DOCUMENT RESUME

ED 079 382 TM 002 982

AUTHOR Haley, John V.
TITLE Effects of Training on the Test of Diagnostic Skills. Publication No. 30.
INSTITUTION Loyola Univ., Chicago, Ill. Psychometric Lab.
SPONS AGENCY Commonwealth Fund, New York, N.Y.
REPORT NO Pub-30
PUB DATE 63
NOTE 16p.; Paper presented at Association of American Medical Colleges' First Annual Conference on Research in Medical Education, October, 1962

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Clinical Diagnosis; Cross Sectional Studies; Longitudinal Studies; *Medical Education; Medical Students; Performance Criteria; *Performance Tests; Professional Training; *Scoring; *Statistical Analysis; Student Evaluation; Technical Reports; Test Results
IDENTIFIERS *Test of Diagnostic Skills

ABSTRACT

This report summarizes research performed on the Test of Diagnostic Skills, used to evaluate the clinical diagnostic skills of medical students. Forms of the test were administered to groups at different levels of medical experience to ascertain the effect of training on performance. A cross-sectional study was conducted with approximately 90 junior medical students, 145 seniors, and 40 physicians. A longitudinal study was conducted with 36 subjects, who were tested at the end of their junior and senior years. Individual and group performances were analyzed. The following were computed: (1) mean number of questions asked by groups; (2) a utility index, the ratio between the number of times a question was asked and the total number in the group, for each question; (3) maximum and minimum performance curves for individuals; (4) utility scores, the average of the utility indexes of all the questions asked by a particular subject. A pattern analysis and a sequential evaluation, in which students were compared to norms developed for physicians and clinicians. (For related document, see TM 002 981.) (KM)

PSYCHOMETRIC LABORATORY

EFFECTS OF TRAINING ON THE TEST OF DIAGNOSTIC SKILLS

by

John V. Haley

Loyola University
Chicago, Illinois

1963

Publication No. 30

Effects of Training on the Test of Diagnostic Skills*

by

John V. Haley

This report summarizes some of the research that took place during a six year period on the Test of Diagnostic Skills. This technique was devised by H. J. A. Rimoldi for the purpose of evaluating medical students. (Rimoldi, 1955). Since this approach is different from the usual methods of evaluation, it was necessary to develop adequate scoring procedures as well as statistical tests to analyze the results. Although this was one of the major problems of the study, the present discussion will treat primarily the conclusions based on these scoring methods and the statistical tests that were adopted.

Forms of the Test of Diagnostic Skills were administered to groups at different levels of medical experience to ascertain the effect of training on performance in this test. Two separate analyses were made. In the first, a cross sectional study, the performances of three groups having different amounts of medical training were compared. In the second study (longitudinal) the performance of a group at an earlier stage of training was compared with the performance of the same group at a later stage.

The sample used in the cross sectional study consisted of approximately 90 junior medical students, 145 senior medical students and 40 physicians, selected from five medical schools in the eastern and midwestern area. In the longitudinal study 36 subjects were selected from a Chicago area medical school.

Two forms of the Test of Diagnostic Skills will be discussed, Test 2 and Test 4. Both are complete transcriptions of actual clinical histories. Medical questions pertaining to each case were included. These questions cover three phases of the diagnostic process: clinical interview, physical examination and laboratory tests.

The students in the cross sectional study were administered both tests in group form toward the end of the school year. The longitudinal group took Test 2 and Test 4 at the end of their junior year and senior year of medical school.

Subjects were presented with folders containing 3 x 5 cards. A question was printed on each card with the corresponding answer given on the reverse side. Each subject was given the chief complaints and admission data and instructed to read over all of the questions. He then selected those which he considered necessary and sufficient to reach a diagnosis. After each

* This research was supported by grants from the Commonwealth Fund of New York.

Paper read at the Association of American Medical Colleges' First Annual Conference on Research in Medical Education, October, 1962.

selection he recorded the number of the question, and read the information contained on the back of the card.

Individual and group performances were analyzed in terms of: number and popularity of questions asked; group agreement concerning the usefulness of the questions; and a criterion group of physicians, both clinicians and surgeons. Other methods have been attempted but will not be reported here either because they did not add significantly to already existing methods or because they had not yet been thoroughly analyzed at the time this research was performed.

Number of Questions Asked

The mean number of questions asked by the groups of both studies are presented in Tables I and II for Test 2 and Test 4 respectively. In general, it was found that all of the subjects regardless of training level selected more questions in Part I (interview) than either Part II or Part III. Significant differences were found between the three levels for the number of questions asked in the total test and for the number of questions asked in Part I. Juniors had to ask more questions in reaching a solution than seniors, and seniors had to ask more questions than physicians.

It appeared that juniors were less able to integrate information acquired during the interview phase of the test than seniors, and physicians were able to integrate this information the best. Juniors asked the largest number of questions and physicians the least number in the interview phase of the test. No differences were evident among the three groups for the number of questions asked which referred to physical examinations or laboratory tests.

Utility Index

The utility index has been defined as the ratio between the number of times a question was asked and the total number in the group. (Rimoldi, 1955). Depending upon the group used for its definition, the utility index for the same question may have different values. This value is independent of the position in which a question is asked. Some questions were asked very frequently by physicians, less by seniors, and the least by juniors, or vice versa. Using Chi square and Fisher's exact probability test it was possible to show that many of these changes are significant. Interpretation of these differences indicates that they are consistent with medical expectations. Questions asked more frequently by physicians were more directly related to the case, more efficient and less redundant.

Table I

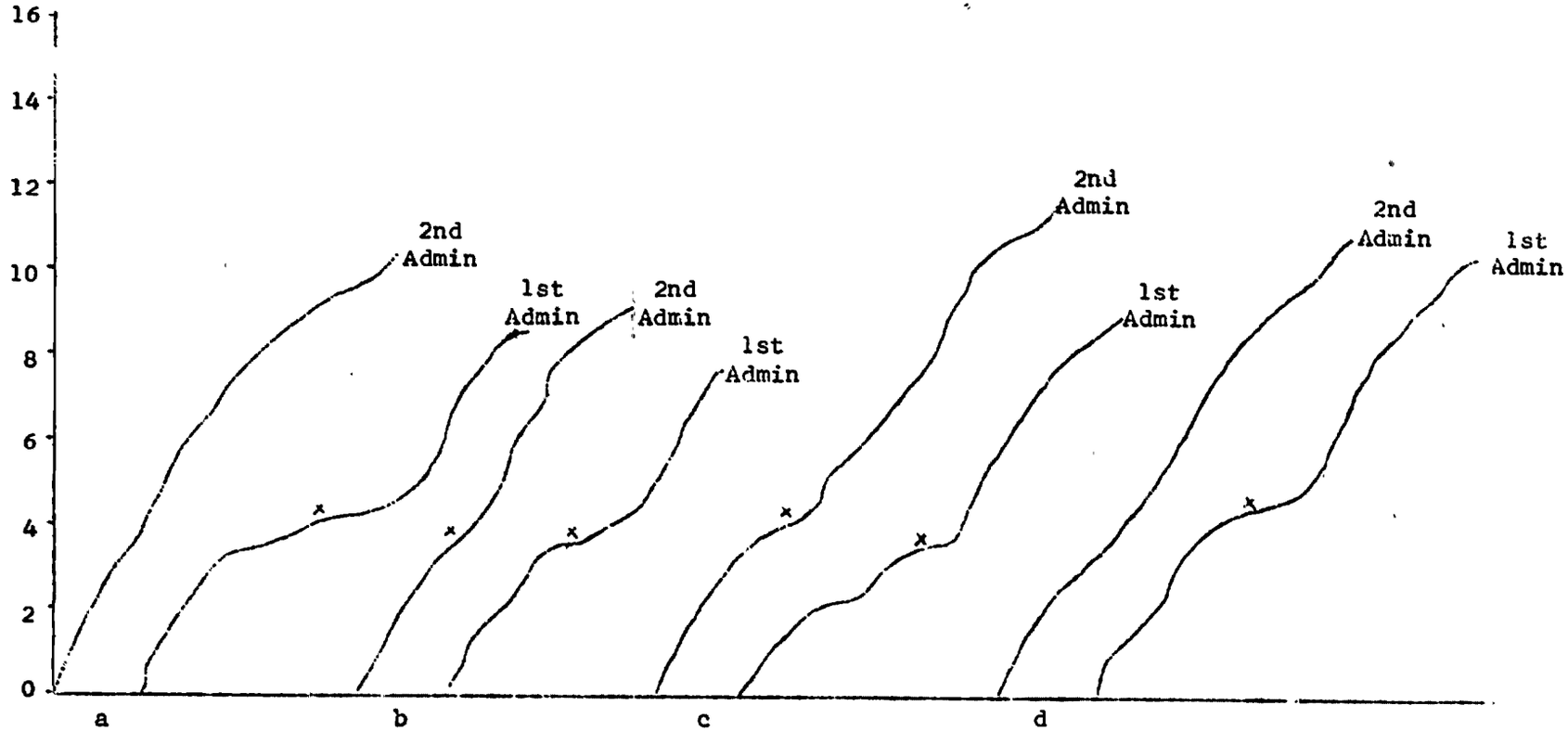
Mean Number of Questions Asked
in Each Part and in the Total
for Test 2

Group	N	Part I	Part II	Part III	Total
Juniors	87	9.26	4.45	5.98	19.69
Seniors	129	8.31	4.10	6.15	18.58
Physicians	41	7.00	4.02	6.73	17.76
First Administration	36	11.28	6.03	6.64	23.94
Second Administration	36	10.03	4.50	6.36	20.89

Table II

Mean Number of Questions Asked
in Each Part and in the Total
for Test 4

Group	N	Part I	Part II	Part III	Total
Juniors	87	10.51	5.80	6.47	22.78
Seniors	147	8.28	5.39	5.82	19.50
Physicians	40	7.22	4.82	6.28	18.32
First Administration	36	12.39	6.61	7.31	26.31
Second Administration	36	8.89	4.81	6.22	19.92



Individual Performance Curves for Two Administrations of the Test of Diagnostic Skills, Test 4 for Four Subjects

Figure I

In the longitudinal study the utility indexes of the questions asked by each subject were accumulated at each successive step in the Test (Figure 1). This was done for both tests and for both administrations. These utility indexes were based on the performance of physicians. The curves in Figure 1 represent the performance of four individuals who took Test 4 in both administrations. The plateaux indicated by an "X" appeared in a large majority of the curves representing performance in the first administration. In the second administration (senior year) these plateaux either disappeared or became less extreme in both tests. The questions asked where these plateaux appeared were from the interview phase of the test in every case. It appears that in the first administration the subjects' perseverance in choosing items not selected by physicians in the interview section of the test was the primary cause for the increase of questions asked and the plateaux. In the second administration the subjects were less persistent in selecting items in Part I; the plateaux disappeared or became less marked, fewer questions were asked and the utility indexes of the questions asked were higher.

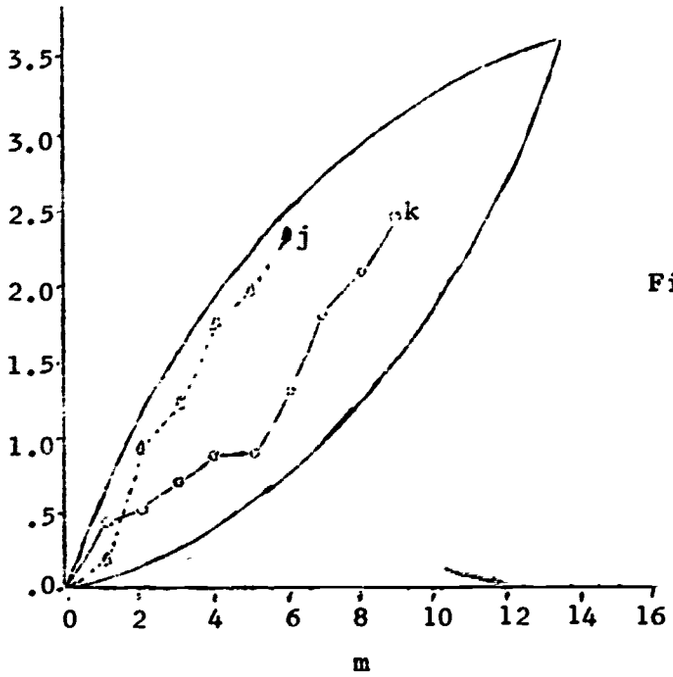
The examples given in Figure 1 were not selected because they emphasized the interpretation given above. However they were selected from Test 4 since this test seemed to be more sensitive to the training that took place. In Test 4, 31 of the 36 subjects followed this pattern, while only 27 of the 36 followed it in Test 2. (Haley, 1960).

Limits of Performance

Group performance in the Test of Diagnostic Skills can be described in terms of maximum and minimum performance curves. These can be defined as the "best" and "worst" possible way in which an individual in a particular group can perform. They are derived by ranking the questions of the test according to their utility index from maximum to minimum and from minimum to maximum. By accumulating the values of the utility indexes on the ordinate against rank order on the abscissa, limits of performance can be described. (Figure 2). Although subjects seldom if ever perform in this fashion, nevertheless, it has proved to be a very useful method for defining the limits of performence of a particular group.

If the utility indexes of all of the questions have the same value then the area collapses into a straight line whose slope is equal to this constant value. However, if a group asks certain questions more frequently than other questions, some utility indexes will be higher than others. This will generate performance curves whose area can be shown to be a function of the discriminative power of the test, or of the agreement of the group.

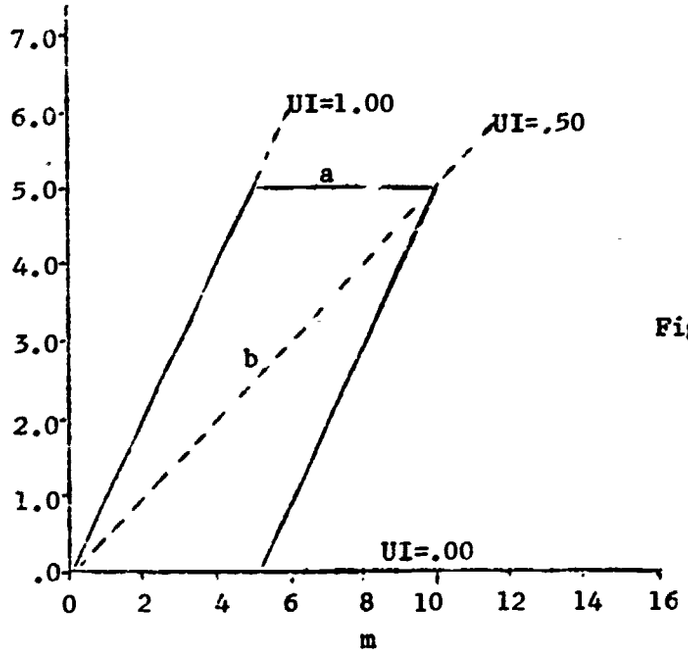
When some questions are asked by every subject in the group and the remaining questions are not asked at all, the utility indexes will be either 1.00 or .00. The resulting limits of performance would generate a parallelogram. (Figure 3). This occurs when there is perfect agreement within the



Σi

Figure II

Limits of Performance



Σi

Figure III

Limits of Performance With Perfect Agreement (a) and With No Agreement (b)

group on which questions are useful in reaching a solution. The ratio between the area of the limits of performance and the area of the parallelogram that would result if there were perfect agreement would then give a measure of agreement.

In all of the problems examined thus far the equation of the maximum and minimum curves is of the form

$$y = c(1 - e^{-bx})$$

where "c" is the "y" asymptote, "y" is the sum of the utility indexes, "x" is the number of questions asked and "b" is the slope. Within this formulation "b" is a constant ratio of the information yet to be gained. It indicates decrement of ignorance. So at the beginning of the test when very little is known, the subject should gain more information or decrease his ignorance at a greater rate than at the end of the process. So the value of "b" reflects the difficulty of the test. (Rimoldi, Devane, Haley, 1961).

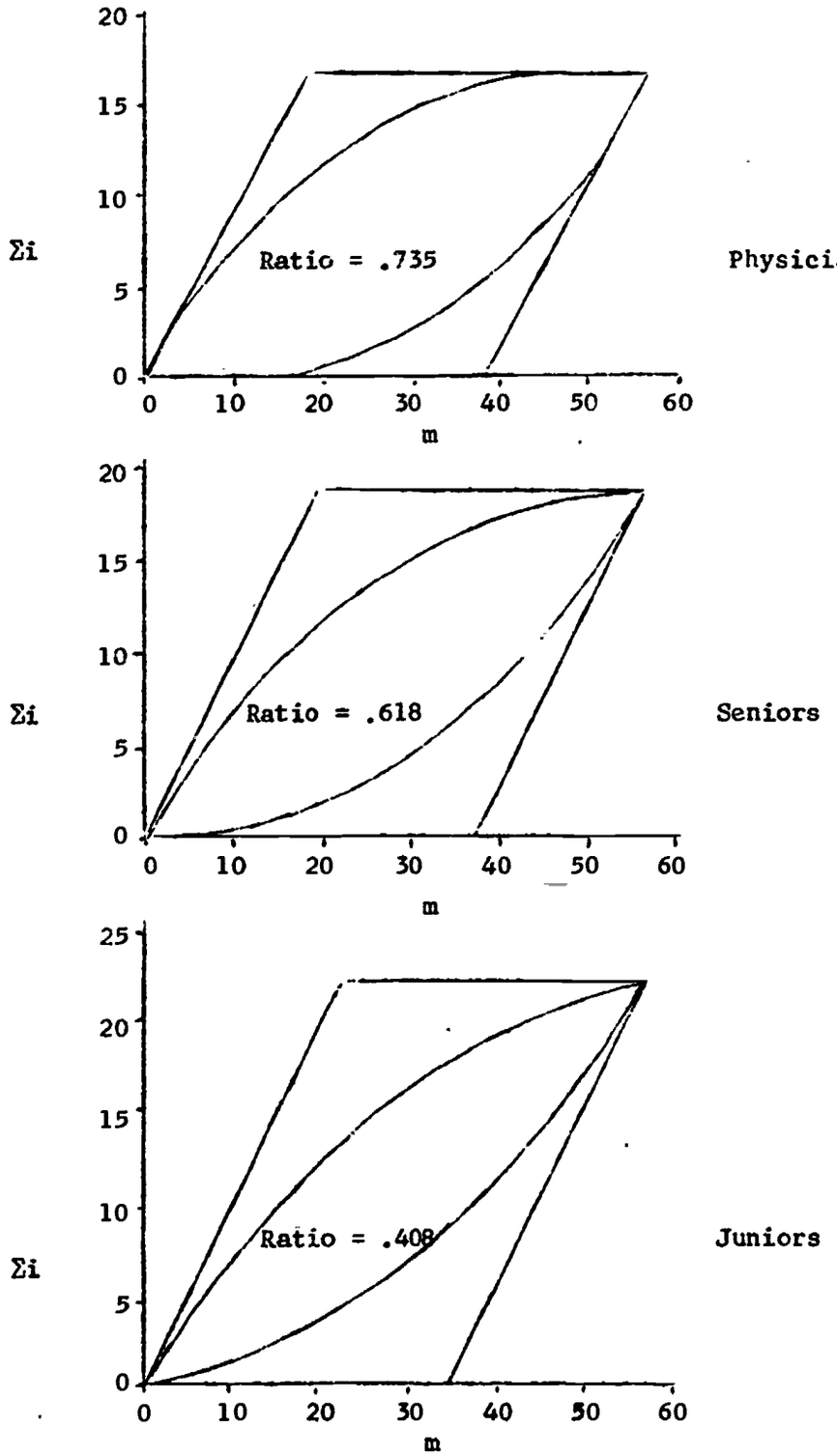
Figures 4 and 5 present the limits of performance for the cross sectional and longitudinal groups respectively. By inspecting these graphs it can be seen that the area becomes larger with increased training. The value of "b" was highest for physicians and lowest for juniors. It can be concluded from these results that the tests were more difficult for juniors than seniors and least difficult for physicians.

The ratios between limits of performance and the corresponding parallelograms are presented in Table III. These values are greatest for physicians and smallest for juniors in both tests. This also holds true for the longitudinal study where the values were higher for the second administrations.

Since these ratios increase when there is more group agreement, it can be stated that the most agreement was found among the physicians while the juniors agreed the least about which questions were more useful in reaching a diagnosis.

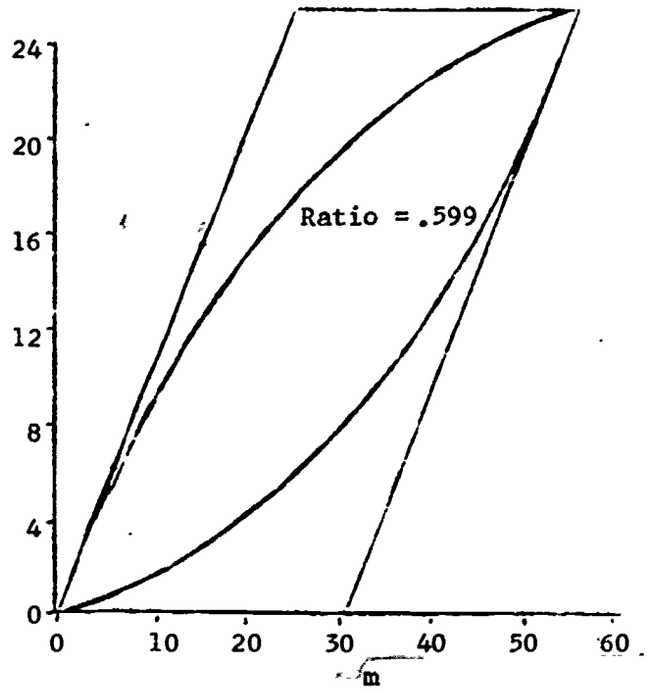
Utility Score

Averaging the utility indexes of all the questions asked by a particular subject will give the utility score. One interpretation that may be given to the utility score is the agreement of a subject with his group, since the utility score will be high when the questions asked by an individual are frequently asked by the group, and low when the questions are seldom asked by the group. In this analysis, the subjects were scored in terms of physicians'

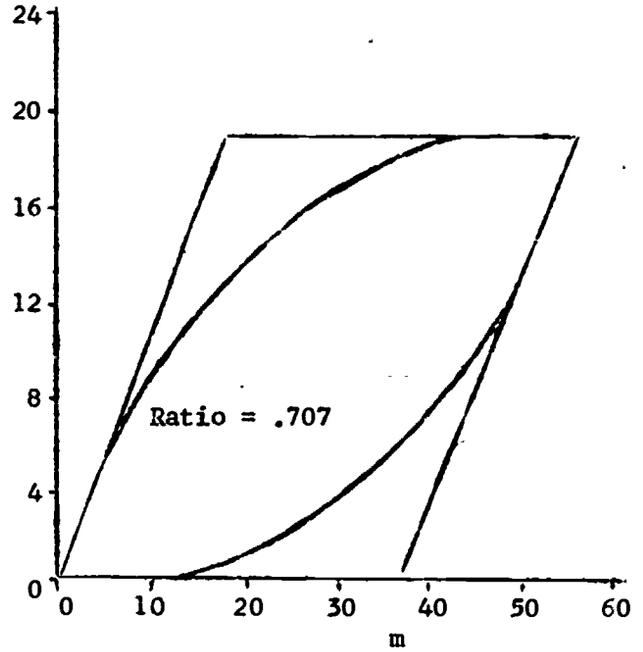


Limits of Performance for Three Groups on the Test of Diagnostic Skills, Test 4.

Figure IV



First Administration



Second Administration

Limits of Performance for the First and Second Administration of Test 4, Total.

Figure V

Table III

Ratios of Obtained Areas to
Maximum Possible Areas

Group	Test 2	Test 4
Juniors	.669	.408
Seniors	.787	.618
Physicians	.820	.735
First Administration	.649	.599
Second Administration	.737	.707

utility indexes. The utility scores, then, can be considered as a measure of agreement with the group of physicians. (Rimoldi, 1955).

Tables IV and V present the mean utility score of each group in each part and in the total for Test 2 and Test 4 respectively. In the total test, the junior mean score was lower than the senior mean in both studies, and the mean for the physicians was the highest ($p < .01$). This is also true in Part I of both tests. It can be concluded from these results that with increased training the subjects tend to ask questions which the physicians consider to be more important in reaching a diagnosis.

Pattern Analysis

The method of pattern analysis was devised in order to characterize patterns of responses. The development of this method has been described in several articles (Rimoldi, Grib, 1960a, 1960b). Essentially this is a descriptive measure which indicates the degree to which a given group agrees with a criterion group. Indexes of agreement were computed for 91 juniors and 143 seniors who took Test 4. The criterion group included 40 physicians. The seniors' pattern agreed more with the physicians' pattern than the juniors' did ($p = .01$). (Rimoldi, Haley, 1962a).

Sequential Evaluation of the Diagnostic Process

From the performance of a group of surgeons who took Test 2 and Test 4, norms were developed based on the proportion of times that each question was asked by the surgeons in every possible order. In the same manner norms were also developed from a group of clinicians. Each question may have different values depending upon the position in the sequence in which it was asked by the criterion groups. The performance of the junior and senior medical students was scored in terms of these norms. (Rimoldi, Haley, 1962b).

Figures 5 and 6 show the cumulative scores that represent the final values of performance curves for Test 2 and Test 4 respectively of all the students scored in terms of the two norms. All of the points that fall below the diagonal line correspond to those subjects who obtained higher values when scored in terms of surgeons' norms. Those above the diagonal scored higher in terms of clinicians' norms. There is a consistent trend for the students scores to be on the surgeons' side in Test 2 and on the clinicians' side in Test 4. This has bearing on the content of the two tests. Test 2 is a case of surgical pathology while Test 4, although a surgical case, presents complex symptamatology that may not be directly interpreted as being predominantly surgical.

The foregoing discussion has been a very brief summary of some of the research that has been performed on the Test of Diagnostic Skills. Although

Table IV

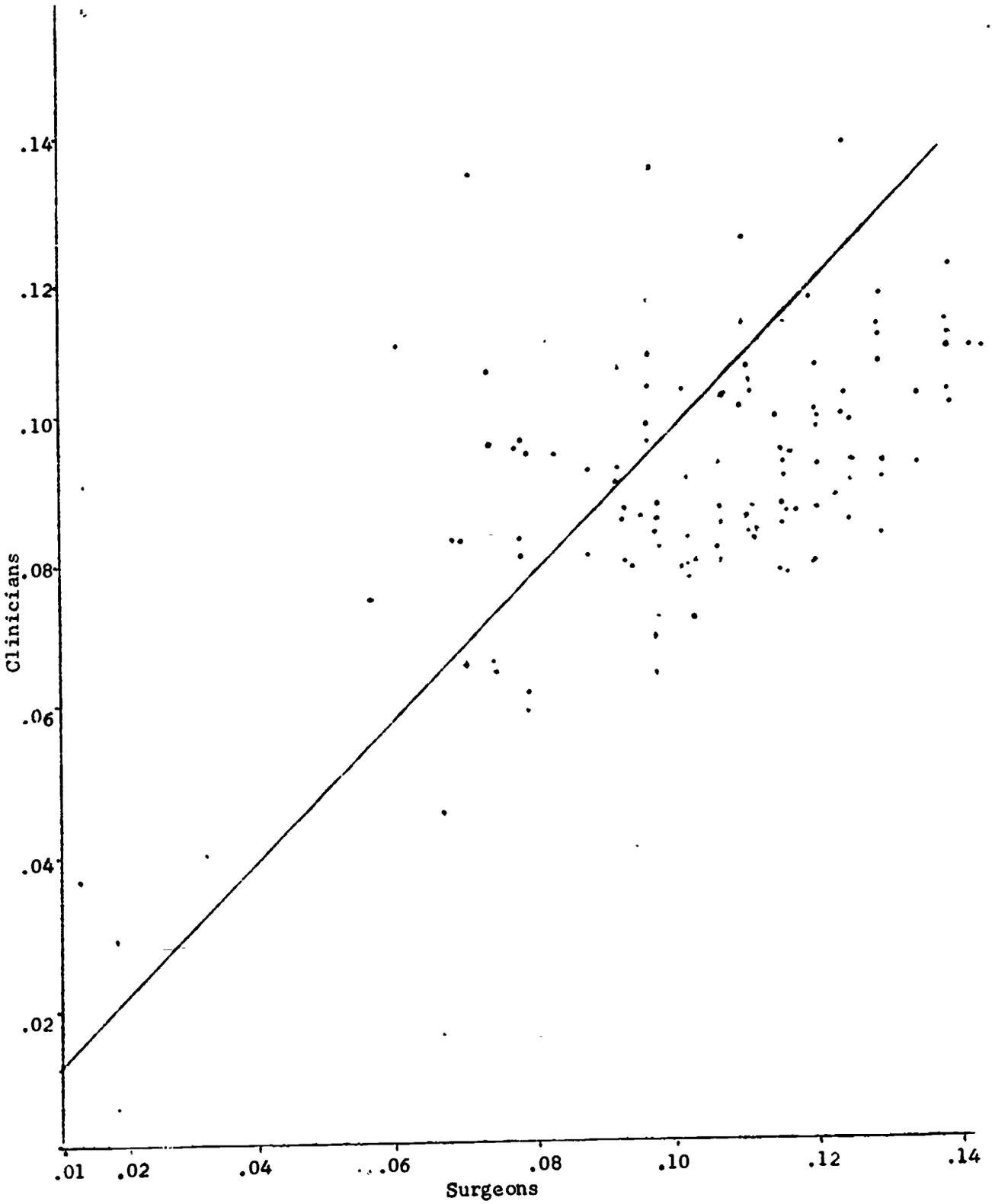
Means of Utility Scores
in Each Part and in the Total
for Test 2

Group	N	Part I	Part II	Part III	Total
Juniors	87	.48	.54	.67	.53
Seniors	129	.53	.55	.67	.57
Physicians	41	.58	.56	.64	.58
First Administration	36	.43	.50	.64	.49
Second Administration	36	.47	.53	.65	.53

Table V

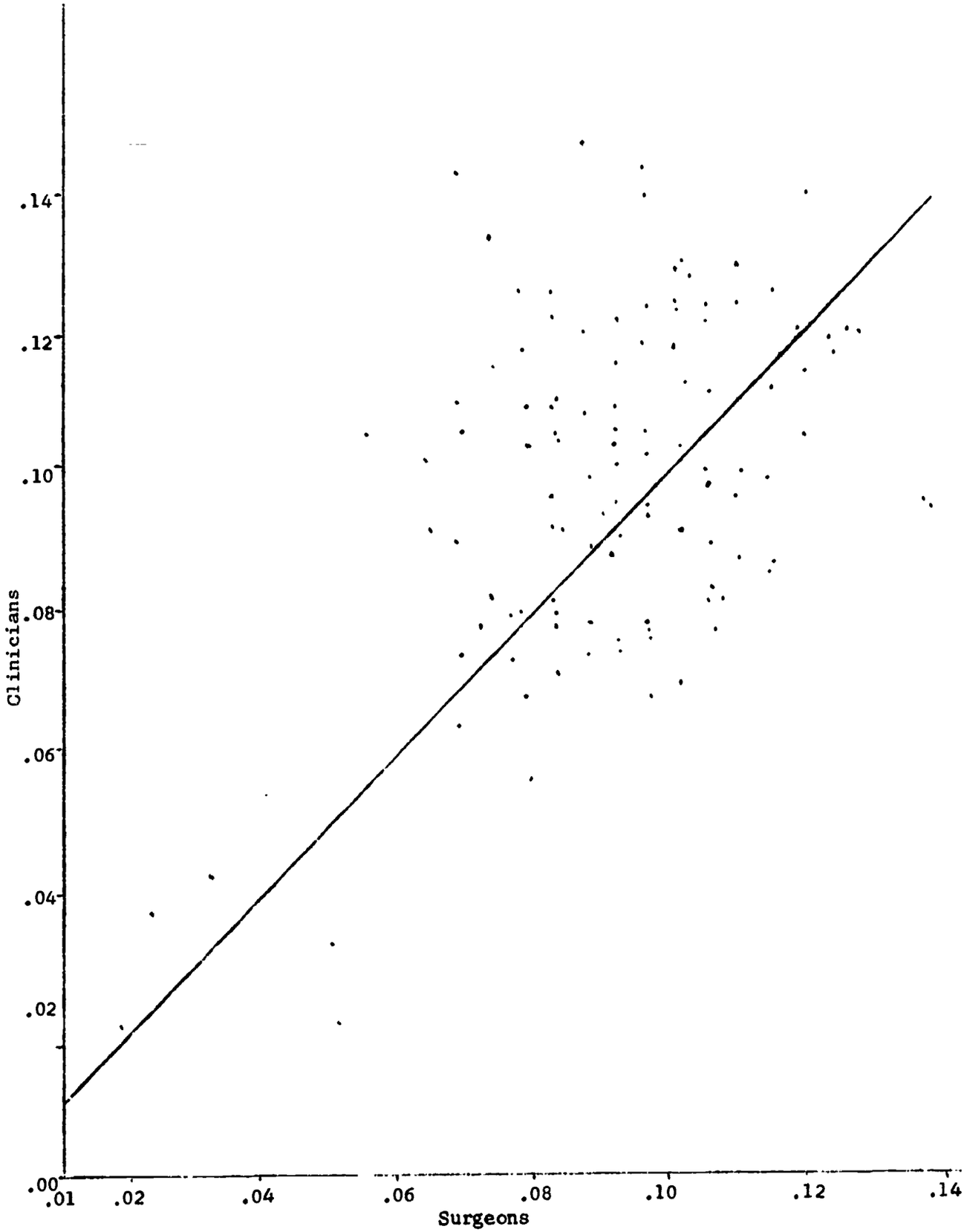
Means of Utility Scores
in Each Part and in the Total
for Test 4

Group	N	Part I	Part II	Part III	Total
Juniors	87	.44	.57	.59	.50
Seniors	147	.52	.61	.59	.56
Physicians	40	.56	.62	.59	.56
First Administration	36	.42	.56	.56	.48
Second Administration	36	.49	.63	.59	.54



Cumulative Scores Representing Final Values of the Performance Curves for Test 2 of Students Scored in Terms of Surgeons' Norms and in Terms of Clinicians' Norms.

Figure VI



Cumulative Scores Representing the Final Values of Performance Curves for Test 4 of Students Scored in Terms of Surgeons' Norms and in Terms of Clinicians' Norms.

Figure VII

some of the scoring methods described seem rather crude when compared to those developed more recently, nevertheless, they indicate that this technique is capable of measuring changes that take place during training. These results also suggest aspects of the diagnostic process that might be emphasized more during medical education. In fact the technique itself may be easily adapted as a training devise.

BIBLIOGRAPHY

1. Haley, J. V., The Effect of Learning on Performance in the Test of Diagnostic Skills. Loyola Psychometric Laboratory Publication No. 11, Loyola University, Chicago, Illinois, Feb. 1960.

2. Rimoldi, H. J. A., A Technique for the Study of Problem Solving. Educational and Psychological Measurement, 1955, 15, 4, 450-461.

3. Rimoldi, H. J. A., Devane, J., and Haley, J. V., Characterization of Processes. Loyola Psychometric Laboratory Publication No. 8, Loyola University, Chicago, Illinois, 1959. Educational and Psychological Measurement, 1961.

4. Rimoldi, H. J. A., and Grib, T. F., Pattern Analysis. Loyola Psychometric Laboratory Publication No. 7, Loyola University, Chicago, Illinois, 1958. British Journal of Statistical Psychology, Nov. 1960a.

5. Rimoldi, H. J. A., and Grib, T. F., Some Properties and Applications of Pattern Analysis. Loyola Psychometric Laboratory Publication No. 14, Loyola University, Chicago, Illinois, 1960b.

6. Rimoldi, H. J. A., and Haley, J. V., Determining Significance Levels in Pattern Analysis. Loyola Psychometric Laboratory Publication No. 23. Psychological Reports, 10, 500, 1962a.

7. Rimoldi, H. J. A., and Haley, J. V., Sequential Evaluation of Problem Solving Processes. Loyola Psychometric Laboratory Publication No. 22, Loyola University, Chicago, Illinois, 1962b.