DOCUMENT RESUME

ED 186 465 TM 800 168

AUTHOR Blumberg, Phyllis
TITLE Issues Concerning the Evaluation of Medical Students' Abilities to Formulate Problem Lists.
PUB DATE Apr 80
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Clinical Diagnosis; Comparative Testing; Cues; Higher Education; Medical Students; Multiple Choice Tests; *Problem Solving; *Recall (Psychology); *Recognition (Psychology); *Test Format

ABSTRACT

To help determine the role that test instrument formats play in evaluation, two parallel examinations were given to 227 second-year medical students. The tests were based on information presented in a medical case history. One required students to generate their own problem lists (the generate group); the other required the students to select problem lists from a list of alternatives (the select group). All the students had difficulty formulating problem lists as indicated by average overall scores of 42% and 57% correct for the generate and the select groups, respectively. Significant quantitative and qualitative differences were noted between the two groups in that the select group usually picked properly integrated problems while the generate group constructed partially correct answers composed of unintegrated cues. As predicted, the select group scored significantly higher than the group generating their own lists. (The relative utility of generate or select response formats for diagnostic and certifying examinations is discussed). (Author/GDC)

ISSUES CONCERNING THE EVALUATION OF MEDICAL STUDENTS' ABILITIES TO FORMULATE PROBLEM LISTS

Phyllis Blumberg, Ph.D.
Center for Educational Development
University of Illinois at the Medical Center

Paper presented at the annual meeting of the American Educational Research Association in Boston, Massachusetts, April, 1980.

The author wishes to acknowledge the assistance of the physicians who reviewed and scored the answers to the exam: Frederick Berlinger, Kenrad Nelson, and Harold Shafter.

ISSUES CONCERNING THE EVALUATION OF MEDICAL STUDENTS' ABILITIES TO FORMULATE PROBLEM LISTS

PHYLLIS BLUMBERG, Ph.D., Center for Educational Development, University of Illinois at the Medical Center

To help determine the role that the examining instrument formats play in evaluation, two parallel exams were given to 227 second-year medical students. One required students to generate their own problem lists (the generate group); the other required the students to select problem lists from a list of alternatives (the select group). All of these second-year medical students had difficulty formulating problem lists as indicated by average overall scores of 42% and 57% correct for the generate and the select groups respectively. Significant quantitative and qualitative differences were noted between the two groups in that they usually picked properly integrated problems while the generate group constructed partially correct answers composed of unintegrated cues. As predicted, the select group scored significantly higher than the group generating their own lists. The relative utility of generate or select response formats for diagnostic and certifying examinations is discussed.

ISSUES CONCERNING THE EVALUATION OF MEDICAL STUDENTS' ABILITIES TO FORMULATE PROBLEM LISTS

Literature on clinical judgments (e.g. Elstein et. al., 1978; Feinstein, 1967; Weed, 1971) indicates that one of the most important aspects of patient care is making clinical judgments about the patient's problems. One essential aspect of clinical judgment is generating diagnostic hypotheses about problems. Therefore, an evaluation of students' clinical judgments should include some assessment of their ability to develop diagnostic hypotheses. The evaluation of clinical judgments is the focus of an exam discussed in this paper. The object of this exam was to evaluate students' abilities to formulate diagnostic hypotheses and patient problems. (For the purpose of this exam both diagnostic hypotheses and patient problems were required on the problem list as explained in the directions and sample case given to the students.) Previous efforts to measure students' abilities to formulate problem lists (Berner, 1976, 1977; Helfer and Slater, 1971) indicate that second-year medical students have difficulty with this task.

Purposes of the Study

Student difficulty in formulating problem lists may be a function of any of several factors including (a) the student's knowledge of the content, (b) the student's clinical judgment ability, and (c) the nature of the particular case. The evaluation of these students can also be influenced by the examination format used. This study addresses the influence of examination formats and, therefore, attempts to minimize the effects of the other three factors. One way of determining the effect of the examining instrument on student performance is through the use of alternative examination formats. This study also considers the utility of two examination formats as they relate to the intended purpose of the examination. More specifically, this study addresses student performance as a function of whether students generate their own problem lists or select problems from a long list of problem-options.

Students who select from a list of possible problems must discriminate the correct problems from the incorrect problems. The processes of recognition and elimination can be used to assist these students. Students who must generate their own lists cannot use the processes of elimination or recognition, but must primarily employ recall processes. It is well documented that item-selection tests are easier than content-parallel tests that require student-generated responses (Anderson, 1972; Kintsch, 1970; Loftus & Loftus, 1976; and McCarthy, 1966). However, the decision to use student generated exams may not be that simple. While the major consideration in selecting an exam format should be the purpose of the exam, feasibility considerations often mediate against the best choice. The evaluation literature has not fully explored the relationships between examination purpose (e.g. admissions decisions, diagnostic counseling, or certification of competence) and the exam format selected. Nor does it test for qualitative differences between

student answers as a function of examination formats. This study analyzes the qualitative and quantitative differences between student-selected and student-generated answers in order to (1) determine if these qualitative differences do exist, and (2) address the possible relationships between examination format and intended purpose.

Method

Data Source

A second-year medical class (N = 227) at a state-supported university was divided in half on the basis of the first letter of students' last names (e.g. A-K, and L-Z). Means difference testing on other sections of the exam was conducted to evaluate the sampling distribution. These results indicated that the two groups did not differ significantly on test scores from the other sections of this exam with patient cases which required the student to determine patients' diagnoses and problems ($\overline{X}_{A-K}$ = 61.893, s.d. = 9.759; $\overline{X}_{L-Z}$ = 62.690, s.d. = 9.812; $t = 0.615$, $p < .54$). Thus, these two groups were initially comparable for the purposes of this study. The first group (i.e., A-K) was designated the generate group; the other group (i.e., L-Z) was designated the select group.

Procedure

The experimental examination designed to evaluate the ability of medical students to generate or select problems was integrated within a comprehensive examination which evaluated ability to make clinical judgments. Other sections of this examination used both generate and select formats to evaluate the students' clinical judgments and knowledge in other areas of medicine. The experimental section was limited to the initial problem list of one patient. Students were given a clinical data base containing history and physical data

for this patient and were expected to determine his problems or diagnostic hypotheses. Briefly, the patient was admitted to the hospital for repair of bilateral inguinal hernia, with a history of congestive heart failure and diabetes mellitus. Half the students generated a list of problems/diagnoses (the generate group) and half of the students (the select group) selected the problems/diagnoses from a list of 67 possibilities. This list was composed of responses from previous second-year students. Both groups were told to indicate no more than 20 problems and/or diagnoses which they felt accurately conveyed the patient's present situation.

Prior to the administration of the exam, physicians developed an "ideal problem list" containing 13 common problems and diagnostic hypotheses based on data presented for this case. Next, they determined the credit (i.e., 1, 2, 3 or 5 points) to be given to each problem or diagnostic hypothesis. One point was given for each relevant family history problem reported on the problem list. Two points were given for each of the patient's diagnostic problems reported in the data base (e.g. inguinal pain). Three points were given for each problem that required a little interpretation of data (e.g. Diabetes mellitus--controlled by diet alone. His history of diabetes was given in the data base as well as the drugs he took. The fact that it was controlled, had to be interpreted from all of the data given.). Five points were given for each problem which required interpretation and synthesis of the data (e.g. organic heart disease as evidenced by several signs and symptoms given separately in the history and physical). The ideal problem list was worth 36 points. The physicians decided that partial credit could be given for responses which identified the relevant signs or symptoms but were not completely integrated into problems or diagnostic hypotheses (e.g. diet control, without giving the reason). No credit was given for signs and symptoms which were merely repeated

as stated in the data base and which should have been integrated further.

Each of the options given to the select group of students had also been categorized by these physicians according to problem and appropriate classification credits. The classification credits were: completely correct (1, 2, 3 or 5 points, as discussed above); partially correct--partially unintegrated problems (1, 2, 3 or 4 points); unintegrated (0 points); and inappropriate problem or diagnosis (0 points). Inappropriate or over-resolutions were diagnoses that could not be substantiated from the data base without the results of laboratory tests not yet available, or complications which could result from the present condition (e.g. post operative pain from the hernia operation). For partially correct problems, only 60% of the total point credit was allowed to be accumulated regardless of how many partial answeres were given.

Once the examination was given, the physicians categorized the student-generated responses in the same fashion according to problem and appropriate classification credits. Identical credit was given to the same response. Any generated responses not on the list of options were assigned appropriate credit according to the same classification scheme. Scores were summed for each student and an item analysis for each problem was conducted for both groups.

Results

Overall, the select group scored significantly higher ($t_{227} = 9.589$, $p < .05$) than the generate group on the list of problems/diagnoses. The mean scores and standard deviations for the generate group was $\bar{X} = 42\%$, s.d. = 11.364; the select group was $\bar{X} = 57\%$ and s.d. = 9.018. The select group did

not identify all of the patient's problems even though they did indicate significantly more ($t_{227} = 2.92$, $p < .01$) individual problems than the generate group. Although this difference is statistically significant, it may not be meaningful since the select group averaged only one more problem than the generate group (generate: $\bar{X} = 8.466$, s.d. = 1.447; select: $\bar{X} = 9.018$, s.d. = 1.643). It is interesting to note that a majority of the students in both groups failed to mention the primary reason for the patient's hospital admission (herniorrhophy) (93% - select; 94% - generate), his inguinal pain (70% - select; 68% - generate), and his nausea (88% - select; 100% - generate), although all of this information was clearly mentioned in the data base. The students selected or wrote an average of 12.823 and 11.826 separate responses.

There clearly is an important difference in the quality of answers between the select group and the generate group, as Table 1 indicates. The select group usually picked the completely correct answers for problems they believed to be relevant, while the generate group wrote partially integrated cues and not completely correct answers for the four problems that required integration of cues. The differences in the quality of answers is especially pronounced with the problem requiring the greatest integration of cues into a diagnosis (i.e., arterioscleratic cardiovascular disease) or the problem requiring discrimination among levels of resolution (e.g. controlled diabetes mellitus). Only 7% of the generate group received full credit for the arterioscleratic cardiovascular disease problem, whereas 76% of the other group selected the completely correct answer ($z = 20.847$, $p < .001$); the generate group wrote unintegrated cues for which they received no credit or partially integrated cues for which they received partial credit; 97% of the select group indicated that his diabetes mellitus was controlled by diet alone (i.e., full credit), whereas only 43% of the generate group said this ($z = 15.503$, $p < .01$); 52% of

the generate group indicated that he had diabetes mellitus but failed to say whether it was controlled or how it was controlled. Thus, giving students a list of problem-options facilitates their integration of cues into diagnostic problems.

However, students in the select group also selected diagnoses and problems that were not appropriate at that time, either unjustified resolution or inappropriate problems. For example, "anxiety" was an inappropriate problem because the data base did not discuss the patient's mood or emotional state; 62% of the select group indicated inappropriate problems for this case, whereas only 22% of the generate group did this ($z = 10.433$, $p < .05$).

Discussion

The majority of students in this study had difficulty formulating a problem list. The most obvious reason for low scores was the omission of about 40% of the correct problems from their problem list. There are several possible explanations for these omissions. Perhaps the students were careless and imprecise in describing problems. It is also possible that the students may not have had a good idea of what a problem and a diagnostic hypothesis were, even though their curriculum emphasizes Weed's (1971) Problem-Oriented Medical Record. The students selected or wrote responses which did not reflect separate problems. Most students restated the same problem by indicating various signs and symptoms without integrating them into a problem or diagnostic hypotheses. These beginning clinical students also sometimes failed to make the appropriate discriminations among levels of integration and resolution in making clinical judgments.

The results of this study suggest that giving a list of problems facilitates students' achievements in that their overall scores are higher and that

for the most part, the students indicate completely correct problem statements rather than partially correct, unintegrated signs and symptoms. The improved performance may be attributable to cueing as an aid to the recall of relevant information. Since it is impossible to determine the extent of the influence of cueing, a scoring correction for the select group was not feasible. However, findings are consistent with earlier work of McCarthy (1966) in the evaluation of clinical performance, which also is indirectly concerned with the effects of cueing. McCarthy (1966) compared student performance on an oral examination and a printed examination using lists of alternatives on several aspects of clinical competence. In general, the scores were higher on the printed examination than on the oral exam for the same students. A second qualitative difference also occurred. The students selected more inappropriate diagnoses when given a list of problem-options than without such a list. Beginning clinical students select some problems at a higher level of resolution than is justified, such as diagnoses that cannot yet be made after reading only the data base and without the supporting laboratory data.

Even when students are given the additional advantage of cues from a list of possible problems, second-year medical students still have difficulty in formulating problem lists. These results indicate they have particular difficulty in (1) integrating cues into problems, (2) selecting the most appropriate levels of problem resolution, and (3) indicating all of the problems for the patient.

Given this difficulty and students' limited clinical experiences, perhaps examinations can be given which facilitate their clinical judgments. Allowing students to select problems rather than requiring them to generate problems facilitates performance because of the processes of cueing and elimination.

Thus, recognition tests may be appropriate for diagnostic examinations for beginning clinical students, since this type of item can identify students' weaknesses, as this study has shown, and is easier to score.

Presenting students with a list of alternatives may not be appropriate for all clinical examinations. Clinical competence is composed of various steps. Data gathering (i.e., taking place during history and physical) is an activity which is cued from what the patient presents and is not directly tied to a predetermined list of possibilities. This raises validity questions on evaluation instruments of data gathering which allow students to select their responses. Diagnostic workups which involve ordering laboratory tests require the integration of many cues. Physicians and other health professionals use the results of earlier information to cue the ordering of laboratory tests. Since numerous tests are available and the ease of performing some tests vary from laboratory to laboratory, physicians may order tests from a standard form listing the alternatives. Thus, allowing students to select laboratory tests from a list, especially if the cost of the test is given, may be an appropriate way to test their ability. Forming a diagnosis or developing a problem list from given alternatives, on the other hand, may not be appropriate, since these are cued from previous information and not a standard list. If all steps of clinical competency are to be evaluated, the instrument may be composed of a combination of student-generate and student-select items depending on the skills involved.

Yet, the most frequently used item formats for certification of students and licensing of health professionals are multiple choice questions and PMP's. Both of these formats allow the examinees to select from a list and employ the processes of recognition, cueing, and elimination. Multiple choice questions and PMP's, therefore, may not be appropriate for certification and

licensing, due to the relationship between the intended purposes of the exam and the item format. Item formats which allow examinees to employ the processes of recognition and elimination may evaluate performance which is less than what is required in actual clinical practice and item formats may inflate examinee scores through cueing. Also, these item formats may not simulate reality as closely as possible. More open-ended item formats requiring generation of answers might be more appropriate for certain sections of certifying and licensing exams since they simulate reality better.

In conclusion, the appropriateness of questions and test formats should be one of the primary considerations in designing an examination, rather than ease of administration and scoring. The level of difficulty of a test is relative depending on the level of discrimination required and on the students' abilities. If a test is too difficult, it cannot discriminate those at the lower end of the distribution. If it is too easy, it cannot discriminate those at the upper end of the distribution. Easier tests may be more appropriate for beginning clinical students. Allowing students to select responses makes the test easier and, therefore, may help to discriminate students at the lower end of the distribution. This may be especially helpful for diagnostic examinations. Thus, the results of this study, together with the scoring convenience factor seem to indicate that student selected item formats are appropriate for evaluating selected types of clinical competence, especially for beginning students or for diagnostic purposes. However, selecting answers may not be appropriate for all examinations of clinical competence.

Table 1

Percentage Point Credit Analyses by Problem and by Group
Showing z Values Which Resulted From the Test of the
Difference Between the Two Proportions

Type and Name of Problem	Percent Completely Correct		Percent Unintegrated Partial Correct		Percent Unintegrated No Credit		Percent Wrong No Credit		Percent Over-resolved No Credit		Percent Omitted Answer	
	Select Group	Generate Group	Select Group	Generate Group	Select Group	Generate Group	Select Group	Generate Group	Select Group	Generate Group	Select Group	Generate Group
Major Acute Problems												
Angina Pectoris	53%	28%	00%	03%	00%	00%	00%	00%	00%	00%	47%	72%
z valve	5.623*		-2.644%								-5.598*	
Arterioscleratic Cardiovascular disease with various symptoms	76%	07%	52%	100%	70%	80%	02%	00%	40%	15%	00%	00%
z valve	20.847*		-14.441*		-2.472*		2.148*		6.200*			
Inguinal Pain	30%	31%	00%	01%	00%	00%	00%	00%	00%	00%	70%	68%
z valve	.231		.662								.460	
Herniorrhophy	03%	03%	00%	03%	00%	00%	03%	00%	01%	00%	93%	94%
z valve			2.641*				2.644		.662		-.431	

* = significant $p < .05$

Table 1 (Continued)

Type and Name of Problem	Percent Completely Correct		Percent Unintegrated Partial Correct		Percent Unintegrated No Credit		Percent Wrong No Credit		Percent Over-resolved No Credit		Percent Omitted Answer	
	Select Group	Generate Group	Select Group	Generate Group	Select Group	Generate Group	Select Group	Generate Group	Select Group	Generate Group	Select Group	Generate Group
Major Chronic Problems												
Diabetes mellitus -												
controlled	97%	43%	04%	52%	00%	01%	03%	01%	02%	00%	03%	01%
z valve	15.503*		-12.493*		-.662		1.522		2.148*		1.522	
Bilateral inguinal hernias	88%	94%	62%	08%	00%	16%	00%	01%	00%	01%	00%	00%
z valve	-1.102		-14.599*		-6.561*		.662		-.662			
Minor Chronic Problems												
Bilateral basal rules	66%	54%	02%	10%	00%	00%	00%	00%	00%	00%	32%	36%
z valve	2.624*		-3.633*								-1.725	
Cigarette Smoking	90%	84%	00%	00%	00%	00%	00%	00%	00%	00%	10%	16%
z valve	1.904										-1.904	
Nausea	12%	00%	00%	00%	00%	00%	00%	00%	00%	00%	88%	100%
z valve	5.554*										-5.551*	

* = significant $p < .05$

Table 1 (Continued)

	Percent Completely Correct		Percent Unintegrated Partial Correct		Percent Unintegrated No Credit		Percent Wrong No Credit		Percent Over-resolved No Credit		Percent Omitted Answer	
	Select Group	Generate Group	Select Group	Generate Group	Select Group	Generate Group	Select Group	Generate Group	Select Group	Generate Group	Select Group	Generate Group
Minor Chronic Problems (Continued)												
Optic fundi with arteriolar nicking	75%	42%	10%	37%	00%	00%	00%	01%	00%	00%	15%	20%
z valve	7.556*		-7.141*				.662				-1.402	
Pigmented raised skin lesion	73%	82%	00%	10%	04%	02%	14%	00%	00%	00%	09%	06%
z valve	-2.305*		-5.011*		1.240		6.066*				1.213	
Family history of Diabetes mellitus, heart disease	81%	59%	00%	00%	00%	12%	00%	12%	00%	00%	19%	29%
z valve	5.2575*				-5.551*		-5.551				-2.422*	
Past history of TUR of prostate	96%	79%	01%	15%	00%	01%	00%	01%	00%	01%	03%	04%
z valve	5.654*		-5.678*		-.662		.662		.662		.058	
Wrong problems indicated					62%	22%						
					10.433*							

* = significant $p < .05$

References

Anderson, R.C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Berner, E.S. Summary of results Phase II Diagnostic Exam - Spring 1976. Abraham Lincoln School of Medicine, University of Illinois at the Medical Center. Chicago: University of Illinois, 1976.

Berner, E.S. Summary of results Phase II Diagnostic Exam - Spring 1977. Abraham Lincoln School of Medicine, University of Illinois at the Medical Center. Chicago: University of Illinois, 1977.

Elstein, A.S.; Shulman, L.S.; and Sprafkra, S.A. Medical Problem Solving. Cambridge: Harvard University Press, 1978.

Feinstein, A. Clinical Judgment. Baltimore: Williams and Wilkins, Co., 1967.

Helfer, R.E. and Slater, C.H. Measuring the process of solving clinical diagnostic problems. British Journal of Medical Education, 1971, 5, 48-52.

Kintsch, W. Learning Memory and Conceptual Processes. New York: John Wiley and Sons, Inc., 1970.

Loftus, G.R. and Loftus, H.M. The Processing of Information. Hillside, N.J.: Lawrence Erlbaum Associates, 1976.

McCarthy, W.H. An Assessment of the Influence of Cueing Items in Objective Examinations. Journal of Medical Education, 1966, 41, 263-266.

Weed, L. Medical Records, Medical Education, and Patient Care. Cleveland: Case Western Reserve University Press, 1971.