DOCUMENT RESUME

ED 259 033 TM 850 422

AUTHOR	Simpson, Deborah E.; Cohen, Elsa B.
TITLE	Problem Solving Questions for Multiple Choice Tests: A Method for Analyzing the Cognitive Demands of Items.
PUB DATE	Mar 85
NOTE	11p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 11985).
PUB TYPE	Spaeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Abstract Reasoning; Cognitive Objectives; Difficulty Level; *Graduate Medical Education; Graduate School Faculty; Higher Education; *Item Analysis; *Multiple Choice Tests; *Teacher Made Tests; Test Construction; *Test Items

ABSTRACT

This paper reports a multi-method approach for examining the cognitive level of multiple-choice items used in a medical pathology course at a large midwestern medical school. Analysis of the standard item analysis data and think-out-loud reports of a sample of students completing a 66 item examination were used to test assumptions related to the differences in cognitive demands pertinent to higher versus lower level multiple-choice items. Items answered by recalling information based exclusively on course content were coded as "knowledge." Items requiring reformulation of course information were coded as "thinking." The validity of the items' cognitive level categorization was assessed by item analyese data (item difficulty, item discrimination, and homogeneity of variance) and categorization of the think aloud responses of 12 students. Results indicated that thinking items were significantly more difficult than knowledge items. Discrimination, when difficulty was held constant, was significantly greater for knowledge items. It was concluded that faculty do write items which assess student ability to reason with what they know; and the method presented can be used by faculty to test their own judgment about the cognitive level of their test items. (BS)

ED259033

BEST COPY AVAILABLE

PROBLEM SOLVING QUESTIONS FOR MULTIPLE CHOICE TESTS: A METHOD FOR ANALYZING THE COGNITIVE DEMANDS OF ITEMS

Deborah E. Simpson
Elsa B. Cohen

The Medical College of Wisconsin

ABSTRACT

This paper reports a multi-method approach for examining cognitive levels of multiple-choice items used in a medical pathology course at a large mid-western medical school. Analysis of the standard item analysis data and think-out-loud reports of a sample of 200 students completing the examination were used to test assumptions related to the differences in cognitive demands pertinent to higher vs lower level multiple-choice items.

**Problem Solving Questions for Multiple Choice Tests:
A Method for Analyzing the Cognitive Demands of Items***

**Deborah E. Simpson, Ph.D.
Elsa B. Cohen, M.D.
The Medical College of Wisconsin****

INTRODUCTION

Medical educators are becoming increasingly concerned that medical school graduates may not be able to use their knowledge to reason efficiently and effectively in a clinical situation (1). One measure of this concern is that the recent AAMC Project Report on the General Education of the Physician (GPEP) urges medical educators to reform their curricula and their evaluation practices in order to emphasize medical problem solving and clinical reasoning (2).

In the preclinical years of medical school, one of the most heavily-used forms of evaluation is the formal examination. As Echina and others, have argued, such "examinations determine how students study (and) what they will learn..." The most frequent criticism of such examinations is that they emphasize memorizing facts rather than thinking and applying these facts. This criticism is often connected with examinations which are composed of multiple-choice items. Giving essay examinations might make it easier to test the ability of a student to synthesize material and solve problems. However, as the GPEP report acknowledges, reform of evaluation systems is constrained by a number of factors, among which are large classes, increasing amounts of information in the bio-medical fields, and the fact that in most medical schools, faculty members are rewarded not for teaching, but for research and patient care (2). These constraints, coupled with the much greater convenience and apparent "objectivity" of multiple-choice examinations seem to mean that the use of multiple-choice examinations will continue. Therefore, an important component of improving the evaluation system, particularly in the preclinical years, is to develop multiple-choice items which test the student's reasoning abilities.

Most approaches to working with faculty on multiple-choice item writing present step-by-step procedures for generating higher level items (6, 7, 8). These approaches assume that faculty do not already generate items which require students to reason. That assumption has not been tested. There are many faculty members who feel that they do write items at the higher cognitive levels. Therefore, to test faculty's beliefs about their item writing ability, an examination of the cognitive demands of test items actually used by faculty must be conducted (9).

It is one thing for the author of an item to believe that he or she has written a question which requires the student to analyze and synthesize material which has been presented in a course. It is nother to decide whether such an item does in fact elicit this behavior from the student.

Requests for reprints should be addressed to: *Office of Educational Services, The Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin 532226

**The authors thank Michael Donnelly, Ph.D., Carol Kuhlmann, M.S. and Kathleen Yindra, M.S. for their assistance in the data analysis.

PURPOSE

This paper reports a multi-method approach for examining the cognitive level of multiple-choice items generated and used by the course director of medical pathology at a large midwestern medical school. Methodology and data analysis approaches were selected to test assumptions related to the differences in cognitive demands related to higher vs. lower level multiple- choice items. More specifically, higher cognitive level items were expected to have higher difficulty and discrimination indices due to their added cognitive demands in comparison to lower level items (10, 11).

METHODOLOGY

Sixty-six items used on pathology course examinations during the 1983-1984 academic year were used in this study. The course director, who was the author of these items, was asked to review each item answering the following questions. First, was the content of the item isomorphic with the content presented in lecture, lab, handouts, or readings? In other words, could the student answer the question correctly by recalling information which was based exclusively on course content? If the answer to this question was no, the course director was asked, "How must the student use the course information to answer the item?" Based on Bloom's Taxonomy of Cognitive Objectives (12), the items were broadly categorized as "knowledge" items (knowledge, comprehension, application) or "thinking" items (analysis, synthesis, evaluation). Items were coded as "knowledge" if the answer to the first question was yes. For example the following item was taken directly from a list found in the required readings:

> In the USA, the type of heart disease most often responsible for death is: (1) hypertensive (2) cogenital (3) traumatic (4) rheumatic (5) none of the above.

If an item required the reformulation of information presented in the course it was coded as "thinking." In other words, thinking items required the student to analyze synthesize, and/or evaluate course information. An example of this type of question is as follows:

> A 60 year old diabetic man with long-standing history of angina enters with chest pain and shortness of breath of two hours duration. He has rales half-way up both lung fields. The most likely explanation for this history and findings is: (1) severe angina with reflex bronchospasm (2) bilateral pulmonary emboli with infarcts (3) acute pulmonary edema secondary to myocardial necrosis (4) pneumococcal pneumonia superimposed on pulmonary edema (5) acute pancreatic necrosis with sympathetic pneumonitis.

The coding resulted in 47 knowledge items and 19 thinking items. Item analysis data from 1983-84 academic year was available for each item.

Six weeks after the course concluded, twelve students were contacted and asked to "think-out-loud" as they answered 7 questions from the final examination. Nine students (4 from the upper 1/3 of the class, 3 from the middle 1/3 of class, and 2 from the lower 1/3 of class) completed the task (1). Students were told that the investigators were interested in how they approached the problem rather than in the correctness of their response. Three "knowledge" questions and four "thinking" questions were selected on the basis of their congruency with the mean difficulty index for items in their respective categories (1). The interviewer was blind to performance ranking of the students.

ANALYSIS OF DATA

The validity of the items' cognitive level categorization was assessed from two perspectives: item analysis data and categorization of the think-out-loud responses of students. More specifically, three indices from a standard item analysis print-out were used to analyze the differences between cognitive item categories: item difficulty, item discrimination, and homogeneity of variance within item responses. One-hundred ninety seven students answered each of the 66 items. To determine if differences in item difficulty and discrimination by item type were significant, t-tests were conducted. In addition, homogeneity of variance was examined using a repeated measures two-way analysis of variance to determine significant differences by student performance on the examination (high, medium, low) and type of item (knowledge, thinking).

Students' think-out-loud responses were analyzed to determine if they exhibited reasoning en route to problem solution or recall/recognition of information. The rater was blind to the performance level of students. Each student response was categorized as "knowledge" or "thinking."

RESULTS

Thinking items were significantly more difficult than knowledge items $t(64) = 8.058$ p .01 with a mean difficulty of .62 for thinking items and .88 for knowledge item (the higher the index, the easier the item). No significant differences in discrimination by item type was obtained ($p > .05$). See Table 1.

Given the relationship of difficulty to discrimination (12) and the expectation that differences in discrimination should occur by item type (i.e., knowledge items being less cognitively demanding resulting in less variability in student responses), an additional analysis was conducted. In order to determine if item difficulty was attenuating the correlation between discrimination and group, a partial correlation was calculated between discrimination and group, holding difficulty constant. This correlation was equal to -.30 ($p < .05$). Note that the zero order correlation was not significant. Unexpectedly however, knowledge items, despite their being less difficult, were more discriminating than thinking items.

Results of the two-way repeated measures analysis of variance for item homogeneity of variance revealed significant differences by item type, $F(1, 64) = 64.302$, p .001 and performance level $F(2,128) = 125.21$, $p < .0001$. See Table 2. Significant interaction effect was also obtained $F(2,128) = 4.053$, $p < .02$.

Figure 1 illustrates the interaction effects for homogeneity of variance scores by item type and performance levels. On thinking items, as compared to knowledge items, middle performance students exhibited more variability in the selection of response alternatives than did high or low performing students. Follow-up analysis using Tukey A contrasts resulted in significant differences between all three performance levels ($p < .01$).

Ratings of the "think-out-loud responses" indicated that students did reason on the 4 items pre-categorized as requiring thinking with knowledge. Of the 36 possible responses (4 items x 9 students, 35 were scored as thinking. The one knowledge response was made by a student who indicated that he "couldn't remember if any of these conditions would cause CHF."

Results of student think-out-loud responses for knowledge items are less straight forward as students indicated that they "did know that on the exam, but just can't remember it right now." Following these self-reported memory losses, students would attempt to reason through the "knowledge" question. For example, in one question which involved remembering the physical conditions in which alcoholism is a known risk factor, students could not recall the connection between alcoholism and pancreatic pseudocysts. Students did recall however, the connection between alcoholism and pancreatitis and used this information to evaluate the relationship of pancreatic pseudocysts to alcoholism.

In summary, students' cognitive responses matched the hypothesized cognitive item demands. All items categorized as requiring thinking for correct solution elicited reasoning responses from students. Memory items, by student self-report, could be identified but not answered by recall alone, due to forgetting.

DISCUSSION OF RESULTS

As expected, a significant difference for item difficulty by type of item was obtained. Students' performance on thinking items was significantly lower than performance on knowledge items. By definition, knowledge items required students to recall/recognize course information in order to correctly respond to the questions. Thinking items required transformation of that knowledge to answer a question. The differences in item difficulty, along with the results of the students' think-out-loud responses, are consistent with the theoretical expectations about the differences in the cognitive demands of the 2 item types. The results also support the assumption that the author of the items could intentionally write questions which demand not just knowledge, but also the ability to think from that knowledge.

Unexpectedly, discrimination when difficulty was held constant, was significantly greater for knowledge items as compared to thinking items. At least two factors may contribute to this finding. First, knowledge items had a greater between-item variability, although lower within item variability compared to thinking items. This variability difference would increase the potential for knowledge items to discriminate more/than thinking items between individuals.

A second factor affecting discrimination of knowledge items may be related to the expectations of students regarding the task of studying for examinations. Students may not prepare for tests which include items that require them to think with course information. Most of their energies may go into memorizing "facts" in preparation for the kinds of questions which they expect to be asked (this strategy would be appropriate for over 60% of the items used in the study). When the appropriateness of study strategy is examined in the context of the impossibility of memorizing all the material which is presented in a course such as medical pathology, one important kind of problem solving, from the students' point of view is to decide <u>what</u> to memorize. The varying abilities of students to decide what to memorize may be reflected in the increased discrimination levels of knowledge items.

CONCLUSIONS AND IMPLICATIONS

This study began with the assertion that one way to encourage probler solving behavior in medical students is for facul to generate multiple-choice questi which require students to think and not just memorize. Two outcomes of this stu e particular attention. First, the results indicate that, contrary to what is imes assumed, faculty do write items which assess the ability of students to use and reason with what they know. Second, the study provides a method which faculty can use to test their own judgement about the cognitive level of their items. This method rests on two distinct elements: a study of the item analysis data readily available to most faculty and analysis of the think-out-loud responses of students.

These results do not address the question of directly rewarding faculty for encouraging problem solving behavior in students. Nor do they address the question of deciding how much emphasis should be placed in problem solving ability in any given course. However, they do provide a way for faculty members to analyze their questions so that they can, given the current constraints of the system, reward the student for the kind of problem solving behavior which most medical educators currently believe must be encouraged.

Table 1

Mean item difficulty index and mean item discrimination index by item type.

	Difficulty	Discrimination
Knowledge	.884*	.249
Thinking	.623	.266

* p .01

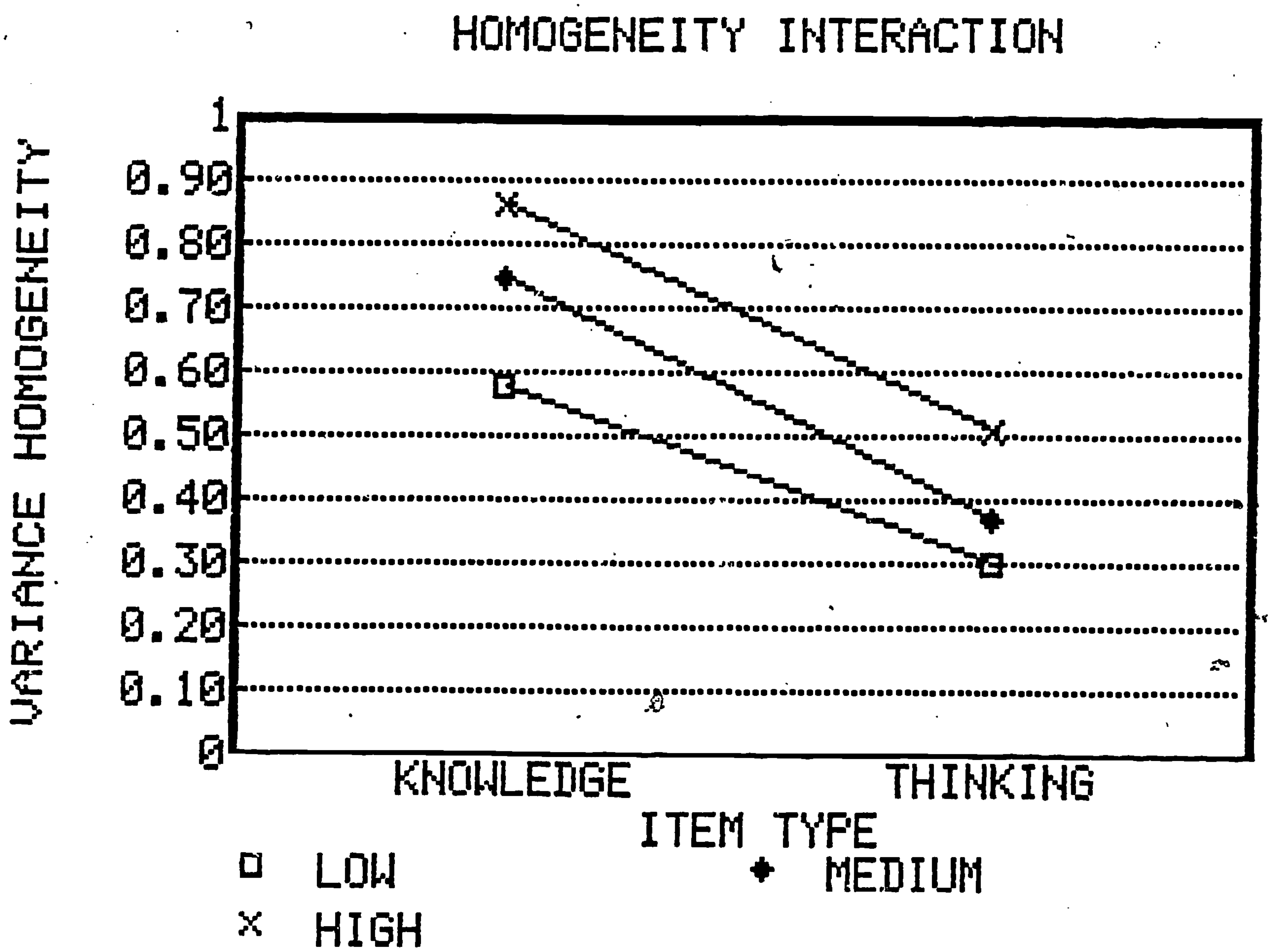
Table 2

Mean variance for within item homogeneity of varia ce by item type and studenc performance level.

Item Type*	Student Performance Level**		
	Low	Medium	High
Knowledge	.58	.75	.75
Thinking	.31	.37	.51

* p .001

** p .0001


HOMOGENEITY INTERACTION
VARIANCE HOMOGENEITY
1
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0
KNOWLEDGE
THINKING
ITEM TYPE
LOW
MEDIUM
HIGH

REFERENCES

1. Research in Medical Education: 1983 Proceedings of the Twenty-Second Conference. Washington, D.C. Nov. 8/9, 1983.

2. AAMC Panel on General Professional Education of the Physician and College Preparation for Medicine, "Emerging Perspectives on the General Professional Education of the Physician." Association of American Medical Colleges October, 1983.

3. Eichna, L. W. "A Medical School Curriculum for the 1980's", New England J. Med., 308: 18-21, January 6, 1983.

4. Bok, D.C. "Excerpts from Bok's Report to Harvard Overseers on Medical Education", Chronicle of Higher Education. Vol 28(10) pp. 14-16 May 2, 1984.

5. Dornhorst, A.C. "Information Overload: Why Medical Education Needs a Shake-up", Lancet, Sept. 5, 1981. pgs 513-4

6. Nu Viet Vu, C. & Cline, W.H. "A Technique to Guide Faculty to Write Exam Items at the Higher Cognitive Levels." Paper presented at Annual Conference for Generalists in Medical Education, 1980.

7. Miller, H. G. Williams, R. G. and Haladyne, T. M. Beyond Facts: Objective Ways to Measure Thinking. Educational Technology Publications: Englewood Cliffs, NJ, 1978.

8. Koewing, J. R. Construction Guidelines for the Single Best Answer Multiple-Choice Question. Health Sciences Consortium: Chapel Hill, N.C. 1979.

9. Carrier, C. A. Dalgaard, K. A., Simpson, D. E. "Theories of Teaching: Foci for instructional improvement through consultation." Review of Higher Education, Vol 6(3), Spring 1983, pgs 195-206.

10. Nu Viet Vu, C. "Medical learning processes and Their Relation to Test Performance and Attitude Toward Item Formats" Proceedings Eighteenth Annual Conference on Research in Medical Education, November, 1979. pgs 311-316.

11. Yinger, R. J., "Can We Really Teach Them to Think? in R. E. Young (ed) Fostering Critical Thinking Jossey Bass Inc: San Francisco, 1980. pgs 11-31.

12. Bloom, B.S. (Ed) Taxonomy of Educational Objectives Handbook: Cognitive Domain. New York: David McKay, 1956

13. Gronlund, N.E. Measurement and Evaluation in Teaching. Macmillan Publishing Co., Inc: New York, 1976.