DOCUMENT RESUME

ED 304 479 TM 012 881

AUTHOR	Case, Susan M.; And Others
TITLE	Evaluating Diagnostic Pattern Recognition: The Performance Characteristics of a New Item Format.
PUB DATE	Feb 88
NOTE	20p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (Miami Beach, FL, February 24-27, 1988).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Clinical Diagnosis; Higher Education; *Item Analysis; Medical Evaluation; *Medical Students; *Pattern Recognition; Physicians; Test Construction; *Test Format; Test Reliability; Timed Tests

ABSTRACT

An item format incorporating pattern recognition was designed to assess medical students' abilities in the area of clinical diagnosis. A group of approximately 20 faculty members of five New England medical schools met in Worcester for half of a day to develop pattern recognition items. Teams of four to six physicians were assigned to work on particular topic areas that represent common chief complaints of patients. They developed a list of approximately 15 common diagnoses that relate to each of the topics. An item describing a patient by listing critical signs and symptoms was developed for each of the diagnoses. Approximately 300 items divided into 21 sets were developed and, subsequently, edited and reviewed by independent physicians before test administration. A modified Angoff procedure was used to set pass/fail standards for the set of items. A total of 336 fourth-year medical students from the five schools were tested using the items. The 21 sets of pattern recognition items were completed by between 112 and 332 examinees. Results indicate that: (1) students performed well on the items--the mean score was 82% correct, and almost 66% of the students passed at least 90% of the sets they took; (2) generalizability analyses indicated that performance in one topic area did not predict performance in other areas very well; and (3) 2 hours of testing time would be required to generate a reasonably reliable score. Four tables and eight figures are provided. (TJH)

Evaluating Diagnostic Pattern Recognition: the Perfomance Characteristics of a New Item Format

Susan M. Case, Ph.D.
David B. Swanson, Ph.D.
Paula Stillman, M.D.

Introduction

Perspective/theoretical framework

Problem solving is traditionally viewed as an application of the scientific method (Fig 1) in which the problem solver formulates tentative hypotheses; collects and synthesizes some data; re-evaluates the hypotheses; and continues with this process, collecting and synthesizing additional data and re-evaluating hypotheses until a solution is found. Although this procedure is believed to be widely used, in some instances the problem solver does not need to follow all these steps. In these cases, the configuration of data elements is so classical that one hypothesis seems to leap into mind almost instantly. This abbreviated form of problem solving has been labeled pattern recognition (Dudley, 1968).

When you think about it, you know that these two forms of problem solving are widely used. We've all had experiences with car mechanics where we describe a set of occurrences; the mechanic asks some questions; he says it might be this or that; he looks under the hood, tries a few things, and says "I think it might be ---, but I have to take the engine apart to be sure." On the other hand, if you say "my car has been increasingly hard to start, and this morning it wouldn't start at all and the headlights are dim." He might say, "aha, it's the battery."

It's clear that some sets of data present a pattern and others don't (Fig 2). It's also clear that data in the hands of an expert mechanic (Fig 3) might generate an immediate "aha"; the same data given to a novice might generate some head-scratching and comments about taking the engine apart.

Pattern recognition is the technique that physicians use most often in arriving at a medical diagnosis. As is true in other professions, expertise in medical diagnostic pattern recognition seems to come with experience. Medical students quite properly view each case as a new experience while senior clinicians have seen some diseases so frequently that the diagnoses appear obvious. A major purpose of clinical training is to provide the concentrated experience necessary for the development of pattern recognition skills.

Presented at the Eastern Educational Research Association (EERA) Annual Conference, Miami Beach, February 25, 1988. Please direct questions to Susan M. Case, Ph.D., Senior Evaluation Officer, National Board of Medical Examiners, 3930 Chestnut Street, Philadelphia, PA 19104.

Project background

A number of years ago, after we had the notion about these two forms of problem solving, we developed a few items (Fig 4) that we thought would get at "pattern recognition" skills (ie, would test the ability to synthesize data and determine the correct diagnosis). Each item briefly described a patient by listing a few critical signs and symptoms that were designed to clearly reflect a particular diagnosis. Answers were to be selected from an alphabetical list of diagnoses; the same list was used for all items. A sample 10-item test was administered to second and fourth year medical students and medical residents (Case and Fabrey, 1984). Results showed a clear difference among groups in the expected direction ($F = 100.9$, $p<.001$). The items were answered much more quickly than standard format multiple choice questions. Examinees indicated that the list of signs and symptoms in each item did form a pattern which was immediately apparent to those who knew the correct answer.

It was the magnitude of the differences between groups, the speed with which the items could be answered, and the positive reaction to the items by the participants that led us to investigate this area further. The purpose of this study was to investigate the performance characteristics of the item format with a larger set of items and subjects.

Method

Item and Examination Development

A group of approximately 20 faculty members from five New England medical schools met in Worcester for half of a day to develop the pattern recognition items. Teams of four to six physicians were assigned to work on particular topic areas that represent common chief complaints of patients (eg, cough, headache). For each of the topics, they developed a list of approximately 15 common diagnoses that relate to it. For example, diagnoses such as pneumonia and bronchitis were included in the list for the topic/chief complaint of cough. An item describing a patient by listing critical signs and symptoms was developed for each of the diagnoses. During the half day, approximately 300 items divided into 21 sets were developed. These were edited and reviewed by independent physicians before test administration; some of the items were deleted so that there more diagnoses listed than there were items.

Standard Setting

Faculty from the participating schools met to set pass/fail standards for the sets of items. A modified Angoff procedure was used. For each set of items related to a particular topic, they independently classified each item as either "of critical importance", "of moderate importance", or "of minimal importance". The following factors were considered in

their ratings: student exposure to cases similar to that described by the item; general importance of the case described by the item; relevance of the case to the curriculum; and technical quality of the item. After classifying all items in a set independently, the group members discussed their responses and achieved a consensus classification for each item.

They then established pass/fail criteria for each set, considering the expected performance of a hypothetical "borderline" student. After working through several sets, they agreed on a standard that required students to pass 90% of the critical items, 50% of the moderately important items, and none of the items of minimal importance. Item difficulties (p-values) were provided to faculty for use in their deliberations, but these did not appear to have much influence on classification.

A pass/fail standard was derived for each set using the formula:

Pass/fail point = 0.9 X (# of critical items)
+ 0.5 X (# of moderately important items)
+ 0.0 X (# of minimally important items)

Similarly, a pass/fail point for the test as a whole was determined by summing the pass/fail points for the individual sets. Pass/fail points were transformed from number right to percent correct scores for purposes of analysis and reporting scores.

Examinees

A total of 336 fourth year medical students from five New England medical schools participated in the study: 84 students came from School 1, 57 from School 2, 78 from School 3, 92 from School 4, and 25 from School 5. Because of the small number of students from School 5, data from this school were excluded from comparative school analyses.

Test administration Procedure

Pattern recognition items were administered as part of a larger study to assess clinical skills. The students worked through a series of simulated patients who were stationed in individual examining rooms. The students rotated among the rooms taking a history or doing a combined history and physical. Following the work-up of a simulated patient, students took a set of pattern recognition items matched to the chief complaint of the patient that they had just seen. Sets contained between 7 and 12 items (average of 10) related to that particular chief complaint.

During the course of the day, each examinee took approximately 12 sets of items out of the total of 21 sets of items that had been developed. The particular sets taken depended upon the simulated patients included in the "test form" on that day of test administration.

The students were allowed two to three minutes to complete a set. Answers

were recorded directly on the test paper and later key-punched. Because of concern that there may have been insufficient time to complete each set, examinees who left more than half of the items blank in a set were excluded from the analysis of that set.

The 21 sets of pattern recognition items were completed by between 112 and 332 examinees. Percent correct scores were calculated for each student for each set that was completed. A total percent correct score was calculated by dividing the total number of questions answered correctly by the total number of questions in the sets taken. The percentage of sets passed was also calculated for each student.

Results

Percent Correct scores

Table 1 shows the number of students who took each set and the average percent correct score obtained on each of the sets. Mean percent correct scores on the sets ranged from 64 to 95.

Figure 5 shows a frequency distribution of the percentage of questions answered correctly. Individual total percent correct scores ranged from 52% to 97%. The overall mean percent correct score was 82% (SD = 8). Figure 6 shows a boxplot of total percent correct scores broken down by school. Although there were significant differences between schools, mean scores were fairly comparable. The score distributions varied.

Table 3 provides pass/fail rates for each set by school. A school by set analysis of variance on percent correct scores yielded a significant interaction. Apparently, school differences in clinical curricula result in characteristic patterns of strength and weakness in students.

Percentage of sets passed

Table 2 shows the percentage of students who passed each set. These percentages varied from a low of 63% of the students passing the set on Foot Pain to a high of 97% of the students passing the set on Fever in Children. The percentage of students passing each set is not directly related to the difficulty of the set, since pass/fail standards were determined individually for each set based of the importance of items in that set.

Figure 7 shows a frequency distribution of the percentage of sets passed by individual students. The percentage of sets passed varied from 9% to 100%. The mean percentage of sets passed was 80% (SD = 18). Over 64% of the students passed at least 90% of the sets. Figure 8 shows a box plot of the total percentage of sets passed broken down by school. Again, while means were fairly consistent across schools, the distributions of scores varied.

Relationships to other measures

Percent correct scores correlated .20 ($p<.01$) with data gathering scores on the simulated patient component of the test battery and .53 ($p<.01$) with Part I of the NBME taken 12 - 15 months earlier. Percent correct scores were not related ($r = .10$) to measures of student interpersonal skills in dealing with the simulated patients.

Generalizability analyses

To obtain information about the generalizability of the pattern recognition scores, a subset of examinees, sets, and items was selected. This subset included 212 examinees, the six most frequently used sets, and the first nine items in each set in a completely balanced design. A Persons X (Items: Sets) random effects analysis of variance was performed to obtain variance components, and a number of decision studies were done using GENOVA statistical software. The results of this analysis are shown in Table 4.

The pattern recognition item format can be used in two ways. First, general ability to recognize diganostic patterns can be of interest, as was the case in this study. Second, ability to recognize diagnostic patterns for a particular complaint can also be of interest, since the item format could be used to identify specific areas of strength and weakness. Generalizability coefficients for both these situations can be derived from the variance components in Table 4.

The decision studies in the bottom of the table are appropriate if the domain of interest (universe of generalization) is general ability to recognize diagnostic patterns. The test as administered to the typical examinee in the study (roughly 12 sets of 10 items) does not yield very reproducible scores: the domain-referenced generalizability (dependability) coefficient was only 0.66. Inspection of the variance components indicates the reason for this: the Persons X Sets variance component is quite large -- more than twice as large as the Persons component.

Examinees are not very consistent in how well they perform from one set to the next, so extensive sampling of sets is necessary to obtain a reproducible assessment. For example, using 25 sets of 5 items each increases the dependability coefficient to 0.73 with very little increase in overall test length. There may be a point of diminishing return in reducing the number of items per set, however, since the time required per item probably increases (due to the additional reading burden in shifting sets) as the number of items per set decreases. Fifty 5-item sets would yield an acceptable level of reproducibility in a testing time of roughly two hours. Required test length for reproducible norm-referenced interpretation of scores is somewhat less.

If the domain of measurement interest is ability to recognize diagnostic patterns for a particular complaint, 30 - 40 items are required to

achieve reasonably reproducible assessment of performance, depending upon whether domain-referenced or norm-referenced score interpretation is desired. A set of this length would require approximately 15 minutes of testing time. Thus, large scale, complaint-by-complaint assessment of an examinee's strengths and weaknesses would be quite practical, and very specific plans for educational remediation of deficits could be derived from the results of such a test battery.

Discussion

The ability to synthesize data to formulate a diagnosis is an important skill for physicians. Recognizing a pattern in the data appears to be one way that physicians solve problems and this ability seems to be related to clinical experience and expertise.

In general, students performed very well on the items. The mean score was 82% correct and almost two-thirds of the students passed at least 90% of the sets that they took. However, using either percent correct scores or pass/fail standards, students who were outliers on the low end of the distribution could be identified. For example, four students answered less than 60% of the items correctly overall (ie, over 3 SDs below the mean) and 16 students passed less than 50% of the sets. For diagnostic or remedial purposes, performance can be examined by content area to determine specific areas of weakness for individual students.

The issues related to measuring this skill are similar to measuring other clinical skills; how can testing time be used most efficiently to obtain reliable and valid scores? How should tests be constructed to obtain scores that validly reflect individual performance in making diagnosis? In this study, results indicated that it is preferable to sample more presenting complaints with fewer items directed at each one, rather than to sample more items within a small number of presenting complaints.

Generalizability analyses indicated that performance in one topic area does not predict performance in other areas very well. For example, students who were relatively expert in diagnosing patients with headaches tend not to be expert in diagnosing patients with chest pain, joint pain, etc. Approximately two hours of testing time would be required to generate a reasonably reliable score (ie, with a generalizability coefficient greater than 0.80).

The next phase of this study will be directed at two issues. First, an investigation will determine whether the format discriminates among students at different levels of training. A second study will determine the benefits of using the current matching format with a relatively long list of response alternatives over a traditional multiple choice item with five choices. It is hypothesized that the shorter list differentially benefits the lower ability students and the more junior students.

References:

Case, SM & Fabrey, LJ. Development of an experimental examination to measure pattern recognition. Presented at the Eastern Educational Research Association Annual Conference, West Palm Beach, February 10, 1984.

Dudley, H.A.F. Pay-off, heuristics, and pattern recognition in the diagnostic process. The Lancet, September 28, 1968, 723-726.

Table 1

Descriptive Statistics for All Pattern Recognition Sets

TOPIC	N	MEAN	SD
ANEMIA	195	70	20
PEDIATRIC BEHAVIOR PROBLEMS	188	91	12
CHEST PAIN	301	85	17
CONFUSION	306	71	19
COUGH	168	86	13
DIARRHEA	173	64	19
DIZZINESS	305	77	16
EASY BRUISING	130	69	18
FEVER IN CHILDREN	125	94	8
FOOT PAIN	153	87	16
HAND PAIN	137	82	16
HEADACHE	169	76	18
JAUNDICE	142	73	17
JOINT PAIN	304	87	15
LOW BACK PAIN	257	78	16
MENSTRUAL DISTURBANCES	200	95	8
SAD AFFECT	332	91	12
SHORTNESS OF BREATH	131	82	10
OCCUPATIONAL RISKS	313	87	14
URINARY FREQUENCY	151	82	17
VAGINAL DISCHARGES/LESIONS	112	85	16
TOTAL	336	82*	8

*Average percent correct scores across examinees

Table 2

Pass/Fail Rates for All Pattern Recognition Sets

TOPIC	N	MEAN	SD	% PASS
ANEMIA	195	70	20	65
PEDIATRIC BEHAVIOR PROBLEMS	188	91	12	78
CHEST PAIN	301	85	17	83
CONFUSION	306	71	19	79
COUGH	168	86	13	75
DIARRHEA	173	64	19	76
DIZZINESS	305	77	16	77
EASY BRUISING	130	69	18	82
FEVER IN CHILDREN	125	94	8	97
FOOT PAIN	153	87	16	63
HAND PAIN	137	82	16	85
HEADACHE	169	76	18	70
JAUNDICE	142	73	17	87
JOINT PAIN	304	87	15	79
LOW BACK PAIN	257	78	16	84
MENSTRUAL DISTURBANCES	200	95	8	86
SAD AFFECT	332	91	12	88
SHORTNESS OF BREATH	131	82	10	89
OCCUPATIONAL RISKS	313	87	14	83
URINARY FREQUENCY	151	82	17	90
VAGINAL DISCHARGES/LESIONS	112	85	16	76
TOTAL	336	82	8	92*

*Percentage of examinees passing their test

Table 3

Pass/Fail Rates for Each Set by School

			SCHOOL							
	ALL		1		2		3		4	
TOPIC	N	%	N	%	N	%	N	%	N	%
-----	---	---	---	---	---	---	---	---	---	---
ANEMIA	195	65	69	57	28	71	16	88	59	64
PEDIATRIC BEHAVIOR PROBLEMS	188	78	64	63	28	93	23	83	56	84
CHEST PAIN	301	83	68	81	57	93	72	75	85	84
CONFUSION	306	79	83	72	48	85	74	80	78	78
COUGH	168	75	37	70	41	76	60	70	30	90
DIARRHEA	173	76	34	76	27	85	60	67	52	83
DIZZINESS	305	77	78	72	52	85	63	70	87	83
EASY BRUISING	130	82	13	92	28	79	59	83	30	80
FEVER IN CHILDREN	125	97	15	100	27	89	54	100	29	97
FOOT PAIN	153	63	36	69	14	57	51	57	27	74
HAND PAIN	137	85	23	83	15	87	52	90	23	87
HEADACHE	169	70	38	55	42	69	26	73	63	79
JAUNDICE	142	87	40	80	29	90	16	94	34	82
JOINT PAIN	304	79	72	76	50	82	78	69	90	88
LOW BACK PAIN	257	84	84	82	43	81	34	88	77	88
MENSTRUAL DISTURBANCES	200	86	51	82	21	86	35	91	68	88
SAD AFFECT	332	88	84	77	57	96	77	94	90	87
SHORTNESS OF BREATH	131	89	39	79	7	100	15	100	50	88
OCCUPATIONAL RISKS	313	83	75	83	56	88	71	82	91	85
URINARY FREQUENCY	151	90	21	90	43	86	25	92	62	92
VAGINAL DISCHARGES/LESIONS	112	76	22	73	36	69	32	81	22	82

Table 4

Results of Generalizability Analyses

EFFECT	DEGREES OF FREEDOM	VARIANCE COMPONENT	STANDARD ERROR
Persons	211	0.0040473	0.0007397
Sets	5	0.0044785	0.0034372
Items: sets	48	0.0163487	0.0033693
Persons: sets	1055	0.0088557	0.0009101
Persons X Items: Sets	10128	0.1054848	0.0014822

NO. OF SETS	ITEMS PER SET	UNIVERSE SCORE VARIANCE	EXPECTED OBSERVED SCORE VARIANCE	NORM-REF ERROR VARIANCE	DOMAIN-REF ERROR VARIANCE	NORM-REF GENER COEFF	DOMAIN-REF GENER COEFF
1	1	0.00405	0.11839	0.11434	0.13517	0.03419	0.02907
1	5	0.00405	0.03400	0.02995	0.03770	0.11904	0.09694
1	10	0.00405	0.02345	0.01940	0.02552	0.17258	0.13689
1	20	0.00405	0.01818	0.01413	0.01943	0.22266	0.17242
6	1	0.00405	0.02310	0.01906	0.02253	0.17518	0.15230
6	5	0.00405	0.00904	0.00499	0.00628	0.44774	0.39177
6	10	0.00405	0.00728	0.00323	0.00425	0.55584	0.48761
6	20	0.00405	0.00640	0.00235	0.00324	0.63216	0.55557
12	1	0.00405	0.01358	0.00953	0.01126	0.29813	0.26433
12	5	0.00405	0.00654	0.00250	0.00314	0.61853	0.56298
12	10	0.00405	0.00566	0.00162	0.00213	0.71452	0.65556*
12	20	0.00405	0.00522	0.00118	0.00162	0.77463	0.71430
25	1	0.00405	0.00862	0.00457	0.00541	0.46947	0.42810
25	5	0.00405	0.00525	0.00120	0.00151	0.77159	0.72854
25	10	0.00405	0.00482	0.00078	0.00102	0.83908	0.79860
25	20	0.00405	0.00461	0.00057	0.00078	0.87746	0.83893
50	1	0.00405	0.00633	0.00229	0.00270	0.63897	0.59954
50	5	0.00405	0.00465	0.00060	0.00075	0.87107	0.84296
50	10	0.00405	0.00444	0.00039	0.00051	0.91250	0.88802
50	20	0.00405	0.00433	0.00028	0.00039	0.93473	0.91241

*Generalizability (dependability) coefficient for the test actually given in this study.

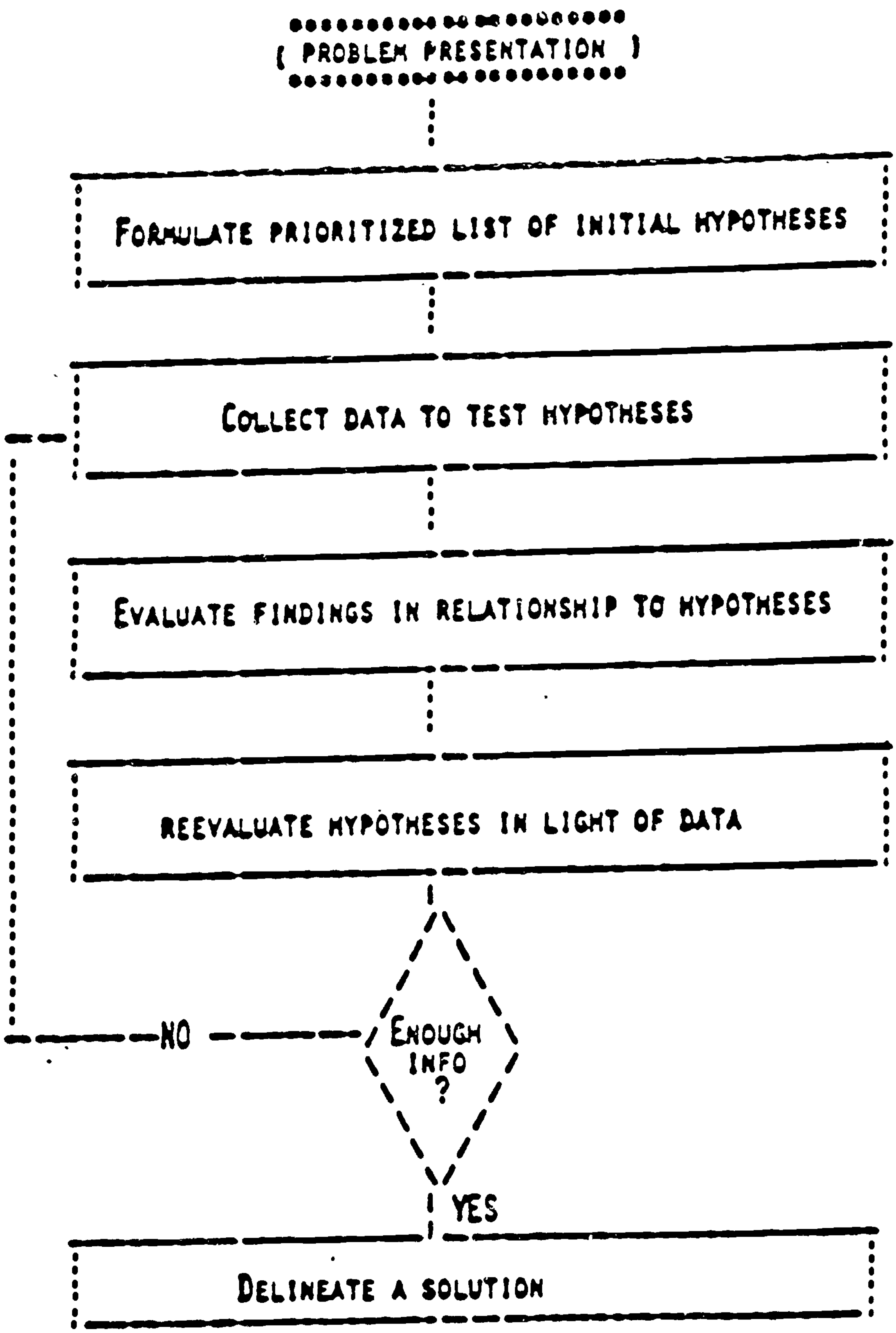

Figure 1. The scientific method.

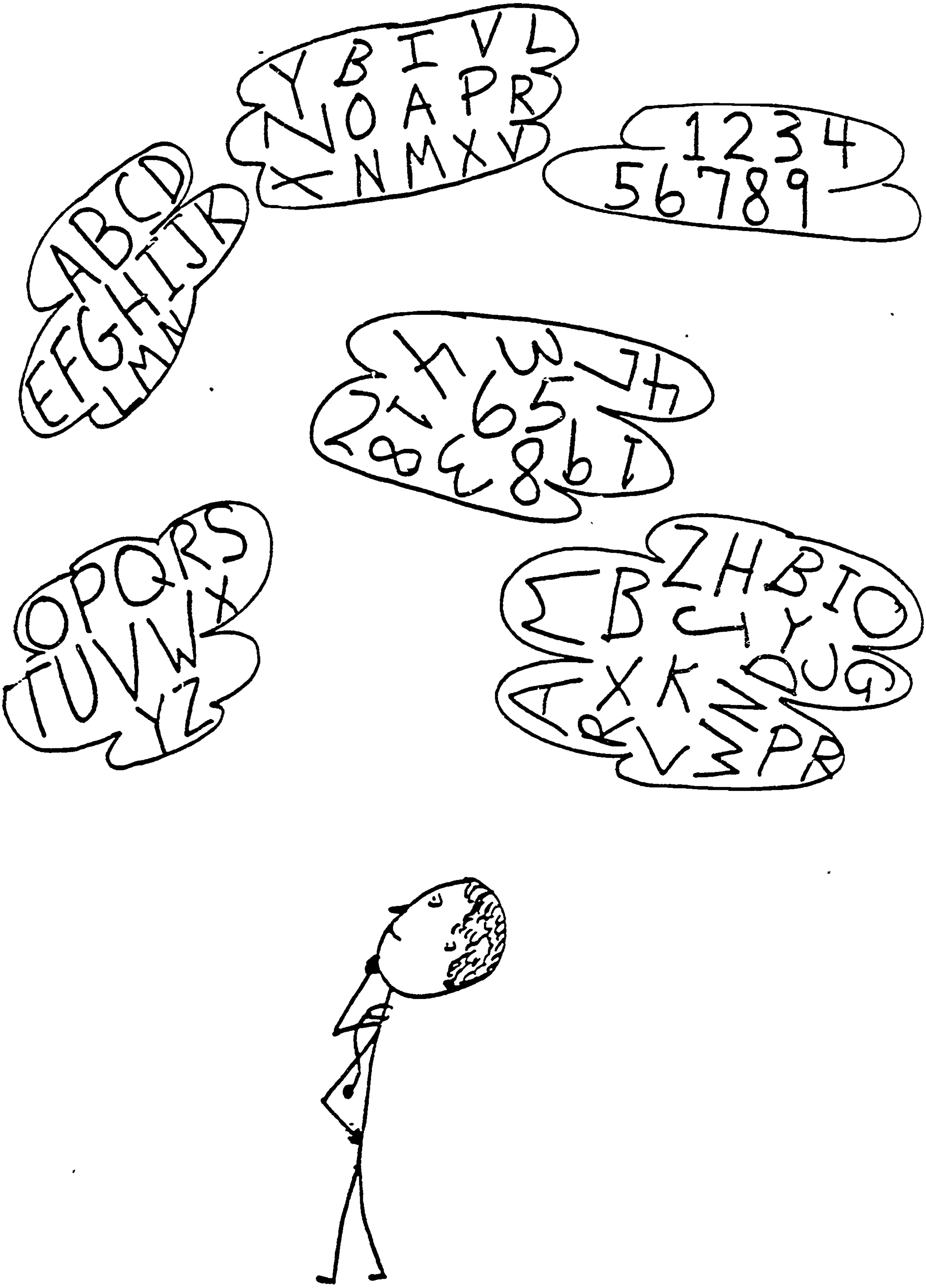

Figure 2. Patterns in data sets.

Figure 3. Expert/novice analyses of data.

For each item listed below (numbers 1-5) select the single best diagnosis (letters A-L). Each diagnosis may be used once, more than once or not at all.

CHEST PAIN CASE 502

A. Angina - stable
B. Angina - unstable
C. Aortic dissection
D. Aortic stenosis
E. Cancer - lung
F. Embolism - pulmonary
G. Herpes zoster
H. Pericarditis
I. Pneumonia
J. Pneumothorax
K. Rib fracture
L. Tuberculosis

1. A 52-year-old man has recurrent, predictable, achy chest discomfort on taking his morning walk; symptoms are relieved by rest

2. A 48-year-old woman who smokes has had increasingly frequent exertional and nocturnal chest discomfort radiating to left arm for three weeks

3. A 30-year-old man has fever, symptoms of upper respiratory infection and nonradiating precordial pain relieved by sitting up and leaning forward

4. An 18-year-old athlete has sudden onset of right-sided pleuritic pain, shortness of breath, and decreased breath sounds on the right

5. A 53-year-old man has fever, chills, right lower pleuritic chest pain, purulent sputum, and bronchial breath sounds over the right lower lobe

Figure 4. Sample diagnostic pattern recognition items

Figure 5

Distribution of Percent Correct Scores

```
 N    SCORE
---   -----
  1     52    *
  1     54    *
  1     56    *
  1     58    *
  1     60    *
  5     62    *****
  4     64    ****
  5     66    *****
  8     68    ********
 13     70    *************
 14     72    **************
 14     74    **************
 29     76    *****************************
 17     78    *****************
 33     80    *********************************
 27     82    ***************************
 36     84    ************************************
 42     86    ******************************************
 24     88    ************************
 26     90    **************************
 22     92    **********************
  9     94    *********
  2     96    **

            I....+....I....+....I....+....I....+....I....+....I
            0        10        20        30        40        50
```

Figure 6

Percentage Questions Answered Correctly by School

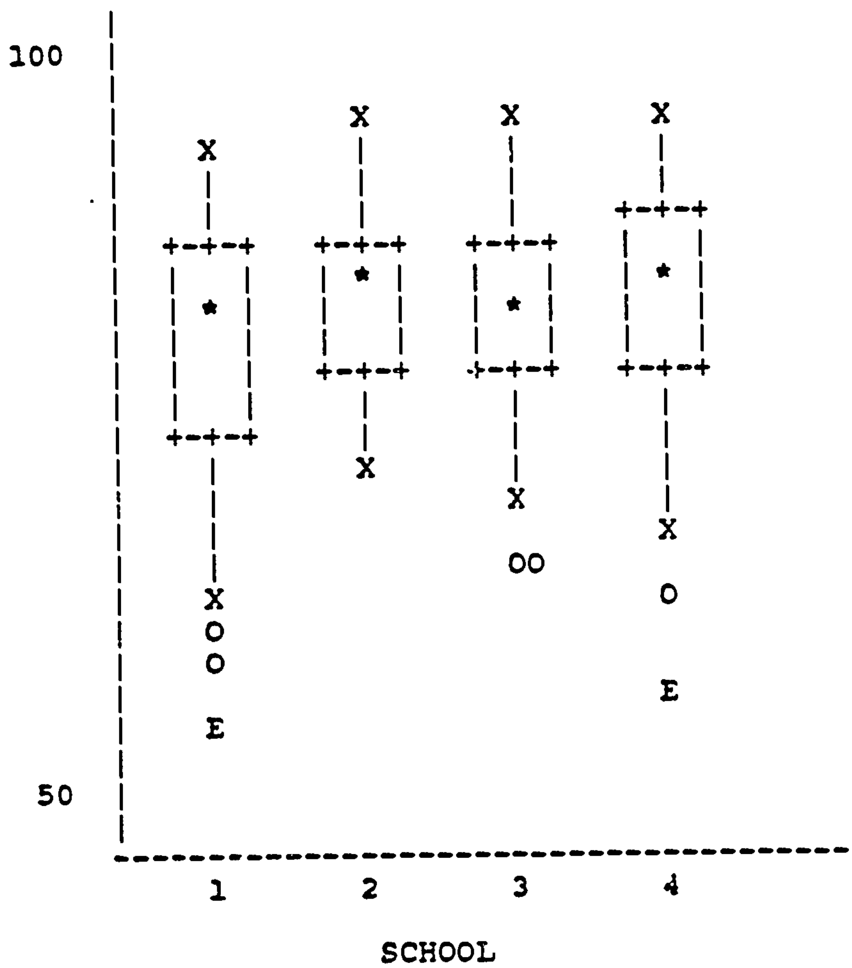

Figure 7

Distribution of Percentage of Sets Passed

```
     % OF SETS
 N    PASSED
---   ------
 2      11    *
 0      15
 0      19
 0      23
 2      27    *
 1      31    *
 3      35    **
 1      39    *
 3      43    **
 4      47    ***
10      51    *******
 9      55    ******
 9      59    ******
27      63    ******************
 6      67    ****
19      71    *************
12      75    ********
32      79    *********************
19      83    *************
47      87    *******************************
44      91    *****************************
19      95    *************
67      99    *********************************************

             I....+....I....+....I....+....I....+....I....+....I
             0         15        30        45        60        75
```

Figure 8

Percentage of Sets Passed by School