Journal of Counseling Psychology
1984, Vol. 31, No. 2, 253-257

Meta-Analytic Evaluation of an Interpersonal Skills Curriculum for Medical Students: Synthesizing Evidence Over Successive Occasions

Fredric M. Wolf
Medical School, University of Michigan (Ann Arbor)

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

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. Larger effects were associated with earlier graduating classes. Sex and entry status did not mediate the effect of the training. Implications of the viability of pretest–posttest designs for evaluation research under certain conditions and of the synthesis of evaluation results over successive offerings of a program are discussed. As data for successive implementations of the program accumulate, trends in student performance may be noted that have implications for curricular planning and development. It is hoped that the study provides one desirable model for approaching the evaluation of counseling psychology training programs.

Methods typically used in meta-analysis to integrate the results of independent tests of the same hypothesis in reviews of research literature (Glass, 1978) are also appropriate for program evaluation in certain situations. One such situation has been referred to as the "independent samples/similar subjects—successive occasions" case (Wolf, 1982). Whereas 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-à-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 excitement 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 permits an examination of such differences. 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. The use of meta-analytic procedures can help to mitigate this difficulty by synthesizing data over successive occasions.

Among the objectives of the 9-week course in interpersonal skills (interviewing and communication) evaluated in the present study was to train students 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 (while being aware of his or her own feelings as the physician). It was hoped that each student would be able to respond accurately to both the feeling and the meaning expressed by patients by the end of this program. Students were assigned partners (within small groups of six) with whom they practiced communication skills and videotaped two interviews with each other outside class. These videotapes were then viewed selectively and discussed in the small groups under the supervision of a behavioral scientist. The behavioral scientists included psychologists, social workers, nurses, and physicians,

---

An earlier version of this article was presented at the meeting of the American Educational Research Association, Montreal, Canada, April 1983.

We would like to thank Charles J. Gelso and an anonymous referee for their helpful comments.

Requests for reprints should be sent to Fredric M. Wolf, Department of Postgraduate Medicine and Health Professions Education, University of Michigan, G1208 Towsley Center, Ann Arbor, Michigan 48109.

most of whom were faculty members in the Department of Family Medicine. This study was conducted to synthesize the evaluation results of the initial experiences of the first 4 years of this new program.

Method

Instruments and Sample

Carkhuff's (1969a) Standard Indexes of Discrimination (DI) and Communication (CI) were administered to first-year medical students as part of the student evaluation of an interviewing and communication skills course (Engler, Saltzman, Walker, & Wolf, 1981). Scores on the DI 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 CI requests students to respond to these same 16 patient statements (i.e., open format) and is scored with a 9-point scale ranging from *feeling and meaning both absent or both inaccurate* (Level 1) to *accurate response to personalized feeling and personalized goal and accurate identification of initial step* (Level 5) 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" (p. 269).

Samples sizes for the four classes were 46 (graduation class of 1981), 43 (1982), 42 (1983), and 72 (1984). Students completed the CI and DI before training in interviewing and communication skills and again after completing training. Characteristics of the participants also were examined as possible mediators of the effects of training. These included sex, entry status (approximately two thirds of each class are admitted into a combined 6-year Bachelor of Science/Medical Doctoral [BS/MD] program), and graduation class.

Design and Analyses

A pretest–posttest preexperimental design was used to evaluate the efficacy of this training program for each class. Glass (1978) noted that this design may be considered primitive yet "adequate *if* the treated group members' pretreatment status is a good estimate of their hypothetical posttreatment status in the absence of treatment" (p. 356). This is an empirical question that can be examined to determine if maturation, pretest 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 pretest–posttest 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). Several recent studies have supported the validity of this design in the present study. McPherson, Wolf, and Sachs (1983) found significant differences on the Carkhuff measure favoring a group that experienced skills training versus a group that received information didactically. The skills-training group improved significantly from pre- to posttesting, whereas the didactic group did not. Thus, no placebo effect resulting from mere exposure to this content was anticipated nor found. Similarly, unpublished results in another study using the Carkhuff DI (McPherson et al., in press) indicated significant improvement for the experimental versus control group. Thus, neither 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. Oetting (1982) pointed out that controls are more important for scientific inquiry in which the purpose is to understand why something occurs as opposed to evaluation research in which the purpose is to determine whether something has occurred.

Results for each of the four independent classes experiencing the interpersonal skills curriculum were synthesized through the use of combined test (Rosenthal, 1978; Winer, 1971) and effect size analyses (Cohen, 1977; Glass, 1978). Wilcoxon matched-pairs ranked-signs tests and dependent t tests were used to examine changes in student performance on each of the Carkhuff indexes for each independent class.

Combined Probabilities and Effect Size Estimation

Although a variety of tests for combining the results of independent studies addressing a common research question 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 scores was used to synthesize the dependent t-test results for each outcome measure for each class. The Stouffer test (Rosenthal, 1978) for summing z scores was used to synthesize results of the individual Wilcoxon analyses.

Statistical tests such as the combined procedures previously described provide a summary index of the statistical significance of the results pertaining to a 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's (1978) exposition and application of meta-analysis relied heavily on the use of measures of effect size that have been eloquently summarized by Cohen (1977). The effect size index d for t tests of means in standard deviation units was used in the present study. Once this effect size was determined, tables provided by Cohen (1977) were used to translate d into measures of nonoverlap (U) between the two groups. 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

Results on the Carkhuff DI indicating 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 (1971) combined test supported the research hypothesis of a significant gain in DI performance (i.e., more accurate discrimination) when the scope of the inference was with respect to the combined populations ($Z_c = 17.85$). The probability of obtaining this value of z or one larger is p ($Z_c \geq 17.85$) $< .001$, one-tailed. Wilcoxon matched-pairs analyses were consistent with the paired t test and Winer combined test results. A significant number of students in each class exhibited improvement ($p < .001$), with zs ranging between -4.76 and -6.75. The Stouffer (Rosenthal, 1978) combined test also supported the research hypothesis; the probability of obtaining this z value or one smaller was p ($Z_c \leq -11.22$) $< .001$, one-tailed. Only 18 of 204 students across all four classes failed to improve on the DI. Effect sizes for performance on the DI ranged between 1.22 and 1.50 standard deviation units, with an average effect size of 1.37 ($SD = 0.13$). 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.

Results for performance on the CI indicated that 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 or one smaller was $p(Z_c \leq -28.51) < .001$, one-tailed. Results of the Wilcoxon tests supported these results, as only 15 of 205 students failed to improve on the CI. Results of the Stouffer combined test likewise supported the research hypothesis ($Z_c = -11.03, p < .001$, one-tailed). Effect sizes for the individual classes ranged between 1.48 and 3.52 standard deviation units, with an average effect size of 2.55 ($SD = 0.81$). Translating this average effect size (d) of 2.55 into the U_3 measure of nonoverlap indicated that the average score (50th percentile) on the CI 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.

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 sex, entry status, and graduation year. Correlations were then computed between each potential mediator and the effect sizes for the pertinent subgroupings.

Sex. The point-biserial correlation between sex and effect size on the CI using effect size for individual classes for males and females was .18 ($n = 8$, *ns*). The correlation between sex and effect size on the DI was likewise nonsignificant ($r_{pbis} = .04, n = 8$, *ns*).

Entry status. Results of point-biserial correlational analyses between effect sizes for students in the combined 6-year BS/MD program and traditional students indicated that whereas the effects of the curriculum were greater for traditional students on both the CI ($r_{pbis} = .44, n = 8, p < .10$, two-tailed) and DI ($r_{pbis} = .37, n = 8$, *ns*), these differences were not statistically significant because of the small sample of effect sizes. Further examination of pre- and posttest average scores indicated that this relationship is the result of the BS/MD students entering the program with better skills. The training program acted as a leveler, as there were no differences between BS/MD and traditional students at posttesting. On the pretest, however, BS/MD students tended to perform significantly better than their traditional cohorts, most likely as a result of a course these students received prior to this program.

Graduating class. Year of graduation and the effect size for each year 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 DI and $-.98$ ($n = 4, p < .01$, one-tailed) for the CI. These results indicate almost a perfect negative relationship between recency of graduation and performance.

Discussion

The effect sizes (d) in this study were translated into the effect size η^2 (really r_{pbis}) used by Haase, Waechter, and Solomon (1982) in determining the average effect size for research reported in the *Journal of Counseling Psychology* over a 10-year period. The effects for performance on the DI and CI placed in the 96.4th (r_{pbis} = .79) and 98.2nd (r_{pbis} = .87) percentiles of this distribution of counseling effect sizes, respectively. 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.

The decreasing trend between effect size and recency of year of graduation has led to an ongoing examination of the training program, which hopefully will lead to "small changes in the way we do things," one of the desired goals of evaluation according to Oetting (1982). Several plausible rival explanations for this decreasing trend are being considered. Because the class of 1981 was the first graduating class at this new medical school and accreditation rested in part on this initial class's performance, perhaps admission procedures have consciously or unconsciously changed over the course of these 4 years. Larger class sizes may be a factor. Also, some changes in instructors and the curriculum have occurred over this period.

Each year the number of physicians in the program who modeled and reinforced the importance of these skills has declined. Training sessions for instructors have decreased from 3 days the first year to 2 hours during the fourth year. This would allow more opportunity for instructors to do "their own thing" rather than adhering strictly to the designed curriculum. The "hidden curriculum" of how to get through the course most efficiently and unscathed has now become oral tradition, perhaps enabling students to invest less in the course and still get by. Sheer time allocation for skill acquisition related most directly to that required on the Carkhuff measures also declined from the first year, as the curriculum was broadened to include initiating responses in addition to empathic responding. Thus the correspondence between training and evaluation may have declined somewhat. Finally, instructors appear now to be less personally involved with students than in the beginning years, with more of an adversarial relationship developing in the students' competition for grades. Instructor enthusiasm for a new and innovative program that in turn impacts upon students' experiences in the course may naturally diminish over time. It is possible that a program may become less impactful once it becomes a standard part of the curriculum. However, it should be kept in mind that the effect of the program was positive, significant, and large for each class even though the magnitude of the effects varied from year to year. As Oetting (1982) noted, "the meaning of a score is as important as whether a difference exists" (p. 67).

Using combined tests and measures of effect size has provided 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-à-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 on smaller sample sizes, as well as providing insight for curricular planning and development.

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. Although the Hawthorne effect might account for some of the evidenced student gains, we believe this is unlikely in as much as neither students nor instructors were aware that they were participants in this evaluation research study; testing was originally designed solely for student evaluation. 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. Obviously performance on the Carkhuff pencil-and-paper measures may not necessarily translate into commensurate performance in on-the-job behavior.

References

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

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

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

Press.
Engler, C. M., Saltzman, G. A., Walker, M. L., & Wolf, F. M. (1981). Medical student acquisition and retention of communication and interviewing skills. *Journal of Medical Education*, *56*, 572–579.
Glass, G. V. (1978). Integrating findings: The meta-analysis of research. In L. S. Shulman (Ed.), *Review of Research in Education* (Vol. 5, pp. 351–379). Itasca, IL: Peacock.
Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology*, *29*, 58–65.
Kraemer, H. C., & Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin*, *91*, 404–412.
McPherson, C., Knopp, W., Sachs, L., & Wolf, F. M. (in press). The doctor–patient relationship: Systematic training in effective communication skills. *Journal of Psychiatric Education.*
McPherson, C., Wolf, F. M., & Sachs, L. (1983). Improving interviewing effectiveness: Positive awareness vs. skills training. *Psychological Reports*, *52*, 741–742.
Oetting, E. R. (1982). Program evaluation, scientific inquiry, and counseling psychology. *The Counseling Psychologist*, *10*, 61–70.
Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin*, *85*, 185–193.
Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.
Wolf, F. M. (1982, August). *Meta-analytic applications in program evaluation.* Paper presented at the annual meeting of the American Psychological Association, Washington, DC. (ERIC Document Reproduction Service No. ED 225 049)

Received July 5, 1983
Revision received October 18, 1982 ■