A Meta-Analytic Evaluation of an Interpersonal Skills Curriculum: Accumulating Evidence Over Successive Occasions

Fredric M. Wolf
University of Michigan
School of Medicine

Mark L. Savickas, Glenn A. Saltzman, Martha L. Walker
Northeastern Ohio Universities
College of Medicine

Running Head: Accumulating Evidence

Prepared for presentation at the meeting of the American Educational Research Association, Montreal, April 1983

Printed in U.S.A.

A Meta-Analytic Evaluation of an Interpersonal Skills Curriculum: Accumulating Evidence Over Successive Occasions

Abstract

Results of individual evaluations of four successive classes of medical students' performance in a communication and interviewing skills curriculum were quantitatively synthesized using combined tests and measures of effect size typically used in literary meta-analytic reviews. The basic conclusion was that this curriculum produced gains on two standardized measures that were large in magnitude. An average student improved 1.37 standard deviation units on the Discrimination Index and 2.55 standard deviation units on the Communication Index from pre- to posttesting. Larger effects were associated with both earlier graduating classes and traditional students (vs. students in a combined six year B.S./M.D. program). Gender did not mediate the effect of the training.

A Meta-Analytic Evaluation of an Interpersonal Skills Curriculum: Accumulating Evidence over Successive Occasions

Methodology typically used in meta-analysis to integrate the results of independent tests of the same hypothesis in reviews of research literature (Glass, 1976, 1978; Glass, McGaw & Smith, 1981) are also appropriate for program evaluation in certain situations. One such situation has been referred to as the "independent samples/similar subjects - succesive occasions" case (Wolf, 1982). While measures of student learning at the conclusion of a course are helpful indexes of the impact of a program and its relative strengths and weaknesses vis-a-vis student performance, results may vary from term to term or year to year. Making curricular decisions based on data from any one term/year may not necessarily be representative of the results for other terms/years. Accumulating evidence over successive presentations of a curriculum would likely provide a more stable and generalizable assessment of both the direction and magnitude of impact.

Often the novelty and excitment of a new curricular effort generates increased interest and motivation on the part of instructors that may influence both the success of the program and student performance. Instructors, content, and/or the characteristics of the students may change from year to year. Systematically accumulating evidence over time would help control for such differences, as well as permit a comparison of differences on measures of student performance (given similar outcome measures across different occasions). Additionally, in programs in which small numbers of students participate at any given time, sample sizes may not be sufficiently large to make valid inferences. Using meta-analytic procedures can help to mitigate this difficulty by synthesizing data over successive occasions.

Goals of the nine-week course in interpersonal skills (interviewing and communication) evaluated in the present study included increasing students' (a) awareness of the physician's personal impact on his/her patient and the healing process and the reciprocity of the physician/patient relationship, (b) skill in establishing a trust relationship with patients, (c) skill in facilitating patient self-exploration and subsequently his/her own understanding of how the patient is experiencing the problem, and (d) skill in providing information, reassurance, support, and direction for the patient (Engler et al, 1981).

These goals can be summarized as increasing the value students place on emotional proximity and sensitivity to barriers in communication. It was hoped that each student would be able to respond accurately to both the feeling and meaning expressed by "patients" by the end of this program. Students were taught to discriminate among three classes of verbal behavior: initiating behavior from the physician's frame of reference, responding to the patient's experience (patient's frame of reference), and helping patients explore their own feelings (and for the physician to be aware of his/her own feelings). The differential application of the classes of verbalizations is demonstrated as the basis of a reciprocal relationship in which the patient feels valued by the physician. This study was conducted to synthesize the results of the initial experiences of the first four years of this new program. Thus, the purpose of this study was (a) to evaluate the effectiveness of an interpersonal skills course for first-year medical students by (b) using methods of meta-analysis (typically used in syntheses of research literature) to summarize results for four successive medical school classes.

Methodology

Instrumentation and Sample

Carkhuff's (1969a) Standard Indexes of Discrimination (DI) and Communication (CI) were administered to first-year medical students as part of student evaluation in a nine-week course in interviewing and communication skills (Engler et al, 1981; Saltzman et al, 1981). Scores on the Discrimination Index are determined by taking the average of the absolute difference between students and experts' ratings of 64 "typical" physician responses to 16 patient statements. The Communication Index requests students to respond to these same 16 patient statements (i.e., open format), and is scored with a 9-point scale ranging from level 1, "feeling and meaning both absent or both inaccurate," to level 5, "accurate response to personalized feeling and personalized goal and accurate identification of initial step" in treatment (Carkhuff, 1969a). Carkhuff (1969b) suggested that "(a) final functioning at level 2.5 or above, or (b) training gains of three-fourths of a level or more were reasonable goals for a successful training program."

Data was obtained from students who participated during the first four years of the interpersonal skills curriculum. Sample sizes for the classes were 46 (graduation class of 1981), 43 (1982), 42 (1983), and 72 (class of 1984). Students completed the CI and DI before participating in the interviewing and communication skills course and again after completing the course. Characteristics of the participants also were examined as possible mediators of the effects of training. These included gender, entry status (approximately two-thirds of each class are admitted into a combined six year B.S./M.D. program, while the remainder are traditional students), and graduation class.

Design and Analyses

A pretest-posttest pre-experimental design was used to evaluate the efficacy of this training program for each class. Glass (1978, p. 356) noted that this design may be considered primitive yet "adequate if the treated group members' pretreatment status is a good estimate of their hypothetical post-treatment status in the absence of treatment." This is an empirical question that can be examined to determine if maturation, pre-test sensitization or other threats to the validity of this design have in fact biased this estimate. Kraemer and Andrews (1982) noted that the effects resulting from a pre-post design would be equal to that from an experimental-control design only "if one has prior certainty of the absence of time effects and of placebo effects" (p. 407). In the present use of the pre-post design in this evaluation, several recent studies support the validity of this design. McPherson, Wolf, and Sachs (1983) found significant differences on the Carkhuff measure favoring a group which experienced skills training versus a group which received information didactically. The skills training group did improve significantly from pre- to posttesting, while the didactic group did not. Thus, no placebo effect resulting from mere exposure to this content was anticipated nor found. Similarly, in another study using the Carkhuff measures (McPherson, Knopp, Sachs & Wolf, 1983), results of a randomized pretest-posttest experimental-control group design indicated significant improvement for the experimental versus the control group. Thus, time nor maturation alone accounted for this effect and no spontaneous improvement was found nor anticipated. Both of these studies support the utility of the pretest-posttest design used in the present study. Indeed, Campbell (1982) indicated that the one group, pretest-posttest design has "now been elevated to a useful quasi-experimental or proto-experimental design" in the planned revision of his classic work on research design (Campbell & Stanley, 1963).

Results for each of the four independent classes experiencing the interpersonal skills curriculum were synthesized through the use of combined test (Fisher, 1932; Rosenthal, 1978; Winer, 1971) and effect size analyses (Cohen, 1977; Glass, 1976, 1978; Hedges, 1982; Rosenthal & Rubin, 1982a, 1982b). Wilcoxan Matched-Pairs Ranked-Signs Tests (Marasuilo & McSweeney, 1977) and dependent t-tests were used to examine changes in student performance on each of the Carkhuff indexes for each class separately.

Combined Probabilities

Statistical methods available for combining the results of independent studies addressing a common research question range from various counting procedures to a variety of summation procedures involving either significance levels (probabilities or their logarithmic transformations) or raw or weighted test statistics such as t's or z's. These later procedures have become known as "combined tests" and were originally developed independently by R. A. Fisher (1932) and Karl Pearson (1933).

While a variety of combined tests are available, the suggestion to select a combined test statistic consistent with the statistics used in the independent tests (Wolf, 1982) for each class was followed in this study. Thus, the combined test offered by Winer (1971) for summing t's was used to synthesize the dependent t-test results for each outcome measure for each class. The Winer procedure for combining independent test results comes directly from the sampling distribution of independent t-statistics in which the t-statistics associated with each test are summed and divided by the square root of the sum of the degrees of freedom (df) associated with each t after each df has been divided by df-2. This is based on df/(df-2) being the variance of a t distribution, which is approximately normally distributed (N(0,1)) when df > 10. This may be expressed in the form of

$$z = \frac{\Sigma t}{\sqrt{df/(df-2)}} \qquad (1)$$

The Stouffer test (Stouffer, 1949; Mosteller & Bush, 1954; Rosenthal, 1978) for summing z's was used to synthesize results of the individual Wilcoxan analyses. It is similar to the Winer procedure with the exception that z's instead of t's are summed. The denominator then simplifies to the square root of the number of tests combined. This procedure is based on the sum of normal deviates being itself a normal deviate, with the variance equal to the number of observations (N) summed. The complete expression takes the form of

$$z = \frac{\Sigma z}{\sqrt{N}} \qquad (2)$$

Fail-Safe N or File-Drawer Problem

Rosenthal (1979) pointed out that published studies more often include results that are statistically significant than do unpublished studies. Thus, it is possible that results of the above combined tests may be biased in favor of significant probabilities resulting. It is therefore possible to estimate the number of studies confirming the null hypothesis that would be necessary to reverse the conclusion of a combined test that a significant effect or relationship exists. When the significance level is set so that p=.05, this fail-safe N, as Cooper (1979) referred to it, can be calculated using the following formula:

$$N_{fs.05} = \left(\frac{\Sigma z}{1.645}\right)^2 - N, \qquad (3)$$

where Σz = sum of the individual z-tests (or t-tests when the Winer procedure is used), N = number of studies combined, and 1.645 is the normal value (z) for p=.05. If p=.01, then 1.645 is replaced by 2.33. A large fail-safe N would suggest that we may place greater confidence in significant results of combined tests, as many additional studies with no effect would be needed to reverse the conclusion of significance. Conversely, a small fail-safe N would call into question the significance of obtained results.

Effect Size Estimation

Statistical tests such as the combined procedures previously described provide a summary index of the statistical significance of the results pertaining to an hypothesis. They do not, however, provide any insight into the strength of the relationship or effect of interest. The desirability of accompanying combined tests with indexes of effect size has been noted by Rosenthal (1978). Glass' exposition and application of meta-analysis relies heavily on the use of measures of effect size that have been eloquently summarized by Cohen (1977). Cohen states, "Without intending any necessary implication of causality, it is convenient to use the phrase 'effect size' to mean 'the degree to which the phenomenon is present in the population', or 'the degree to which the null hypothesis is false'. Whatever the manner of representation of a phenomenon in a particular research in the present treatment, the null hypothesis always means that the effect size is zero" (pp. 9-10).

The goal is to obtain "a pure number, one free of our original measurement unit, with which to index what can be alternatively called the degree of departure from the null hypothesis of the alternative hypothesis, or the ES (effect size) we wish to detect. This is accomplished by standardizing the raw effect size as expressed in the measurement unit of the dependent variable by dividing it by the (common) standard deviation of the measures in their respective populations, the latter also in the original measurement" (Cohen, 1977, p. 20). This may be accomplished in the form of

$$d= \frac{|\bar{x}_1 - \bar{x}_2|}{\sigma} \qquad (4)$$

where d = ES index for t-tests of means in standard unit, $\bar{x}_1$ and $\bar{x}_2$ = sample means in original measurement units, and σ = standard deviation of either sample (as homogeneity of variance is assumed). The means, $\bar{x}_1$ and $\bar{x}_2$, are typically the experimental and control group means in posttest-only control group experimental designs, or pre- and post means in one group pretest-posttest pre-experimental designs, as used in this study.

Once the effect size, d, is determined, Cohen provides tables to translate d into measures of nonoverlap (U) between the two groups, which translate rather nicely into graphical displays which facilitate interpretation of the results. Perhaps the most useful index of nonoverlap is Cohen's U_3, which translates average performance in percentiles (area under the normal curve) of the posttest (or experimental) group to the equivalent percentile of the pretest (or control) group.

Results and Discussion

Data were analyzed separately for each of the two criterion measures. This is consistent with some meta-analytic studies (e.g., Kulik, Kulik & Cohen, 1979; Mazzuca, 1982) but inconsistent with those that have combined all outcome measures in one analysis (e.g., Smith & Glass, 1977). The former approach was taken in that more precise information would be available for future curricular planning than if results for the two Carkhuff measures had been combined. It is possible that the training program may have influenced performance on the two measures differentially. This would be obscured in one larger analysis where effects might even cancel each other.

Overall Combined Results and Effect Sizes

Results for each of the four classes on the Carkhuff Discrimination Index are summarized in Table 1. Significant ($p < .001$) gains in performance from pre to posttesting were exhibited by each class, with paired t-tests ranging between 7.14 and 10.53. Results of the Winer combined test supported the research hypothesis of a significant gain in discrimination performance (i.e., more accurate discrimination) when the scope of the inference is with respect to the combined populations ($z=17.85$). The probability of obtaining this value of z or one larger is $p\ (z \geq 17.85) < .001$, one-tailed.

Insert Table 1 about here

Wilcoxan matched-pairs analyses reported in Table 2 were consistent with the paired t-test and Winer combined test results. A significant number of students in each class exhibited significant improvement ($p < .001$) with z's ranging between -4.76 and -6.75. The Fisher combined test also supported the research hypothesis, with the probability of obtaining this z value or one smaller being $p(z \leq -11.22) < .001$, one-tailed. Only 18 of 204 students across all four classes failed to improve on the Discrimination Index.

Insert Table 2 about here

Effect sizes in Table 1 ranged between 1.22 and 1.50 standard deviation units, with an average effect size of 1.37 (SD=.13). Cohen (1977) provides interpretative guidelines for effect size, with d=.2 indicative of a small effect, d=.5 indicative of a medium effect, and d=.8 indicative of a large effect. Each of the individual class effects, as well as the average effect, may be considered large in magnitude. Translating these effect sizes d into a measure of overlap (U) is accomplished by referring to tables in Cohen's text (1977).

Alternatively, a normal distribution table may be used as these values are equivalent to Cohen's U_3 tabled values. An average d value of 1.37 translates into a U_3 value of .915. This means that the average score (50th percentile) on the posttest was equivalent to the 91.5th percentile on the pretest. This is depicted graphically in Figure 1.

Insert Figure 1 about here

Results for performance on the Communication Index are summarized in Table 3. Students in each class again exhibited significant improvement ($p < .001$) with paired t-tests ranging between -8.55 and -24.18. Results of the Winer combined test were also significant and indicated the probability of obtaining this value of z (-28.51) or one smaller is $p < .001$, one-tailed. Results of the Wilcoxan tests summarized in Table 4 supported these results, as only 12 of 205 students failed to improve on the Communication Index. Results of the Stouffer combined test likewise supported the research hypothesis ($z=-11.03$; $p < .001$, one-tailed).

Insert Tables 3 and 4 about here

Effect sizes for the individual classes ranged between 1.48 and 3.52 standard deviation units, with an average effect size of 2.55 (SD=.81). Translating this average effect size (d) of 2.55 into the U_3 measure of non-overlap indicated that the average score (50th percentile) on the Communication posttest was equivalent to the 99.5th percentile on the pretest. The average student could expect to improve 2.55 standard deviation units as a result of this program. These results are depicted graphically in Figure 2.

Insert Figure 2 about here

Fail-Safe N Results

For results pertaining to the Discrimination Index, the number of tests supporting the null hypothesis necessary to reverse the findings reported above (i.e., to find a combined test result of $\underline{p} > .05$) was 488 if t-tests were used, or 182 studies if the more conservative Wilcoxan tests were used. For the Communication Index, 1,245 additional studies with null results would be needed to reverse the conclusion of a significant effect in the t-test analyses. Approximately 180 null results would be needed to reverse the Wilcoxan findings. Thus, the findings reported here appear to be robust and well above Rosenthal's (1979) "tolerance level" for null effects.

Analyses of Mediating Effects

Hypothesis tests were categorized according to the potential mediators of the effects of the interpersonal skills curriculum. Effect sizes were computed for each subgrouping of gender, entry status, and graduation year. Correlations were then computed between each potential mediator and the effect sizes for the pertinent subgroupings. Average effect sizes for the subgroupings are summarized in Table 5, except for graduation year which is included in Tables 1 and 3.

Gender. The average effect sizes on the Communication Index for males was 2.25 (SD=.80) and for females it was 2.06 (SD=1.47). The point-biserial correlation between gender and ES using ESs for individual classes for males and females was .18 (n.s.; n=6). Average effect sizes on the Discrimination Index for males was 1.40 (SD=.29) and for females it was 1.21 (SD=.65). Again the correlation was non-significant (r_{pbi}=.08; n=6). Thus, gender does not appear to mediate the effect of the program, as no differences of significance between males and females were found.

Entry Status. Average effect sizes on the Communication Index for students in the combined six-year B.S./M.D. program and for traditional students were 1.91 (SD=1.03) and 3.10 (SD=1.33) standard deviation units, respectively. Average effect sizes on the Discrimination Index were 1.12 (SD=.16) for B.S./M.D. students and 1.64 (SD=.12) for traditional students. Results of point-biserial correlational analyses indicated that the effects of the curriculum were greater for traditional students on both the Communication (r_{pbi}=.45; n=6; $p < .10$, two-tailed) and Discrimination (r_{pbi}=.88; n=6, $p < .05$) Indexes. Further examination of pre- and posttest average scores indicated that this finding is the result of the B.S./M.D. students entering the program with better skills. The training program acted as a "leveler", as there were no differences between B.S./M.D. and traditional students at posttesting (independent t-tests ranged between -0.92 and 1.19, n.s.). On the pretest, however, B.S./M.D. students in each class consistently performed significantly better than their traditional cohorts on the Communication Index (independent t-tests ranged between 2.38 and 3.18, $p < .03$) and on the Discrimination Index (class of 1984 only). This is likely the result of a course the B.S./M.D. students receive prior to this interpersonal skills program.

Graduating Class. Year of graduation and the effect size for each year, summarized for the Discrimination Index in Table 1 and for the Communication Index in Table 3, were correlated to test the research hypotheses of significant declines in performance for more recent classes. Correlations were -.91 (n=4; $p < .05$, one-tailed) for the Discrimination Index and -.98 (n=4; $p < .01$, one-tailed) for the Communication Index. These results indicate almost a perfect negative relationship between recency of graduation and performance on the two Indexes.

It is noteworthy, however, that the effects of the program for each of the classes may be considered large based on Cohen's (1977) criteria. Each class attained Carkhuff's (1969b) two criteria for a successful training program, training gains of three-fourths of a level and final communication scores of 2.5 or above. There are several plausible rival explanations for this decreasing trend between effect size and recency of year of graduation. Because the class of 1981 was the first graduating class at this new medical school, perhaps greater care was taken in selection procedures. Thus, admission policy may have changed over the course of these four years. Larger class sizes may be a factor. However, because this course is taught in small groups of 10-12 students, this would most likely not be an influence unless the addition of new instructors affected the quality of instruction. Finally, some changes in the curriculum may have occured over this period. Thus, an exmaination of the stability and change in (a) admission standards, (b) class size, (c) instructors or their performance, (d) curriculum, and (e) other student characteristics is necessary to more fully understand the meaning of the relationship between graduation year and effect size.

Conclusions

Even though student performance on the CI and DI increased significantly for each of the medical school classes, the magnitude of the effect varied. Using the Winer and Stouffer combined tests and Cohen's measure of effect size provide more stable summary indexes of impact of the course on student performance. As each successive class completes the course, data may be added to previous years' results and the Winer, Stouffer, and Cohen statistics may be recalculated. Thus, the curriculum may be compared vis-a-vis these student performance measures from year to year, as well as over all the years it is offered. Combining and synthesizing these individual class findings permits greater generalizability and confidence in the evaluation results of the program than do individual results based upon smaller sample sizes. As data accumulate, trends in student performance may be noted that have implications for curricular planning and development. Two such mediating relationships were found in the present study. First, tranditional students' learning gains as evidenced on the Carkhuff measures were superior to gains of the six year combined B.S./M.D. students. This was a result of B.S./M.D.

students entering the program with greater skills, most likely as a result of educational training earlier in their program that is related to the content of this course. Secondly, the effect of training, while significant and large in magnitude for each class, appears to be declining with each succesive class. This trend merits closer examination and understanding.

Evidence was cited from other studies to support the validity of the findings of the pretest-posttest design used in the present study. It is unlikely that the large effects of this curriculum were the result of maturation, time, spontaneous improvement, or a placebo effect of merely attending didactic sessions. Over 180 additional studies with no effect would be necessary to reverse the conclusion of a significant effect of training. Clearly, the program did have a significant positive impact on first year medical students' communication and interviewing skills. Whether these skills become more fully developed, refined, and eventually used in interactions with patients in the future are important issues that merit examination.

References

Campbell, D.T. Can we be scientific about policy research? Award address presented at the meeting of the American Educational Research Association, New York, March 1982.

Campbell, D.T. & Stanley, J.C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Carkhuff, R.R. Helping and human relations: A primer for lay and professional helpers (Vol. 1). New York: Holt, Rinehard & Winston, 1969. (a)

Carkhuff, R.R. The prediction of the effects of teacher-counselor education: The development of communication and discrimination selection indexes. Counselor Education and Supervision, 1969, 8, 265-272. (b)

Cohen, J. Statistical power analysis for the behavioral sciences (Rev. ed.), New York: Academic Press, 1977.

Cooper, H.M. Statistically combining independent studies: A meta-analysis of sex differences in conformity research. Journal of Personality and Social Psychology, 1979, 37, 131-146.

Engler, C.M., Saltzman, G.A., Walker, M.L. & Wolf, F.M. Medical student acquisition and retention of communication and interviewing skills. Journal of Medical Education, 1981, 56, 572-579.

Fisher, R.A. Statistical methods for research workers (4th ed.). London: Oliver and Boyd, 1932, pp. 99-101.

Glass, G.V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5, 3-8.

Glass, G.V. Integrating findings: The meta-analysis of research. In L.S. Shulman (Ed.) Review of Research in Education (Vol. 5). Itasca, Illinois: F.E. Peacock, 1978.

Glass, G.V., McGaw, B. & Smith, M.L. Meta-Analysis in social research. Beverly Hills, CA.: Sage, 1981.

Hedges, L.V. Estimation of effect size from a series of independent experiments. Psychological Bulletin, 1982, 92, 490-499.

Kraemer, H.C. & Andrews, G. A nonparametric technique for meta-analysis effect size calculation. Psychological Bulletin, 1982, 91, 404-412.

Kulik, J.A., Kulik, C.C., & Cohen, P.A. A meta-analysis of outcome studies of Keller's personalized system of instruction. American Psychologist, 1979, 34, 307-318.

Marasuilo, L.A. & McSweeney, M. Nonparametric and distribution-free methods for the social sciences. Monterrey, CA: Brooks/Cole, 1977.

Mazzuca, S.A. Does patient education in chronic disease have therapeutic value? Journal of Chronic Disease, 1982, 35, 521-529.

McPherson, C., Knopp, W., Sachs, L. & Wolf, F.M. The doctor-patient relationship: Systematic training in effective communication skills, 1983, submitted for publication.

McPherson, C., Wolf, F.M., & Sachs, L. Improving interviewing effectiveness: Positive awareness vs. skills training. Psychological Reports, 1983, accepted for publication.

Mosteller, F.M. & Bush, R.R. Selected quantitative techniques. In Handbook of social psychology: Vol. 1. Theory and method, G. Lindzey (ed.). Cambridge, Mass.: Addison-Wesley, 1954.

Pearson, K. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random, Biometrika, 1933, 25, 379-410.

Rosenthal, R. Combining results of independent studies. Psychological Bulletin, 1978, 85, 185-193.

Rosenthal, R. The "file drawer problem" and tolerance for null results. Psychological Bulletin, 1979, 86, 638-641.

Rosenthal, R. & Rubin, D.B. Comparing effect sizes of independent studies. Psychological Bulletin, 1982, 92, 500-504(a).

Rosenthal, R. & Rubin, D.B. Further meta-analytic procedures for assessing cognitive gender differences. Journal of Educational Psychology, 1982, 74, 708-712(b).

Saltzman, G.A., Wolf, F.M., Savickas, M.L. & Walker, M.L. Dogmatic thinking and communication skills of student physicians. Psychological Reports, 1981, 48, 853-854.

Smith, M.L. & Glass, G.V. Meta-analysis of psychotherapy outcome studies. American Psychologist, 1977, 32, 752-760.

Stouffer, S.A. et. al. The american soldier: Vol. 1. Adjustment during army life. Princeton: Princeton University Press, 1949, p. 45, footnote 15.

Winer, B.J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971, pp. 49-50.

Wolf, F.M. Meta-analytic applications in program evaluation. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C., August 1982. (ERIC Document Reproduction No. ED)

Table 1

Means, Standard Deviations, Paired t-Tests and Effect Sizes for Medical Student Performance on Pre and Post Standard Indexes of Discrimination

		Pre		Post		Paired		
Graduation Year	n	M	Sd	M	Sd	t	d	U_3(%)
1981	46	.99	.23	.65	.17	8.95*	1.48	93.1
1982	43	.95	.20	.65	.15	9.86*	1.50	93.3
1983	42	1.00	.23	.71	.16	7.14*	1.26	89.6
1984	73	1.04	.27	.71	.18	10.53*	1.22	88.9
Average							1.37	91.5

*p <.001, two-tailed test

Table 2

Wilcoxon Matched - Pairs Test for Change in Performance
on Standard Index of Discrimination

		Number of Students			
Graduation Year	n	Declined	Same	Improved	z
1981	46	3	0	43	-5.34*
1982	43	2	0	41	-5.58*
1983	42	4	0	38	-4.76*
1984	73	9	0	64	-6.75*

*$p < .001$, two-tailed test

Table 3

Means, Standard Deviations, Paired t-Tests and Effect Sizes for Medical Student Performance on Pre and Post Standard Indexes of Communication

		Pre		Post				
Graduation Year	n	M	Sd	M	Sd	t	d	U_3(%)
1981	46	1.55	.30	2.60	.22	-24.18*	3.52	99.9
1982	44	1.32	.39	2.54	.48	-14.16*	3.12	99.9
1983	42	1.47	.52	2.55	.59	-8.55*	2.07	98.0
1984	73	1.73	.50	2.47	.29	-11.28*	1.48	93.1
Average							2.55	99.5

*p <.001, two-tailed test

Table 4

Wilcoxon Matched - Pairs Test for Change in Performance on Standard Index of Communication

		Number of Students			
Graduation Year	n	Declined	Same	Improved	z
1981	46	0	0	46	-5.91
1982	44	1	1	42	-5.69*
1983	42	2	1	39	-4.71*
1984	73	7	0	63	-5.75*

*$p < .001$, two-tailed test

Table 5

Average Effect Sizes for Subgroupings of Study Characteristics for Communication and Discrimination Indexes

Characteristics	Communication		Discrimination		N
	$\bar{x}_d$	SD_d	$\bar{x}_d$	SD_d	
Gender					
Males	2.25	.80	1.40	.29	3
Females	2.06	1.47	1.21	.65	3
Entry Status					
Combined B.S./M.D.	1.91	1.03	1.12	.16	3
Traditional	3.10	1.33	1.64	.12	3

Note: N is the number of studies on which the average effect size ($\bar{x}_d$) and SD_d are based.

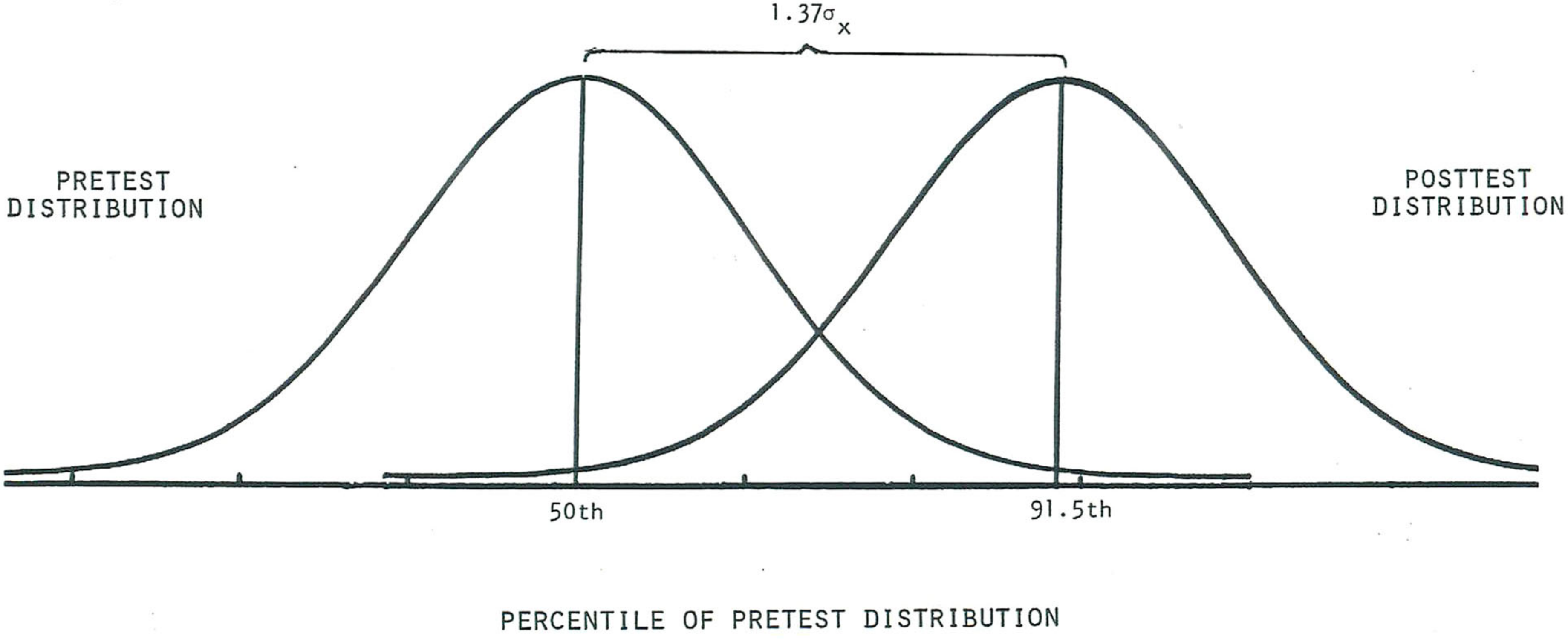
$1.37\sigma_x$
PRETEST DISTRIBUTION
POSTTEST DISTRIBUTION
50th
91.5th
PERCENTILE OF PRETEST DISTRIBUTION

FIGURE 1

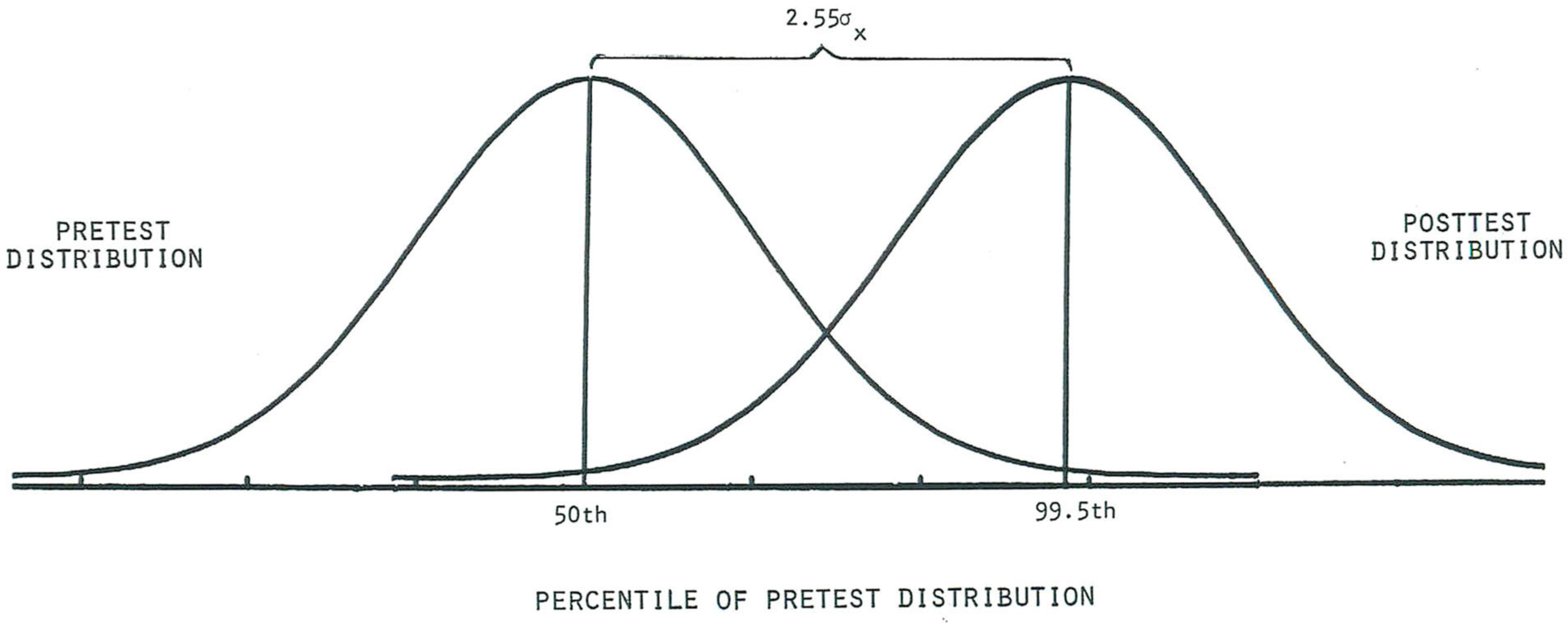
$2.55\sigma_x$
PRETEST DISTRIBUTION
POSTTEST DISTRIBUTION
50th
99.5th
PERCENTILE OF PRETEST DISTRIBUTION

FIGURE 2