type of distribution to be compared to a reference distribution much in the manner of χ^2 goodness of fit tests. It does allow the use of natural ordering present in the data which is otherwise lost. In data sets where the observations can be divided into categories which are sequentially related, but in which further definition as to absolute rank is not possible, or of spurious value, the use of the traditional ordinal tests is complicated by the necessity of correcting for multiple ties. These tests may yield more approximate results, are difficult to represent graphically, and yield no ample index of the value (sic) of a sample in which the variability of the sample is shown (as in the mean and standard deviation of interval measurements).

A distinct advantage in using Ridit analysis is that the relative value of different treatments can be estimated by their relative average ridit value. This will be further demonstrated in the section which explains the calculation of Ridit values.

The technique for conversion of data to Ridit values is illustrated below using the distribution of Evaluator I of the results of the direct vision evaluation.

The total population (96 observations) forms a frequency distribution which totals to 100%.

The technique for determining the value of each category is illustrated below using the distribution of Operator I using the original direct vision evaluation.

Category	# In Each Group 1	% In Each Group 2	Midpoint of Each Group 3	Percentile of Midpoint of Each Group 4
IV	38	39.58	19.79	80.21
III	28	29.17	14.59	45.27
II	23	23.96	11.98	19.27
I	7	7.29	3.65	3.65

Column 1 is the frequency of each cell. Column 2 is the relative frequency expressed as a percent. Column 4 is the so called "Ridit's" value of each category which is the sum of all the lower group percentage plus half the cell percentage.

It is these values in Column 4 which are used to calculate the average Ridit value for an experimental distribution.

For example, using the distribution assigned by Evaluator 1 to Technique A by direct vision:

			Ridit Value Previously Calculated From Reference Distr.	Total
IV	Excellent	7 X	.8021	5.6147
III	Good	5	.4583	2.2915
II	Acceptable	4	.1927	.7708
I	Poor	0	.0365	0.0
			divided by # of observations	16)8.67700
			average Ridit for Technique A for Evaluator 1	.542313

Thus, the average Ridit value of Technique A, as evaluated by Operator 1 by direct vision, is .542313

****

The average Ridit of any reference distribution is always .5000.

For example:

	Reference Distribution		Ridit Value	
IV	38	x	.8021	30.4798
III	28	x	.4583	12.8352
II	23	x	.1927	4.4321
I	7	x	.0365	.2555
	total 96		divided by	96)47.8915
				.4989

(The value of .4989 differs from .5000 only due to round-off error)

The value of .5423 can be interpreted as meaning that the average value of specimens produced by Technique A is extremely close to the average specimens from the reference distribution.

It is possible to compare between experimental groups by comparing their Ridit value.

For example:

Technique A has an average Ridit value of .5423 and Technique C has a value of .1415. The probability of A producing a better (sic) specimen is .5423-.1419 = .4004 +.5 or .9004. (If the Ridits were equal, the chance of either producing a better specimen than the other would be .5; i.e., .6523 - .6523 = 0 + .5 = 5).

The Ridits for different categories can then be compared using t-tests[5] or, as one investigator has done, by using parametric ANOVA techniques.[6]

RESULTS

Ordinal Testing:

The rank ordering of the 96 photos by the 3 evaluators at 2 different times yielded 6 data sets that could be compared. A Kendall coefficient of concordance (Table 3) amongst the sets indicated that all the ranking sets were similar.

Other correlations done between evaluators and within operators between times of evaluations were also strongly significant (Table 3). Based on the agreement amongst evaluators at one time period, the 3 sets of ranks for each time period were averaged to yield an average rank for each observation. This data set analyzed with ordinal tests to attempt to determine differences in the techniques. Mann-Whitney U tests were done to compare all parts of treatment groups and a simultaneous test procedure was done based on the results (Table 4).

The mean ranks are shown connected by underlinings where they cannot be shown to be different. Groups d, e, b, a were different from Group f which was different from Group c at the .05 level of probability.

Categorical Testing:

Table 5 lists the means and standard deviations of the Ridit scores for both direct and photographic evaluation.

This same data is displayed in Figures 1, 2, and 3 to illustrate the graphic possibilities of this type of data summary.

The results of the 3-way ANOVA of the Ridits are shown in Table 6. The 3-way interaction was significant, indicating that the Ridit scores for the various cells were affected differently by the various combinations of mode of evaluation and operator. This interaction precludes testing the main effects. The data was separated by operator and method of evaluation, and reanalyzed using 1-way analyses of variance. The post hoc tests based on these ANOVAS are presented in Table 7.

An arcsine transformation was done on the data; an accepted technique to normalize percentage or proportional data. The 3-way ANOVA results were similar; the 3-way interaction was significant (Table 8).

Paired-t comparisons were made between the Ridit values for the techniques for each operator according to the mode of evaluation. There was no significant difference for any operator that could be attributed to the manner of evaluation, direct or photographic.

DISCUSSION

The comparison of the same data base by two different statistical techniques provided some valuable insights about several problems that face clinical evaluators. 1) Can categorical ranking scales be compared between operators? 2) Are ordinal rankings more worthwhile and powerful in separating sample groups in clinical studies? 3) Is the Ridit statistic, which is intuitively attractive and easily understandable, as useful as the less intuitive ordinal test methods?

The ranked data, 6 sets of observations from 1-96, was evaluated first by measures of correlation (Table 7) which indicated that each operator at each time ranked the samples in essentially the same order. There was a statistically significant agreement within all possible pairs.

The degree of agreement amongst evaluators ranged from .57 to .85 and by evaluators with themselves from .77 to .91, agreeing fairly well with Goldman's figures.[1]

When the pooled ranks at each time period were analyzed using the post-hoc test (Table 4), the analysis at the two-time periods yielded the same mean rank orders for the treatment groups.

The influence of time or operator seems to be negligible using photo ranking techniques. This is in agreement with other conclusions on this technique.[9]

The Ridit values were analyzed in a 3-way design - (technique x operator x mode of evaluation). The avtable indicated significant interaction (Tables 7 and 9). This conclusion is not intuitively acceptable when the Figures 1-3 are compared. These results may be ascribed to characteristics peculiar to Ridit scores.

In our data base, the homogeneity of results within each technique group resulted in some techniques having all observations in the same rank category. This yields a cell mean for analysis of variance with a zero variance. This seems to have unfavorable effect on the analysis of variance by causing small differences in actual operator/mode evaluation levels to be unnecessarily prominent statistically.

In effect, the small number of categories implies a precision of evaluation that does not exist.

The pitfall of Ridit analysis seems to be that it implies an accuracy of evaluation that is not really present. Since the Ridit value for any category can be extended to as many significant digits as is convenient, the illusion of precision may be increased at will. In a practical sense, if Evaluator 2 were to assign all of one set of specimens to Category 4, the average Ridit would be .8021 with a standard deviation of 0. If, on subsequent evaluation, he evaluated the same group and assigned 13 to Category IV, and 3 to Category III, the average Ridit for the group would be .7376 with a s.d. of .13859. This difference ($t_s = 1.86$) is statistically significant at between .05 and .10, and intuitively there seems to be a difference between the numbers .8021 and .7376, yet with the loose criteria, it is very easy to allow certain borderline cases to fit into either one of two categories. Thus a single observation could be given to a value of either .8021 or .7376; this will cause an effect in the mean and standard deviation out of proportion to the deviation of opinion which caused it.

The precision of accuracy, which is implied by the four significant figures after the decimal point, should be recognized as being only a construct unrelated to the precision of the evaluation technique.

Because of the significant interaction in the main ANOVA, the data was reevaluated to judge the degree of separation of the techniques when the operators and mode of evaluation were all considered separately.

Individual analysis of the Ridit scores for each mode/evaluator combination showed that distinctions drawn between the treatment groups are essential the same for each operator and mode. Ridit scores for the different techniques can be seen to be approximate the same between operators. This is because the Ridit analysis is essentially a rank-ordering technique[8] related to Wilcoxon rank sum test. Minor differences between Ridit scores for the same categories are due to minor discrepancies in applying the "sliding scale" of criteria by the evaluators.

The Ridit analysis done by direct vision gave the same relative Ridit scores to the various techniques resulting in the same separations as the Ridit evaluation by photo (Table 7). It can be inferred that even such simple criteria as listed in Table 1 can be applied evenly both in pictures and by direct vision with some confidence that undue error is not introducted.

Thus, although the Ridit analysis is "self-correcting" to some extent, the appearance of precision implied by the Ridit score for each category is actually specious. The very broadness of the category decreases the discriminative ability of the Ridit analysis, and the implied precision makes complex mathematical treatment (as in multiway ANOVA) problematical.

Ridit analysis seems to be best suited to graphical display of data, simple inference, and as an intuitively appealing index.

CONCLUSIONS:

The conclusions that were drawn from this comparison are as follows: 1) Ridit analysis is a remarkably attractive tool for use in categorical-ordinal situations. It is partially self-correcting between operators, but the overly precise "appearance", e.g., 8021, .0365, etc. of the Ridit score implies an accuracy of measurement which does not exist. 2) The use of Ridit analysis in graphic representation of results seems to be far more trustworthy then extensive analysis using parametric tests. 3) The Ridit analysis provides a manageable way for characterizing the value of non-normal distributions and provide an useful measure of "central tendency" for this type of data. The traditional ordinal tests provide as much, or more, information about the relative worth of different samples but, in this case, required a great expenditure in time and resources to provide the pictures for ranking. 4) The Ridit characterizations were done with equal certainty and with equally sound results by direct vision without the need for reproduction. 5) Classification of root canal fillings into categories based on subjective assessment by direct vision through a microscope was shown to provide the same relative results as evaluation of pictures of the same field. Differences between operators, time evaluated, and method of result analysis produced statistically significant, but small differences that did not affect the relative rankings of the various root canal fillings technique employed.

References

1. Goldman, M., Pearson, A., and Dargenta, N. Reliability of Radiographic Interpretation. *O.S.* 38(2):287-293, August 1974.
2. Wong, M., Peters, D.D., and Lorton, L. Comparison of Gutta-Percha Filling Techniques, Compaction (Mechanical), Vertical (Warm), and Lateral Condensation Techniques, Part 1. *J. Endod.* 7(12):551-558, December 1981.
3. Wong, M., Peters, D.D., Lorton, L., and Bernier, W.F. Comparison of Gutta-Percha Filling Techniques:Three Chloroform--Gutta-Percha Filling Techniques, Part 2. *J. Endod.* 8(1):4-9, January 1982.
4. Conover, W. J. 1971. Practical Non Parametric Statistics, John Wiley & Sons, Inc., New York, p. 256.
5. Sokal, R. and Rohlf, F.J. 1969. Biometry, W. H. Freeman & Co., San Francisco, p. 392.
6. Bross, I. How to Use Ridit Analysis. *Biometrics* 14:18, March 1958.
7. Mahler, D., Terkla, L.G., and VanEysden, J. Marginal Fracture of Amalgam Restorations. *J. Dent. Res.* 52:823, July 1973.
8. Conover, W. J. 1971. Practical Non Parametric Statistics, John Wiley & Sons, Inc., New York, p. 270.
9. Selvin, S. A Further Note on the Interpretation of Ridit Analysis. *Am. J. Epidemiol.* 105:16, January 1977.
10. Osborne, J. et al. Three Year Clinical Comparison of Three Amalgam Types. *JADA* 93:784, October 1976.

Table 1

Criteria of Nominal Classification

Class I :	(Poor)	No Apical Replication	Wrinkles & Folds	No Fins
Class II :	(Acceptable)	Poor Apical Replication	Some Wrinkles & Folds	Minimal Fins
Class III:	(Good)	Good Apical Replication	Few Wrinkles & Folds	Fins
Class IV :	(Excellent)	Excellent Apical Replication	No Wrinkles & Folds	Fins

Table 2a

Example DISTRIBUTION OF ORDINAL/CATEGORICAL DATA

DIRECT VISION EVALUATION

Technique	I Poor	II Fair	III Good	IV Excellent	Total
Operator I					
a		4	5	7	16
b		2	3	11	16
c	7	9			16
d			7	9	16
e		1	5	10	16
f		7	8	1	16
TOTAL	7	23	28	38	96
Operator II					
a	1	0	9	6	16
b		1	2	13	16
c	12	4			16
d			5	11	16
e			3	13	16
f		10	6		16
TOTAL	13	15	25	43	96
Operator III					
a	1		10	5	16
b		1	3	12	16
c	11	5			16
d			4	12	16
e			2	14	16
f		10	6		16
TOTAL	12	16	25	43	96

Table 2b

Photographic Evaluation

Technique	I Poor	II Fair	III Good	IV Excellent
Operator I				
a	1	3	4	8
b	1	2	5	8
c	11	5		
d			1	15
e		2	8	6
f		7	8	1
Operator II				
a	1	5	3	7
b	1	2	7	6
c	14	2		
d			1	15
e	2	2	11	1
f	2	8	4	2
Operator III				
a	3		7	6
b				16
c	16			
d			6	10
e				16
f			6	10

Table 3

Analysis of Ordinal Data

Kendall Coefficient of Concordance

degree of agreement amongst all 3 operators of pictures ranked 1-96

initial evaluation	.7746	$p < .0000$
second evaluation	.8427	$p < .0000$

Spearman Rho

degree of agreement between evaluators

initial evaluation

Evaluators	Rho value	t value	
1 vs 2	.8496	11.4	$p < .000$
1 vs 3	.4869	6.88	$p < .000$
2 vs 3	.5703	7.32	$p < .000$

second evaluation

1-2	.6916	8.32	$p < .000$
1-3	.86667	12.05	$p < .000$
2-3	.7332	8.84	$p < .000$

degree of agreement within operator between initial and second evaluation

1	.6730	8.13	$p < .000$
2	.9119	14.65	$p < .000$
3	.7721	9.46	$p < .000$

Table 4

Separation of Techniques by STP Based on
Mann-Whitney U-Tests Using Ordinal Data

Initial evaluation

technique	d	e	b	a	f	c
mean pooled rank	64.4	110.3	126.0	127.5	202.4	241.2
					$p < .05$	

Second evaluation

technique	d	b	e	a	f	c
mean pooled rank	73.4	105.6	112.8	132.0	194.3	262.6
					$p < .05$	

Table 5

Ridit Scores

		Evaluator 1		Evaluator 2		Evaluator 3	
		Direct	Picture	Direct	Picture	Direct	Picture
Technique A	$\bar{x}$	.5423	.5938	.5326	.5801	.5101	.4753
	s	.2578	.2578	.2136	.2698	.2052	.2471
B	$\bar{x}$	.6615	.6136	.6966	.6078	.6731	.7400
	s	.2282	.2401	.1763	.2302	.1884	.0000
C	$\bar{x}$	.1419	.1146	.1042	.1302	.1081	.099
	s	.0593	.0746	.0652	.0711	.0648	.0000
D	$\bar{x}$	.6417	.7878	.6653	.8308	.6875	.6191
	s	.1775	.0573	.1695	.0729	.1584	.1745
E	$\bar{x}$	.6351	.6172	.7096	.4883	.7317	.7500
	s	.2056	.1837	.1427	.1896	.1209	.0000
F	$\bar{x}$	.3636	.4414	.2917	.4147	.2884	.3066
	s	.1767	.1921	.1042	.2214	.1068	.0755

Table 6

3-Way Analysis of Variance for Ridit Scores
(raw data)

Main Effects	Sum of Squares	DF	MS	F	P
technique	23.372	5	4.674	155.034	.000
evaluation	.004	2	.002	.061	.941
mode of evaluation	.000	1	.000	.000	.989
2-way interaction					
technique x evaluation	.924	10	.092	3.066	.001
technique x mode	.864	5	.173	5.731	.005
evaluation x mode	.004	2	.002	.058	.943
3-way interaction					
technique x evaluation x mode	1.029	10	.103	3.414	.000
residual	16.281	540	.030		

Table 7

Separation Produced by Individual ANOVA
+Student Neuman Keuls Test (at .05 Confidence Level)

Evaluator	Technique	F Value of ANOVA	P
	Direct Vision		
1	d b e a f c	6.517	.0000
2	e b d a f c	3.540	.0037
3	e d b a f c	3.681	.0027
	Pictures		
1	d b a e f c	2.904	.0134
2	d b e a f c	7.217	.0000
3	e b d f a c	3.028	.0104

Table 8

3-Way Analysis of Variance
(data transformed with arcsine junction)

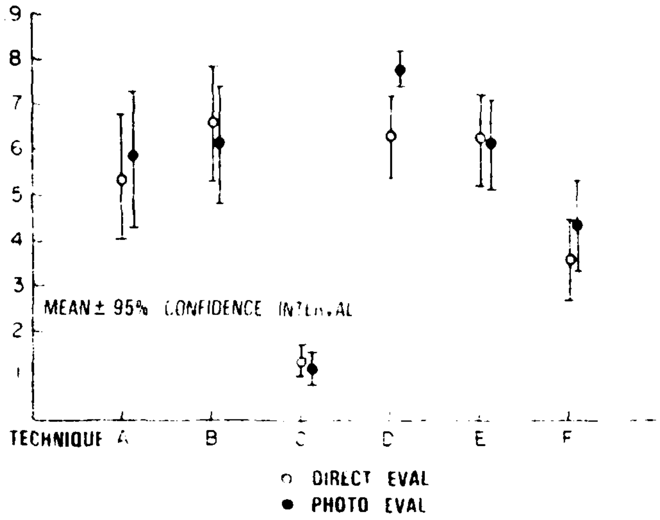
Main Effects	Sum of Squares	DF	MS	F	P
technique	31.671	5	6.334	140	.0000
evaluation	.00414	2	.002	.05	.9553
mode of evaluation	.00004	1	.0004	.00	.9769
2-way interaction					
technique x evaluation	1.44449	10	.14445	3.19	.0005
technique x mode	1.30621	5	.27724	6.13	.0000
evaluation x mode	.00412	2	.00206	.05	.9554
3-way interaction					
technique x evaluation x mode	1.60539	10	.16054	3.55	.0001
residual	24.42137	540	.04522		

Legend

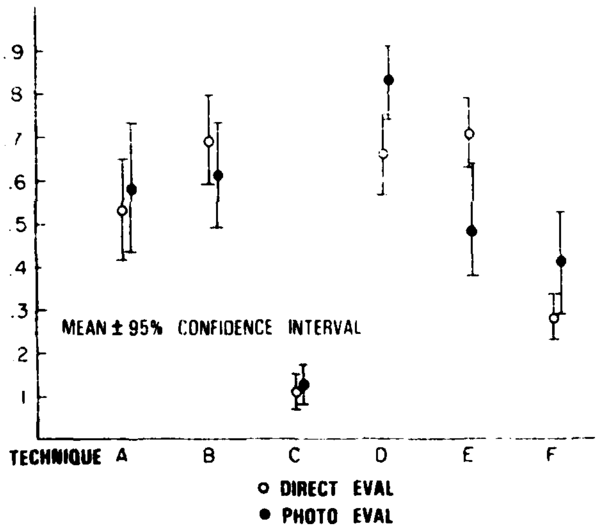
Figures 1, 2, and 3

Average Ridit scores and 95% confidence intervals illustrate ease of graphic interpretation of a Ridit score.

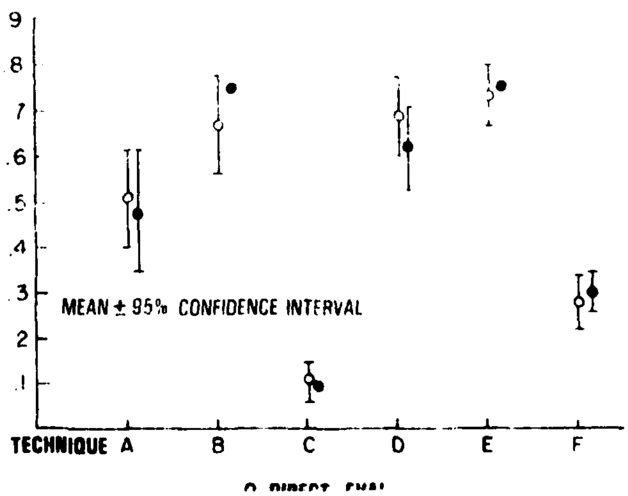
RIDIT SCORES
EVALUATOR I



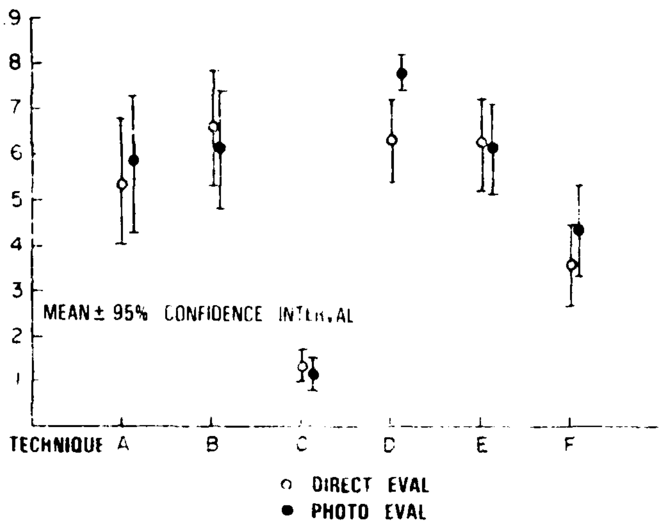
RIDIT SCORES
EVALUATOR II



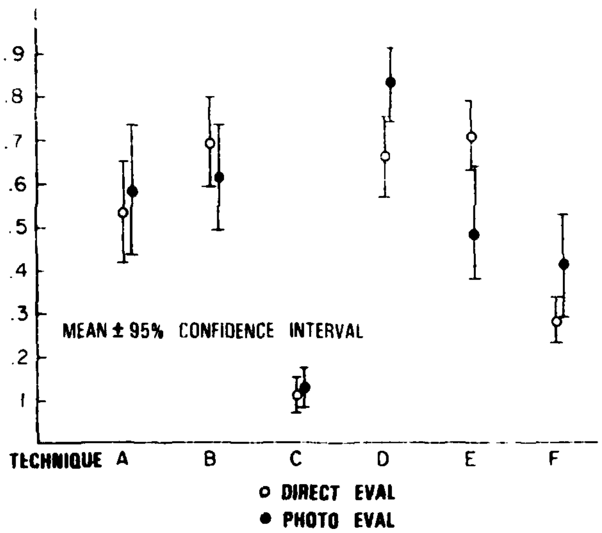
RIDIT SCORES
EVALUATOR III



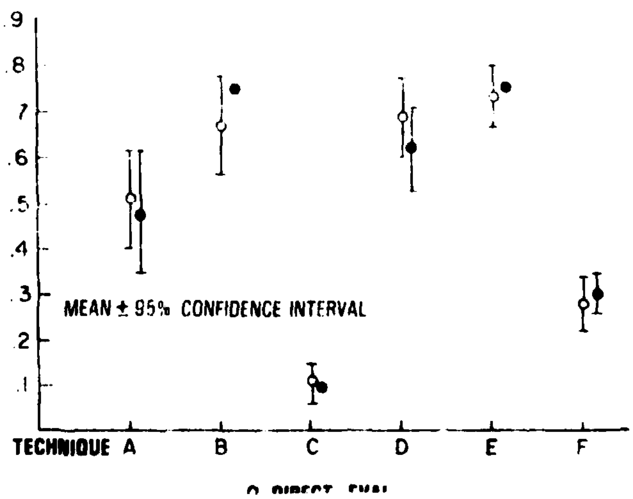
RIDIT SCORES
EVALUATOR I



RIDIT SCORES
EVALUATOR II



RIDIT SCORES
EVALUATOR III



4-8

DTIC