DOCUMENT RESUME

ED 175 368 HE 011 698

AUTHOR Sheehan, T. Joseph
TITLE Moral Judgment as a Predictor of Clinical Performance.
SPONS AGENCY Commonwealth Fund, New York, N.Y.; Connecticut Univ. Research Foundation, Storrs.; National Fund for Medical Education, Cleveland, Ohio.
PUB DATE 7 Apr 79
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, California, April 7, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Background; Comparative Analysis; Ethics; Foreign Medical Graduates; Geography; *Graduate Medical Students; *Integrity; Internal Medicine; *Medical Services; *Moral Values; *Performance Factors; Physicians; *Predictor Variables
IDENTIFIERS Family Practice (Medicine)

ABSTRACT

Over a 4-year period moral reasoning and performance data were studied on 350 resident pediatricians, internists, and practitioners of family medicine from seven different institutions. Clinical performance was measured by faculty ratings, and integrity (moral reasoning) was measured by Kohlberg's Standard of Moral Judgement Interview and Rest's Defining Issues Test (DIT). Findings of the resident pediatricians showed that there was a significant difference between American and foreign residents, and that there was a high correlation between performance ratings and integrity scores. The residents in internal and family medicine were a much smaller group and none of the correlations were statistically significant. Limitations of the study included the nature of judging values, sampling procedures, and the use of adjusted performance ratings. It was concluded that while the study indicated a relationship exists between integrity and clinical performance, it represented indirect evidence. Extensive tables analyzing the findings are appended. (PHR)

MORAL JUDGMENT AS A PREDICTOR OF CLINICAL PERFORMANCE*

T. Joseph Sheehan, Ph.D.
Professor and Head
Dept. of Research in Health Education
UConn Health Center
Farmington, CT 06032

*The research reported here was based at the University of Connecticut Health Center, Farmington, and is supported in part by grants from the National Fund for Medical Education, the Commonwealth Fund, and the University of Connecticut Research Foundation.

Collaborators on this project are T. Joseph Sheehan, Ph.D., and Susan D. R. Husted, B.S., Department of Research in Health Education, University of Connecticut Health Center, Farmington, CT; Mr. Mark Bargen, Sperry Univac, Windsor, CT; Dan Candee, Ph.D., Center for Moral Education, Harvard University, Cambridge, MA; and Charles D. Cook, M.D., Department of Pediatrics, Downstate Medical Center, Brooklyn, New York.

Presented at the Annual Meeting of the American Educational Research Association, San Francisco, California -- April 7, 1979.

Although pre-eminent physicians such as Harvey have had little difficulty in characterizing the excellent physician, most attempts to quantify these characteristics or to otherwise predict clinical performance have been unsuccessful.

Harvey listed (Bennett, 1973) six desirable physician characteristics: integrity, intellectual ability, capacity for work, common-sense and judgment, grasp of the scientific method, and knowledge of medicine. At face value, these characteristics, along with empathy, fit the common sense definition of the good physician. We wish our physicians to have them and expect that poorer physicians are deficient in one or more of these areas. That's what intuition says. However, where studies have been done there has been little or no correlation between the measures used and estimates of clinical performance.

School grades, for instance, have not been good predictors of performance. Wingard and Williamson (1973) reviewed some 27 studies performed between 1955 and 1972 and found little relationship between school grades and subsequent performance. This was true for other professionals as well as physicians.

Brown (1970) studied the prescribing habits of physicians and found that antibiotics were inappropriately prescribed in a large proportion of their patients, and yet he could find no knowledge deficit when he inquired. Sanazaro (1976) reports other examples of poor and good performance that are not correlated with physician knowledge. Williamson and others (1975) reports a study where a systematic audit showed that a medical staff failed to properly follow up almost 90% of major laboratory abnormalities. The same staff most enthusiastically received an educational conference

directed toward these shortcomings, but their overall level of performance did not improve. Although it is contrary to common sense there seems to be a clear gap between knowledge and performance, and certainly a lack of correlation between medical knowledge and clinical performance.

The failure to predict physician performance is dramatically summarized in the work of Price and Taylor who graphed some 3,000 correlations between a wide variety of predictors and a variety of performance measures. What emerged, Figure 1, was a bell-shaped curve, closely approximating the theoretical distribution of correlations between two random measures, where the true correlation is zero and the standard deviation is the standard error of randomly generated correlations.

Insert Figure 1 about here

The histogram of actual correlations is hardly distinguishable from the theoretical curve, as can be seen in Figure 1. Ths histogram represents the actual distribution of Price and Taylor's 3,000 correlations. As is evident, the mean of their distribution is centered on the theoretical mean of zero, while their standard deviation fits closely the standard error of a random distribution whose true mean is zero. So, just about all of their correlations could have been produced by chance.

Despite the widespread failure of others to predict successfully physician performance, we were convinced that good and poor performance could be distinguished and predicted. We were not surprised to see that medical knowledge and grades were poor predictors, because we all knew of brilliant professionals who performed poorly for any number of reasons. Although we believe that knowledge of medicine is still

important, we viewed it more as a necessary condition for adequate performance, not as a sufficient condition. The characteristic which we felt was most important on Harvey's list was integrity. We reached that conclusion before seeing Harvey's list. Harvey may have meant to describe honesty and high moral standards in his use of the term, but there is more implied. The latin root of the word integer means whole. And the physician with integrity is whole, or in the modern idiom, together. We were convinced that the lack of integrity or wholeness could partially explain the gap commonly documented between physician knowledge and physician performance, and sought our initial research funding in 1974 using a phrase coined by Voytovich, the knowledge-performance gap among physicians.

METHODS

Although we wished to study the relationship between integrity and physician performance, we faced major obstacles. How do we measure integrity? How do we measure clinical performance? No satisfactory measure of either exists. We therefore adopted the principle that it is more valuable to get an imperfect answer to an important question than to get no answer at all. Since we did not have exact measures we would work with approximations.

To measure clinical performance we used faculty ratings. We felt that ratings would be based upon a broad range of activities related to performance and would cover a much wider array of experiences since the ratings would be based upon daily encounters over a whole year or longer. Such ratings would lack the objectivity of a simulated case, for example, but assuming adequate reliability, they would represent clinical performance across multiple clinical problems and over an extended time period.

The rating form based upon earlier work by Cook and Margolis (1974) had reliabilities in the .75 range. We adapted their scale to a semantic differential format

for 18 performance characteristics as well as a rating of overall performance. Each house officer in our study was rated by three to eight faculty members. On one sample of 26 residents rated by four common faculty members the reliability of the mean rating was 0.86 while the average intercorrelation among four common raters was 0.67. Different raters were used at different institutions and even within the same institution it was not always possible for the same raters to rate all residents. In order to rate a resident a faculty rater had to know the resident fairly well and had to have sufficient clinical experience with the resident. Ratings were done independently and averaged.

Lacking a direct measure of integrity, we were attracted to Kohlberg's theory of moral development and the measures which have been used in his and related research. Although Kohlberg has never claimed that his theory and measures could be used in this way, we felt comfortable with the logic and coherence of his theory and the soundness of his measures.

Kohlberg's theory identifies three levels of moral reasoning: preconventional, conventional, and principled. There are two stages of reasoning within each of these levels, for a total of six possible stages. According to Kohlberg, individuals develop their moral reasoning through a series of six sequential stages. At stages one and two, the preconventional level, reasoning about right and wrong is mainly in terms of reward and punishment. Such reasoning is typical of young children up through the early teen years. At stages three and four, reasoning is mainly focused on maintaining harmony in interpersonal relationships, loyalty to peers and preserving the social order. The postconventional level, stages five and six, is characterized by principled thinking, that is, reasoning about right and wrong based upon values and principles which have validity over and beyond the authority of the groups and persons who hold these values.

Although Kohlberg distinguishes sharply between the structure of moral reasoning and the rightness of the actual choices and behaviors in solving moral dilemmas, there is some evidence of a relationship between measures of moral reasoning and behavior (McColgan, 1975; Jacobs, 1977; Froming and Cooper, 1977; Gunzburger, et al, 1977; G. Rest, 1978). To the extent that measures of moral reasoning are related to behavior, we felt that moral reasoning would be related to physician performance. That is, if we were able somehow to observe and measure the whole range of clinical performance, from the excellent to the poor physician, and if we were able to measure the whole range of moral reasoning, or better yet, integrity, we believed that the two would be related. The highly principled physician would govern his actions, in large part, in terms of what was right and just for himself, his patient, and society; these values would be reflected in his performance. Physicians at the conventional and pre-conventional stages would be motivated more by what they found personally rewarding, the expectations of their peers, and institutional norms, rather than by what was best for the patient.

The measures of moral reasoning were Kohlberg's Standard Moral Judgment Interview (Kohlberg, 1969, 1975, 1976) and Rest's Defining Issues Test (D.I.T.), (Rest, 1974a, 1974b, 1975, 1976a, 1976b). The first is a structured interview in which six moral dilemmas are presented and systematically explored. The D.I.T. is a paper and pencil derivation of the structured interview. Instead of responding to general questions about how each dilemma is solved, the D.I.T. presents a series of statements to be rated in terms of their importance in solving each dilemma. Each statement represents some stage of moral development so that a person's stage of development can be estimated by scoring responses to statements across stages and across each of six dilemmas.

The D. I. T. had a .78 correlation with the Kohlberg interview for the 45 subjects who were given both measures in this study. The D. I. T. reliability is reported in the high .70's and low .80's (Davison and Robbins, 1978).

The sample consisted of 348 house officers, 257 from pediatric residency programs, 68 from medicine, and 23 from family medicine. These physicians were from seven different institutions and data were gathered over a four year period. The samples were not chosen randomly but rather from available institutions where the necessary cooperation could be obtained and which were felt to be representative of the range of house officers in U.S. residency programs, at least in pediatrics. Participation was voluntary following standard human experimentation committee guidelines.

RESULTS

From an early sample of performance ratings we drew a random sample of residents from among the higher and lower performers. We administered the Kohlberg interview to these 48 subjects and found a statistically significant correlation of .47 between their performance rating and their moral maturity scores, derived from the interview.

In order to determine whether responses to the D. I. T. stage scores and ratings on the eighteen performance characteristics were correlated, we performed a canonical correlation. A canonical correlation indicates whether two sets of measures are related.

The canonical correlation between the six D. I. T. stage scores and the eighteen performance characteristics for 257 pediatric house officers was .68, which was statistically significant at $P < .0001$. For the 68 internal medicine residents, the canonical correlation was .75 which also was statistically significant at $P = .02$.

There was one significant canonical correlation for the medical residents and one for the pediatric residents which would indicate that the six D. I. T. stage scores and the eighteen performance characteristics are related. It also means that there is a low probability that the relationship could be attributed to chance and because there is only one correlation it means that they are related along a single dimension.

Intuitively, the D. I. T. is supposed to measure along a single dimension, the development of moral reasoning.

What about the eighteen performance characteristics? How many dimensions are needed to represent them?

In order to examine the structure of the performance characteristics we performed a factor analysis of the correlation matrix of performance characteristics. The factor analysis extracts from this matrix the weighted combination of performance characteristics which accounts for the largest source of variability shared in common by all eighteen performance characteristics. This is the first factor. Then removing the variability due to the first factor, it extracts the next largest source of shared variation and repeats this process untill all meaningful variation is accounted for.

The factor analysis of the eighteen performance sub-scales with a sample of 314[1/] pediatric house officers is shown in Table 1. Three factors account for 82% of the common variance, with the first factor accounting for some 72% of that total. Table 1 shows those physician characteristics which contribute to each of the three factors.

1/ The discrepancy between 314 and 257 is due to the rigorous criteria for screening the D. I. T. for validity. We lost 57 cases, or 18%, because of inconsistencies in responses, incomplete responses, or because the selection of nonsense items was too high. According to Rest a 20% loss at screening is about average.

The first factor consists mostly of cognitive characteristics, the second factor mostly attitudinal and the third factor mostly emotional characteristics. Thus, most of the variability in performance rating is accounted for by one major cognitive component, a single dimension.

Insert Table 1 about here

The D.I.T. stages are meant to constitute a single scale on the basis of developmental theory, and they empirically do so (Davison, 1977). Since both the performance ratings and the moral reasoning measures are uni-dimensional, it is not surprising that the association shared jointly by the performance characteristics and the D.I.T. measures can be summarized along a single dimension, by one significant canonical correlation.

Table 2 shows the summary data for the pediatric house officers. The first column shows the number of residents from each of seven residency programs. Four programs are university-related and three are based in community hospitals. The residents in the first three programs are graduates of American medical schools. There are foreign medical graduates in programs four and five, and a mixture of American and foreign graduates in programs six and seven. The second column contains summary data from the D.I.T. P-score represents the amount of principled reasoning, responses from stages 5A, 5B, and 6. P-percent scores expresses these scores as a percent of the maximum P-score, 57. The third column contains the mean overall performance ratings for the residents in each program. These ratings were gathered by the faculty in each institution and range from 1.0, excellent, to 4.0, unsatisfactory. The fourth column contains adjusted performance ratings; this adjustment will be explained

below. The fifth column shows the correlations between P-score and performance, all of which are in the predicted direction.

Insert Table 2 about here

From Table 2 it is clear that there are institutional differences on P-score, with the most striking difference between American and foreign residents, that performance means and standard deviations are fairly similar across institutions and that P-score is consistently correlated with performance ratings.

Table 3 is a display of American and foreign residents on the same variables, with obvious major differences in P-score. The P-score differences are further described in another paper (Husted, 1978). The correlations between P-score and overall performance are statistically significant for both groups, and when both groups are merged, the overall correlation between P-score and performance is -.33, in the predicted direction, moderate in size, and highly significant. The third column shows the mean adjusted performance for both groups. These adjustments are based upon ratings of each of the seven training programs by a group of 28 professors of pediatrics.

The reason for adjusting performance ratings can be seen by returning to Table 2 and examining column two, overall performance ratings of residents. The ratings of residents in each program have similar means and standard deviations. There is little discrimination across institutions. This occurred despite explicit directions to faculty raters in each institution to rate each resident against national rather than an institutional norm group. Well known differences among residency programs were, therefore, concealed. The seven programs were listed with eleven other residency programs and each program was rated separately as being

in the top 10% of programs nationally, the top 11-25% of programs better than average, below average, bottom 10% and don't know. Table 4 shows the means and standard deviation of these ratings for the seven programs.

Insert Table 4 about here

The program ratings shown in Table 4 were used to standardize the ratings of residents in each program.[2/] In order to perform this standardization we had to assume a one-to-one correspondence between the rated quality of the training program and the resident trainees within these programs.

Returning to Table 3 there is both a clear P-score and adjusted performance difference between the American and foreign residents. These differences are analyzed more completely later, but it is clear that foreign and American residents differ on both P-score and performance. Despite these differences, the correlations between P-score and adjusted performance are similar in size and statistical significance for both groups. When the groups are combined, which provides greater variability in both measures, the correlation between P-score and adjusted performance rises to .57. This correlation is statistically significant and practically significant.

2/

The adjusted performance rating for any resident is found using the following expression:

$$x'_{ij} = \left[(x_{ij} - \bar{x}_j)/s_j\right] s'_j + \bar{x}'_j \qquad \text{where:}$$

$\bar{x}'_j$ = mean rating of program "j" by professors of Pediatrics

s'_j = standard deviation of ratings of program "j"

x'_{ij} = adjusted performance rating of resident "i" in program "j"

x_{ij} = mean rating of resident "i" in program "j"

$\bar{x}_j$ = mean rating of all residents in program "j"

s_j = standard deviation of ratings in program "j"

In order to examine more closely the relationship between stages and performance level, we rescored all D.I.T. responses to determine whether a predominant stage of reasoning existed for each resident. According to Rest (1974), a predominant stage exists if responses at that stage are greater than one standard deviation above the mean for the norms at that stage. From our sample of 257 pediatric residents, we could stage 227 of them. The residents split with about half above stage four and half at stage four or below. In detail 4.5% are at stage 2, 6.3% at stage 3, 35% at stage 4, 12% at stage 5A, 12.7% at stage 5B, and 29.4% at stage 6. Table 5 shows the mean and standard deviation for adjusted performance and the number of residents at each stage. A one-way analysis of variance on these means yields a highly significant F-ratio indicating non-chance difference in performance among residents who are grouped according to the predominant stage of their moral judgment. There is a clear trend of improved performance with higher stage scores and a clear-cut split between the non-principled and principled stages, that is, groups 2, 3 and 4 as compared to groups 5A, 5B and 6.

Insert table 5 about here

Examining these data in a different way, we divided performance into three levels and P-score into three levels to produce Table 6. The number of residents in the highest P-score group and the highest performance group was 45. The Chi-square on this table is 53.26 and is highly significant and indicates that performance and P-score are related. The most interesting data are in the lower left and upper right hand sections of the table. There is only one resident in the highest P-score group rated as a low performer, and only six in the low P-score group who are rated as high performers.

Insert Table 6 about here

Tables 7 and 8 contain similar analyses for American residents and foreign residents separately. One table is the mirror of the other with the Americans generally scoring higher on both performance and P-score, and the foreign residents lower on both.

Insert Tables 7 and 8 about here

Is it possible that the association between moral reasoning and clinical performance is due to the discrepancies in both moral reasoning and clinical performance between American and foreign medical graduates?

Because of the large number of foreign medical graduates in the pediatric sample and because this group had a noticeably lower mean P-score, 19.7 as compared to 32.6 for the American graduates, we wished to see whether P-score by performance correlations might somehow be explained by this factor, that is, by the American versus foreign graduate differences. We therefore ran a two-way analysis of variance, with foreign versus American as the first factor, and principled versus pre-principled as the second factor, i.e., comparing those whose primary stage of reasoning was 5A, 5B and 6 to those whose primary stage was 2, 3 or 4.

The results are summarized in Table 9 using overall performance as the response variable, and in Table 10 using adjusted performance as the response variable.

Insert Tables 9 and 10 about here

Both tables 9 and 10 show statistically significant differences between principled and pre-principled physicians on performance, and statistically significant differences between foreign and American physicians. The interaction mean square, which would show whether the two factors are correlated, is extremely small in both tables. This analysis shows a strong difference between foreign and American medical school graduates on both overall performance and adjusted performance, an equally strong difference between pre-principled and principled moral reasoners on performance, with the principled group out-performing the pre-principled group, and no interaction, which is to say that these conclusions are not conditional. So, the possibility that foreign versus American differences could explain the observed correlation between moral reasoning and performance is eliminated.

Another way of stating this interpretation is that there are clear-cut differences in performance between the principled and pre-principled groups which are independent of the fact that some residents are American and some are foreign.

There were two other ways in which we examined the relation between stage scores and performance. The first was to correlate the six stage scores to overall and to adjusted performance. How much does each stage score correlate with performance? These correlations are shown in Table 11. The correlations follow the same pattern for overall performance as they do for adjusted performance, but are higher for adjusted performance. The shift from negative to positive direction occurs between stages four and 5A. These correlations also support the earlier results showing that moral reasoning and performance are related along a single dimension.

Insert Table 11 about here

The fact that stages represent points along a continuum, at least in theory, raises the question of redundance in the measure, i.e., how much does each stage score contribute to explaining the variance in performance? Which stage contributes most heavily? After removing the variance due to the heaviest contributing stage, is there very much variance left to explain by the other stage scores? These questions led to the second approach which was step-wise multiple regression analysis.

We performed the regression analyses using six different orderings and stage combinations, but the same results emerged from all analyses. First, when all six stage scores are used, the multiple correlation reaches .31 with overall performance and .54 with adjusted performance, which is about the same as when using P-score alone. So, P-score very satisfactorily summarizes the information contained in the D.I.T. at least in predicting clinical performance. Secondly, stage 5A emerges as the key contributor, accounting for 7.3% of the variability in overall performance, and 20% of the variability in adjusted performance. The other stages together account for another 2.1% of the variance in overall performance and 9.1% of the variance in adjusted performance. All regression equations are highly significant.

Results for Medical Residents

The medical residents and family medicine residents were selected from two university-affiliated residency programs. Tables 12 and 13 show summary data from a two-year period. For institution 1, the relationship between P-score and overall performance is .37, which has a significance level of .058. For institution 2, the P-score performance correlation is .06, which is not significant. Upon closer inspection of the scatter-plots, institution 2 contained two very deviant observations, one with a very high P-score and a very low performance rating, a P-score of 6 and a performance rating of 1.4, and the other with a P-score of 40 and a performance

rating of 3.1, both of which were highly atypical observations. When these points were omitted from the computations, the correlation between P-score and performance was .32, significant at the .025 level. None of the correlations between P-score and performance are statistically significant for the Family Medicine residents, but two of three are in the predicted direction, and in the third group there are only four observations.

Insert Tables 12 and 13 about here

In considering the correlations presented in Table 13 there are two points to keep in mind. First, the canonical correlation between the six stage scores and the 18 performance sub-scales was .75 and was statistically significant at the .02 level. Second, although null hypotheses were not rejected for these correlations it is important to avoid the trap of concluding that there is no relationship between P-score and performance in these samples, (Freiman, et al, 1978). It is appropriate, therefore, to look at the 95% confidence limits on the observed correlations. The limits for the correlation of .37 are from -.07 to .67; the limits for the correlation of .06 are from -.22 to .035; the limits for the correlation of .32 are from .04 to .55; and the limits for the correlation of .16 are from -.09 to .38. While zero correlation is within each of these limits, it is generally at the tall end of these limits. The possibility exists, especially with these rather small samples of medical residents who are much more homogeneous than the pediatric residents, that we could be reporting a type II error, or falsely accepting a null hypothesis.

DISCUSSION

Before discussing some of the limitations of this study, it is important to understand

its significance, not in a statistical sense, but in the sense that finding reliable predictors of clinic ' performance has been so frustrating and success has been so rare. Essayists, on the other hand, have had little trouble describing the good doctor, and even the research of Price and Taylor (1971) concluded with a set of eight characteristics of the excellent physician.

The above results firmly support our rationale that moral reasoning is a predictor of clinical performance. The association between moral reasoning and clinical performance shows up consistently across many approaches to the data: simple correlation, multiple regression, analysis of variance, and chi-square. The correlation cannot be attributed to differences between American and foreign medical graduates. The correlation is stable across both groups, analyzed separately or together. The correlations are stronger in the pediatric samples than in the internal medicine or family medicine samples.

The present study was prospective. We began with a set of expectations based upon theory and experience. We knew that medical knowledge was a poor predictor of clinical performance, that medical school grades and MCAT (Medical College Admission's Test), and a host of biographical and personality variables were poor predictors. We were convinced, however, that integrity, in the sense of wholeness, not self-righteousness, was related to performance. It is in Harvey's list. It emerges from the Price and Taylor studies.

The title of our initial proposal was, "The Knowledge-Performance Gap: A Possible Explanation." We believed there was more to performance than knowledge and problem solving skills. Clinical judgment, in the sense of integrating all of the available patient data and evaluating the patient's subjective status, including the patient's attitudes and values, had to be involved. The values of the physician and his priorities,

his responsibilities to his peers, his institution, himself, and his sense of what is right -- all would play a role. The difficulty, for us, was to find some way to quantify these influences.

Kohlberg distinguishes strongly between moral reasoning and moral behavior. We earlier identified some studies suggesting a relationship between moral reasoning and behavior. Our use of the Kohlberg and Rest measures was predicated on a belief that they would be related to behavior, at least weakly related. In the performance area we chose to use ratings of overall performance. We were more interested in habitual performance observed over an extended period of time than in response to a single measure, such as a simulated case, at a single point in time. We preferred to be as unobtrusive in measuring clinical performance as possible. And faculty are certainly part of the working environment for residents.

Recent work at our school (Harper, 1979) shows that it is now possible to obtain accurate assessment of a student's ability to formulate clinical problems directly from the medical record. Assessment of performance from the medical record, when it is more fully developed, will constitute a much more precise, objective, and perhaps a more valid assessment of performance than faculty ratings. But, we also have evidence that there is some correlation between record audit and faculty ratings (Voytovich, 1975).

Given the measures we used, it is remarkable that we were able to observe any relationship at all. It would seem, in fact, that any relationship estimated from our data might even be considered an underestimate of the true relationship between the underlying constricts that interest us, that is, between integrity and physician performance. After all, we know of no physician in our study who were either charged with or guilty of medical malpractice, or who were judged to be overtly dishonest and

unethical. Clearly we did not include extreme values. Nonetheless we found relationships that are statistically significant, psychologically meaningful, and perhaps practically useful.

Moral behavior is a combination of moral reasoning plus such other non-moral factors as one's perception of the probability of success in a given problem, one's emotional makeup, e.g., brave versus cowardly, ego strength, willingness to act on a decision, desire for publicity or fame. Since moral judgment is only one factor contributing to moral behavior, we should not expect a one-to-one correspondence between the decisions a physician makes and his moral behavior. However, the fact that we continually found correlations between moral reasoning and a general measure of physician performance indicates that moral reasoning itself is an important component of clinical behavior.

Damon (1977) did find that when he examined children's reasoning about distributive justice, it was closely related to their moral reasoning about a real dilemma involving distributive justice. However, he also found that neither measure was related very well to actual patterns of behavior, Damon's results, although with children, do illustrate that other factors are involved in the transition from thought to action, from deliberation to decision.

One of our future projects will be to look more closely at the influences on the clinical behavior of physicians; the intent is to estimate more precisely the impact of these other influences. Our future work focuses on trying to observe more specifically the ways in which moral reasoning influences performance.

Other limitations of this study, in addition to the fallibility of available measures, might include sampling procedures. Although the study was prospective, we were unable to identify a population of physicians from which we could randomly

sample. We chose residency programs that would be cooperative, that would be accessible to our research team, and that we could afford within the limits of our funding. There is the possibility, moreover that random selection may be over emphasized. Many worthwhile studies may not get done because random selection of subjects is impossible or because random assignment of subjects to treatment once selected, is not feasible. What is even more important than randomization is replication. Are the results reproducible in different samples and over time? We did replicate. We studied pediatric house officers over a four year period, and medical house officers over a two year period, replicating within and across medical specialities.

To the extent that our samples were not randomly selected, however, we must be careful about the extent to which we can generalize the results beyond the kinds of residents we studied. On the other hand, the results were fairly stable over time and across samples, so that within these kinds of residency programs we can be confident of the stability of our findings.

Finally, we would have preferred not to have introduced the notion of adjusted performance ratings. If there were some practical and economical way of gathering performance data across institutions that would be reflective of differences in the way residents perform, we would have done so. The raw data did not discriminate, although, we were convinced that there were real and observable differences among institutions. The ratings of programs by professors of pediatrics did seem to reflect well known program differences. When these adjustments are used, the correlation between P-score and performance is rather spectacular, .57. When those adjustments are not used, the correlation is still rather respectable, .33.

CONCLUSIONS

Over a four year period we have gathered moral reasoning and performance data on a total of 350 house officers from pediatrics, internal medicine and family medicine. We have repeatedly confirmed our hypothesis that moral reasoning is a predictor of clinical performance. Although we believe that integrity is causally related to clinical performance, and although this study may be regarded as confirmatory, it represents indirect evidence. Moral reasoning may be related to integrity, but it is conceptually distinct. To confirm our hypothesis more directly will necessitate a more direct methodology.

TABLE 1

FACTOR ANALYSIS OF PERFORMANCE RATINGS

Factor 1		Factor 2		Factor 3	
Organized	.72	Admits Mistakes	.79	Empathy	.89
Knowledge	.85	Responsible	.62	Compassionate	.52
Teaching Skills	.71	Honest	.67	Seeks Consultation	.83
Seeks Knowledge	.72	Dependable	.63		
Decision Making	.83	Works Hard	.71		
Clinical Judgment	.79	Relates Well to Patients	.65		
Acts in Emergency	.68	Compassionate	.62		
		Works Well With Others	.77		
		Knows Own Limits	.73		
% of Variance	71.9		5.8		5.0

TABLE 2

MEANS, STANDARD DEVIATIONS, AND CORRELATIONS FOR PEDIATRIC SAMPLES

Institution	Principled Reasoning P-Score		P%	Overall Performance	Adjusted Performance	Correlation of P-Score and Overall
American Graduates						
(1) N=105/80*	$\bar{X}$	33.3	58%	1.9	1.8	r=.17**
	S.D.	7.9		0.6	0.8	P=.04
(2) N=6/4	$\bar{X}$	28.0	49%	1.7	1.8	r=.60
	S.D.	10.4		0.3	0.1	P=.10
(3) N=38/36	$\bar{X}$	31.2	55%	2.0	3.2	r=.28
	S.D.	7.3		0.6	0.6	P=.05
Foreign Graduate						
(4) N=17/17	$\bar{X}$	17.2	30%	2.0	4.4	r=.32
	S.D.			0.5	0.4	P=.11
(5) N=91/80	$\bar{X}$	20.1	36%	2.3	4.3	r=.23
	S.D.	8.2		0.4	0.7	P=.02
Mixed						
(6) N=4/4	$\bar{X}$	33.3	58%	2.4	4.5	r=.41
	S.D.	9.5		0.3	0.4	P=.30
(7) N=9/9	$\bar{X}$	26.22	46%	1.9	4.3	r=.04
	S.D.	11.2		0.6	1.1	

* N=;05 is the number of residents taking the DIT; 80 is the number with performance ratings and DIT.

** The algebraic signs have been changed for ease of interpretation.

TABLE 3

AMERICAN VERSUS FOREIGN MEDICAL GRADUATES

	Principled Reasoning P-Score		Principled Reasoning P%	Overall Performance	Adjusted Performance	Correlation of P-Score & Overall	Correlation of P-Score Adjusted
American 1, 2, 3 N=147	$\bar{X}$	32.6	57%	1.9	2.13	.20** P=.007	r=.21 P=.005
	S.D.	7.8		0.6			
Foreign 4, 5* N=97	$\bar{X}$	19.7	34%	2.3	4.0	.20 P=.028	r=.25 P=.006
	S.D.	7.9		0.4	0.9		
ALL N=244	$\bar{X}$	27.45	48%	2.1	2.9	.33 P=.001	r=.57 P=.001
	S.D.	10.1		0.6	1.3		

*Residents in 6 and 7 are omitted because they are mixed samples.
**The algebraic sign has been changed for ease of interpretation.

TABLE 4

RATINGS OF RESIDENCY PROGRAMS BY PROFESSORS OF PEDIATRICS

Residency Program	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean	1.86	1.77	3.13	4.35	4.12	4.33	4.47
Standard Deviation	0.79	0.69	0.61	0.99	0.99	1.18	0.92

Note: Raters have been removed from ratings of their own institution.

TABLE 5

ANALYSIS OF VARIANCE ON ADJUSTED PERFORMANCE BY STAGE SCORES

	Predominant Stage Score					
	2	3	4	5A	5B	6
Mean Adjusted Performance	4.03	3.86	3.46	2.25	2.68	2.34
Standard Deviation	1.1	.98	1.33	1.14	1.23	1.10
# of Residents Per Stage Score	N=10	N=14	N=77	N=27	N=28	N=65

$F = 11.5$, $DF = 5$ and 215, $P < .0001$

TABLE 6

THREE-WAY COUNT OF P-SCORE BY ADJUSTED PERFORMANCE

P-Score / Performance	High Principled* P > 35	Medium Principled 20 to 34.9	Low Principled P < 19.9	Row Sum	Row Percent
High Performance < 2.5	45	44	6	95	38.9%
Medium Performance 2.5 to 4.5	24	60	36	120	49.2%
Low Performance > 4.5	1	10	18	29	11.9%
Column Sum	70	114	60	244	
Column Percent	28.7%	46.7%	24.6%	100%	

$\chi^2 = 53.26$, DF = 4, P < .00001

* Principled P-Score

TABLE 7

THREE-WAY COUNT OF P-SCORE BY ADJUSTED PERFORMANCE
FOR AMERICAN GRADUATES

P-Score / Performance	High Principled* P > 35	Medium Principled 20 to 34.9	Low Principled P < 19.9	Row Sum	Row Percent
High Performance < 2.5	45	43	5	95	63%
Medium Performance 2.5 to 4.5	19	32	2	53	36%
Low Performance > 4.5	0	0	1	1	.7%
Column Sum	64	75	8	147	
Column Percent	43.5%	51%	5.4%		100%

$\chi^2 = 20.2$, DF = 4, P = .0005

* Principled P-Score

TABLE 8

THREE-WAY COUNT OF P-SCORE BY ADJUSTED PERFORMANCE
FOR FOREIGN GRADUATES

P-Score / Performance	High Principled* P > 35	Medium Principled 20 to 34.9	Low Principled P < 19.9	Row Sum	Row Percent
High Performance < 2.5	0	1	1	2	2%
Medium Performance 2.5 to 4.5	5	28	34	67	69%
Low Performance > 4.5	1	10	17	28	29%
Column Sum	6	39	52	97	
Column Percent	6%	40%	53%		100%

* Principled P-Score

TABLE 9

ANALYSIS OF VARIANCE ON OVERALL PERFORMANCE

	Mean Square	F-ratio	P-value
Foreign vs. American	2.51	8.98	.003
Pre-Principled vs. Principled	2.35	8.43	.004
Interaction	.09	.32	.999

TABLE 10

ANALYSIS OF VARIANCE ON ADJUSTED PERFORMANCE

	Mean Square	F-ratio	P-value
Foreign vs. American	115.47	132.46	.001
Pre-Principled vs. Principled	6.59	7.56	.006
Interaction	.124	.142	.999

TABLE 11

CORRELATIONS OF STAGE SCORE WITH PERFORMANCE

	Overall Performance	Adjusted Performance	
Stage 2	-.12*	-.24***	
Stage 3	-.18**	-.39	
Stage 4	-.24	-.29	
Stage 5A	.29	.47	
Stage 5B	.19	.28	
Stage 6	.22	.39	
P-Score	.34	.56	N=220

* P = .03
** P = .003, all other correlations are significant at P = .001.
*** The algebraic signs have been reversed for ease of interpretation.

TABLE 12

MEANS, STANDARD DEVIATIONS, AND CORRELATIONS FOR RESIDENTS FROM INTERNAL MEDICINE

	P-Score	P%	Overall Performance	P-Score Overall	95% Confidence Limits
(1) N = 19	33.8 (6.5)	59%	1.8 (0.5)	.37* P=.06	-0.07 to 0.67
(2) N = 49	30.4 (7.6)	53%	2.0 (0.5)	.06 .32** P=.025	-0.22 to 0.35 0.04 to 0.55
Both (1) & (2)	31.3 7.4		1.9 0.5	.16 P=.09 .32 P=.04	-0.04 to 0.38 0.07 to 0.45

* The algebraic signs have been reversed for ease of interpretation.
** Computed with two outliers omitted.

TABLE 13

MEANS, STANDARD DEVIATIONS, AND CORRELATIONS FOR RESIDENTS FROM FAMILY MEDICINE

Institution	P-Score	P%	Overall Performance	P-Score Overall
(1) A, N = 7	33.0 8.8	58%	1.9 0.5	.11*
(1) B, N = 12	35.2 8.9	62%	1.8 0.3	.17
(2) N = 4	28.0	49%	1.7 0.6	-.37

* The algebraic signs have been reversed for ease of interpretation.

REFERENCES

Aronfreed, J. Moral development from the standpoint of a general psychological theory. In T. Luckona (Ed.), Moral development and behavior. New York: Holt, Rinehart, & Winston, 1976, 54-69.

Bandura, A. Social learning theory. Englewood Cliffs, NJ: Prentice-Hall, 1977.

Bennett, Ivan L., Jr., M.D. Trends and objectives in medical education. Quoting the late Dr. Samuel Harvey, Professor of Surgery, Yale University. Bulletin of the NY Academy of Medicine, Vol. 49, No. 4, April 1973.

Brown, C. R., Jr., Uhl, HSM: Mandatory continuing education: Sense of nonsense? JAMA 213: 1660-1668, 1970.

Damon, W. The social world of the child. San Francisco: Jossey-Bass, 1977.

Davison, M. L. On a unidimensional, metric unfolding model for attitudinal and developmental data. Psychometrika, 1977, 42, 523-548.

Davison, M. L. and Robbins, S. The reliability and validity of objective indices of moral development. Unpublished manuscript, University of Minnesota, 1977.

Freiman, Jennie, et al. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. The New England Journal of Medicine, Sept. 28, 1978, pp 690-694.

Froming, Wm. J. and Cooper, Robert G. Jr., Predicting compliance behavior from moral judgment scales. Journal of Research in Personality, 11, 368-379, 1977.

Gunzburger, D. W., Wegner, D. M., Anooshian, L. Moral judgment and distributive justice. Human Development, 1977, 20, 160-170.

Harper, Laura. Transforming the Subjective Evaluation: A Reliable Audit of the Problem List. M.D. Thesis, University of Connecticut,School of Medicine, 1979.

Husted, S. Using a stage profile to assess judgment of moral issues: examining the development of moral reasoning in pediatric faculty, house officers, and medical students. (In preparation.)

Jacobs, M. K. Women's moral reasoning and behavior in a contractual form of prisoners' dilemma. In J. Rest (Ed.), Development in judging moral issues -- a summary of research using the Defining Issues Test. Minnesota Moral Research Projects, Technical Report #3, 1977. (ERIC Document Reproduction Service No. ED144980).

Kohlberg, L. Stage and sequence: The cognitive-developmental approach to socialization. In D. Goslin (Ed.), Handbook of socialization theory and research. Chicago: Rand McNally, 1969, 347-480.

Kohlberg, L. and colleagues. Moral stage scoring manual: Part I, introduction to interviewing and scoring, and Part II, form A-I standard scoring manual, and Form B-I, standard scoring manual, Center for Moral Education, Harvard Graduate School of Education, 1975 (unpublished).

Kohlberg, L. Moral stages and moralization: The cognitive-developmental approach. In Moral Development and Behavior (Lickona, T., Ed.), New York: Holt, Rinehart, and Winston, 1976.

Krebs, R. L. Some relations between moral judgment, attention, and resistance to temptation. Unpublished doctoral dissertation, University of Chicago, 1967.

Margolis, C. Z. and Cook, C. D. Rating pediatric house officer performance. Pediatric Research, 8:472, 1974.

Mischel, W. Processes in delay of gratification. In L. Berkowitz (Ed.), Advances in Social Psychology, Vol. 7, New York: Academic, 1974.

McColgan, E. Social cognition in delinquents, predelinquents and nondelinquents. Unpublished doctoral dissertation, University of Minnesota, 1975.

Price, P. B., Taylor, C. W., et al. Measurement and predictors of physician performance. Salt Lake City: L. L. R. Press, 1971.

Rest, J. Manual for the defining issues test. University of Minnesota, 1974.

_____. Manual for the defining issues test: An objective test of moral judgment development. Available from the author, 1974a.

_____. The cognitive-developmental approach to morality: The state of the art. Counseling and Values, 18: 64-78, 1974b.

_____. Longitudinal study of the defining issues test of moral judgment: A strategy for analyzing developmental change. University of Minnesota, 1975 (unpublished).

_____. New approaches in the assessment of moral judgment. In Moral Development and Behavior, Lickona, T. (Ed.), New York: Holt, Rinehart, and Winston, 1976a.

_____. Moral judgment related to sample characteristics. Final report to NIMH, available from the author, 1976b.

Rest, G. J. Voting preference in the 1976 presidential election and the influences of moral reasoning. Unpublished manuscript, University of Michigan, 1978.

Sanazaro, Paul J., M.D. Medical audit, continuing medical education and quality assurance. The Western Journal of Medicine, 125 3, September 1976, 241-252.

Voytovich, A. E. An analysis of student performance as reflected in the non-threatening audit of the structured medical record. Academic Decision Making -- Issues and Evidence, published by AAMC (1976) pp 37-42.

Williamson, J. W., Aronovich, S., Simonson, L., et al. Health accounting: An outcome-based system of quality assurance: Illustrative application to hypertension. Bulletin, NY Acad. Med. 51: 727-738, 1975.

Wingard, J. R., and Williamson, J. W., M.D. Grades as predictors of physicians' career performance: An evaluative literature review. JAMA, Vol 48, April 1973, 311-322.