DOCUMENT RESUME

ED 397 078 TM 025 078

AUTHOR	Papa, Frank J.; Schumacker, Randall E.
TITLE	Assessing Disease Class-Specific Diagnostic Ability: A Practical Adaptive Test Approach.
PUB DATE	[95]
NOTE	17p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Artificial Intelligence; Classification; *Clinical Diagnosis; *Diseases; *Graduate Medica. Students; Monte Carlo Methods; Pilot Projects; Probability; *Robustness (Statistics); Simulation; Symptoms (Individual Disorders); Test Construction; Test Items
IDENTIFIERS	Rasch Model

ABSTRACT

Measures of the robustness of disease class-specific diagnostic concepts could play a central role in training programs designed to assure the development of diagnostic competence. In the pilot study, the authors used disease/sign-symptom conditional probability estimates, Monte Carlo procedures, and artificial intelligence (AI) tools to create test items (case vignettes) representing varying levels of typicality for the disease class known as myocardial infarction (heart attack). The typicality estimate assigned to each test item was converted to a Rasch logit scale value representing its difficulty level. Selected test items were then embedded within a paper-based examination and the performance of 628 first-year postgraduate residents-in-training determined for each item. The residents' performance was then simulated in the context of a practical adaptive testing (PAT) format. Results from residents for the actual paper-based and simulated PAT are compared and discussed. These two testing formats are also discussed in terms of their use to measure the robustness of disease-specific diagnostic concepts. An appendix explains a simulation procedure. (Contains 1 figure, 1 table, and 17 references.) (SLD)

ED 397 078

ASSESSING DISEASE CLASS-SPECIFIC DIAGNOSTIC ABILITY: A PRACTICAL ADAPTIVE TEST APPROACH

ABSTRACT Medical diagnostic performance (accuracy) appears to be both disease class-specific (performance against one disease class can not be used to predict performance against a different class), and, a function of a case presentation's 'typicality' (typical disease class case presentations are more likely to be correctly diagnosed than atypical presentations). Given this, diagnostic performance could be said to simply reflect the robustness of the subject's diagnostic concept for a given disease class. Interestingly, medical educators have demonstrated little interest in measuring the robustness of disease class-specific diagnostic concepts. The authors suggest that such measures could play a central role in training programs designed to assure the development of diagnostic competency.

In this pilot investigation, the authors utilized disease/sign-symptom conditional probability estimates, Monte Carlo procedures and artificial intelligence (AI) tools to create test items (case vignettes) representing varying levels of typicality for the disease class known as myocardial infarction (heart attack). The typicality estimate assigned to each test item was converted to a Rasch logit scale value representing it's difficulty level. Selected test items were then imbedded within a paper-based examination and the performance of PGY1 (first year postgraduate residents-in-training) determined for each item. The authors then simulated the residents' performance in the context of a practical adaptive testing (PAT) format. Results of the actual paper-based and simulated PAT residents are compared and discussed. The authors will also discuss these two testing formats in terms of their use to measure the robustness of disease class-specific diagnostic concepts.

AUTHOR INFORMATION

Frank J Papa, DO, Ph.D.
Professor, Department of Medical Education
University of North Texas Health Science Center
3500 Camp Bowie Blvd.
Fort Worth, Texas, 76107-2699
817-735-2440 Voice

Randall E. Schumacker, Ph.D.
Associate Professor
Educational Research
University of North Texas
Denton, TX 76203-6857
817-565-3962 Voice
817-565-2185 FAX
E-mail: schmckr@unt.edu

ASSESSING DISEASE CLASS-SPECIFIC DIAGNOSTIC ABILITY: A PRACTICAL ADAPTIVE TEST APPROACH

FRANK J. PAPA, D.O., Ph.D. [1]
RANDALL SCHUMACKER, Ph.D. [2]
DAVID ALDRICH, B.A. [3]
ROBERT STONE, D.O. [1]

[1] UNIVERSITY OF NORTH TEXAS HEALTH SCIENCE CENTER- FORT WORTH
[2] UNIVERSITY OF NORTH TEXAS
[3] EDUCATION RESEARCH LABORATORY INCORPORATED

Paper submitted to Mid-Western Educational Research Association

Division I
Chicago, Illinois

INTRODUCTION

Classification. Classification, the identification of the set of objects to which a given instance belongs, is one of the most fundamental and useful capabilities of the human intellect. Following several decades of research, cognitive scientists have developed two competing theories (exemplar and abstraction) attempting to describe how humans perform classification tasks.[1,2] For the most part, their investigative methodologies have involved college students, required to perform artificially contrived classification tasks based upon exposures to instances of classes consisting of predefined combinations of abstract symbols (bar charts, letters, etc.). Despite a number of limitations, these investigations have increased our understanding of how humans form the internalized 'concepts' which enable them to perform classification tasks.

Differential diagnosis as a classification task. Many similarities exist between these artificial classification tasks and medical differential diagnosis (DDX). However, the applicability and utility of classification theories in forwarding DDX research remained largely unexplored until recently. Over the last decade, a small number of investigators have utilized medically trained personnel and medical data (real patient cases, medical literature, disease/feature relationships) to determine if cognitive sciences-derived classification theories and methodologies can account for the behaviors of clinicians performing DDX.[3,4,5,6,7] By better understanding how clinicians form and utilize disease class concepts, medical educators could create more effective training and assessment methodologies.

To this end, Norman and colleagues,[8] utilize medical students, residents-in-training and medical practitioners, and, stimuli such as slides picturing actual cases of patients with dermatologic conditions, to study the effects of exemplars upon DDX performance. Through careful selection and presentation of these slides, they have successfully mounted arguments explaining how exemplars form the basis of the disease class concepts which underlie DDX performance.

In studies designed to model abstraction-based DDX performance theories, Papa and co-workers,[9,10] acquire knowledge in the form of abstractions consisting of disease/feature frequency (conditional probability) estimates from medical students, residents and board certified practitioners. These conditional probabilities are then transformed into a knowledge base sufficient to enable the investigators' artificial intelligence tools to simulate each subject's DDX performance against test cases. Their findings demonstrate that expert/novice differences in DDX (accuracy) can be accounted for solely on the basis of abstraction-derived disease class concepts.[11]

Disease class concepts and the assessment of DDX competency. There is no doubt that clinicians remember, and likely use exemplars (specific case instances) to form disease class concepts for use in DDX. It is also true that clinicians create abstractions (e.g., disease/feature conditional probability estimates) and likely use this knowledge to perform DDX. Nonetheless, it remains to be seen whether researchers can determine if exemplars or abstractions play the primary role in the formation and use of disease class concepts during DDX.

The authors suggest that enough is known about the role of disease class concepts in DDX to investigate whether DDX competency assessments can be produced via testing procedures designed to measure 'concept' robustness. Disease class concept robustness estimates would be of great benefit to both physicians-in-training and educators engaged in the process of developing and assuring DDX competencies. In this presentation, the authors develop a rationale for diligently pursuing the development of disease class DDX competency assessment procedures, and, describe their methodology and findings in a pilot study designed to assess the robustness of disease class DDX concepts.

RATIONALE

DDX competency: Case or disease class-specific. Before launching recent investigations into the applicability of classification theories in understanding how physicians perform DDX, medical education-oriented researchers focused their attention on a more fundamental question. These efforts attempted to determine whether 'skill' or 'knowledge' accounted for the development of competency in the keystone medical task known as DDX. Identification of either skills or knowledge in general, or better yet, a specific skill or knowledge base as preeminent in the development of diagnostic competency would be of critical importance for both assessment and instruction.

Up through the late 1970's, it was widely assumed that DDX competency was derived from the development of intellectual skills in general and problem solving skills in particular. However, in 1978 Elstein et al found that board certified clinicians (who were presumed to have superior problem solving skills) did not out-perform non-certified clinicians when challenged with the same battery of test cases. Rather, subject performance varied from case to case. Elstein therefore determined that DDX performance was not dependent upon any particular *skill* (including problem solving) but rather was dependent upon *knowledge* based constructs.[3] In 1988, Case et al provided evidence substantiating the knowledge based supposition by demonstrating for example, that a subject's performance against a case presentation of myocardial infarction could not be used to predict their performance against a case presentation representing pneumonia.[12]

It is critical to note that these findings would eventually be used as evidence supporting the notion that DDX performance was not only knowledge based but also 'case-specific'. Adoption of the case specificity hypothesis would have important implications for medical educators. Specifically, educators would resort to the use of a small number of case vignette test items (with only one item used to represent any given disease class) and Generalizability theory in an attempt to derive a single, *global* assessment of a subject's DDX competency.

The authors suggest however that the 'DDX performance is case-specific' interpretation is not only a conservative but perhaps an erroneously narrow interpretation of earlier findings. Rather, the authors suggest that the findings be more broadly re-interpreted as a evidence that DDX performance is 'disease class-specific' (i.e., DDX performance against one class of diseases cannot be easily used to draw inferences regarding potential DDX performance against another disease class). The authors now offer further arguments in support of a shift towards the assessment of disease class-specific DDX concepts.

First, patients presents with signs and symptoms due to the presence of an underlying pathophysiologic process. The clinician's responsibility is to correctly diagnose the disease class responsible for the signs and symptoms. Secondly, for any given disease class, the constillation of signs and symptoms with which the patient presents can be some factorial combination of signs and symptoms. Put another way, for the vast majority of diseases, there is no defining set of signs and symptoms with wihc to clinically diagnose the presence of a given disease, and, there an numerous sets of different combinations of signs and symptoms with which a given disease class can manifest itself.

If one accepts the argument that DDX performance is disease class-specific, and, given that the clinician's primary task is to diagnose which class of diseases a patient suffers from, then competency assessments should focus on the subject's abilities within any given disease class. Furthermore, given the various signs and symptoms with which a disease class can present, then competency assessments for a given disease class should not be allowed to depend upon a 'single' dichotomous correct/incorrect response to a case vignette test item. Instead, the authors suggest that assessments of DDX competency be based upon some estimate of the degree to which a subject can correctly diagnose typical through atypical case representations of a given disease class. Such estimates could therefore be said to reflect the 'robustness' of the subject's disease class-specific DDX concepts.

The need to represent the natural variation of case presentations within a given disease class. Consider the disease class known as myocardial infarction. Legitimate myocardial infarction test cases may be portrayed by any combination of the following possible (but not exhaustive) signs and symptoms; dull chest pain, pain mid-chest in location, pain duration > than 30 minutes, radiation of pain to arm, and/or neck and/or jaw, dyspnea, diaphoresis, nausea, rales and wheezes.

Given the potentially large number of distinctly different myocardial infarction case portrayals such combinations of signs and symptoms can generate, it seems neither valid nor plausible to assume that correct performance against one case vignette can serve as the basis for making any generalization as to how many other different myocardial infarction case vignettes an examinee would correctly diagnose. Simply put, the basis for making a reliable assessment of disease class-specific DDX competencies would need to be predicated upon the use of a number of different disease class-specific case vignettes. If one accepts this premise, then it would appear that medical educators would need to create an examination instrument which makes it possible to produce logistically feasible and reliable assessments of disease class-specific DDX competencies.

The authors suggest that two distinctly different testing formats could provide reliable, disease class-specific DDX competency estimates with only one capable of providing logistically feasible estimates with currently available technologies. The first format, a traditional paper-based approach could be designed to include a sufficiently large number of various case vignette presentations representing a given disease class. The number of test items needed to produce a reliable assessment might likely prove to be logistically prohibitive.

The second, a practical adaptive testing format could be used to more efficiently identify the level of case typicality at which the subject's disease class concepts fail to provide them with a correct response. Within the context of a practical adaptive testing format, the use of Rasch logit scale values to represent a case's typicality (and level of difficulty) could serve as the basis for determining the robustness of the subject's disease class DDX concept. A critical precursor to the implementation of practical adaptive testing however, is the need to demonstrate the existence of a relationship between case typicality and DDX performance, and, the ability to create case vignettes with sufficiently fine and varying levels of 'typicality'. These two issues are now addressed.

DDX performance as a function of 'case typicality': The creation and use of case typicality estimates as the basis for measuring disease class concept robustness.

Common to both exemplar and abstraction classification theories is the notion that diagnostic performance is a function of a case's 'typicality'. In a recent investigation, Papa et al[13] demonstrated that a case typicality/performance gradient applied in DDX (i.e., the more atypical the case, the less likely it would be correctly diagnosed, and, the more typical the case, the more likely it would be correctly diagnosed). These findings are the results of the authors' ability to use artificial intelligence-derived tools (named KBIT) to carefully construct case presentations with precisely varying levels of typicality.

Given the ability to create case presentations of varying levels of typicality, the authors hypothesize that the development of DDX competency might be characterized (and measurable) as the ability to correctly diagnose increasingly atypical case presentations. If this is true, then by challenging subjects with case presentations with varying (and known) levels of typicality, it might be possible to draw inferences regarding the robustness of the subject's disease class-specific DDX concept (range of typical through atypical case presentations over which the subject's disease class concept enables correct DDX). It would appear that a practical adaptive testing tool using a well calibrated Rasch item bank (representing case's ranked in terms of their typicality) could assess DDX ability directly and disease class-specific concept robustness indirectly.

Evidence of the feasibility and utility of a practical adaptive testing format for assessing disease class-specific DDX competencies could lead to important and pragmatic implications for medical instruction and curricular design. The following section provides background briefly describing the artificial intelligence-derived tools and associated methodologies used in constructing a framework for investigating a practical adaptive testing approach to assessing the robustness of a subject's disease class DDX concepts.

METHODOLOGY

Investigative tool. A team of researchers at our institution have attempted to integrate current abstraction theory-derived assumptions into the design of an artificial intelligence research tool called KBIT (Knowledge Base Inference Tool). The goal of the KBIT project is to enable investigators to simulate the DDX performance of clinicians by creating knowledge base structures and decision making processes theorized as existing in, and utilized by clinicians. Successful simulations of actual performance should enable the investigators to generate more precise inferences regarding the structure of the knowledge base and the inferencing processes underlying DDX performance.

KBIT consists of three components: 1) a knowledge base acquisition module, 2) a knowledge base transformation module, and 3) an inferencing (decision making) module.[14] With the first module, KBIT acquires knowledge from subjects in the form of conditional probability estimates for a predefined number of diseases/signs-symptoms in a given problem area. Consistent with abstraction theories, these conditional probability estimates are believed to represent 'summarized generalizations' which reflect the subject's knowledge of the frequency with which a given disease class is associated with various given signs-symptoms.

These generalizations may be stored in long term memory (derived from formalized instruction wherein an authority suggests that *'95% of patients with asthma have wheezing'*). Generalizations may also be generated on-line in working memory (i.e., based upon a clinicians personal, case-based experiences wherein *'20% of the patients he/she has seen with pneumonia have chest pain with deep breathing')*.

Based in part upon the work of Kellogg,[15] KBIT's second module contains a 'normalization' routine which transforms each subject's generalized estimates into weights representing the strength of each disease/sign-symptom relationship. (The use of conditional probability estimates occasionally gives rise to concerns regarding their 'correctness or accuracy'. The authors respond by noting that this normalization procedure is designed to diminish such concerns by shifting the basis for subsequent inferences from a dependence upon *absolute* estimate correctness to weights which reflect the subject's generalizations regarding the *'relative* nature of the strength of disease/sign-symptom relationships').

KBIT's third module contains several different inferencing mechanisms. One mechanism (prototype emulation routine) is designed to transform the normalized weights into theoretically idealized disease class 'prototypes' which are used for research measurement purposes to determine the degree to which a given test case is both similar to an idealized disease class prototype (lies within the class - also referred to as 'pattern match'[10]), and different from each competing, idealized disease class prototype (distance between classes - also referred to as pattern discrimination[10]) in the problem area.

These within-class (WC) and between-class (BC) measures can be used by the authors to achieve three distinct research objectives. First in terms of simulating a subject's DDX performance, they have been used to confer a diagnosis upon a given test case. Second, they have also been used to estimate the degree to which any given combination of case vignette signs/symptoms represents a typical through atypical disease class case presentation.[13] Third, the authors now suggest that they can be used in the context of a practical adaptive test to generate inferences regarding the robustness of a subject's disease class concepts (as will be described in this presentation). (For a more detailed review of how KBIT creates and utilizes idealized prototypes, within-class and between-class measures in arriving at a diagnosis see Papa et al.[14]).

Study design: Overview. This investigation consisted of four separate phases.

The first phase involved the use of conditional probability estimates derived from a panel of subjects, to generate via the use of a Monte Carlo procedure, a large number of potential test cases representing various presentations of the disease class known as myocardial infarction (MI). The second phase consisted of the use of KBIT's within-class and between-class measures to generate estimates of each potential test case's typicality, the calculation of a Rasch logit scale value for each test item, and finally, the selection of eight MI test case items for use in this study. The third phase consisted of the administration of the selected test items to PGY1 residents and the compilation of individual and group performance against the items. The fourth phase consisted of the construction of a practical adaptive test (PAT) which simulated the PGY1 residents performance against the eight selected test items.

Research hypotheses. Hypothesis # 1: There is a positive correlation between typicality/logit case values and PGY1 group performance on a paper-based examination containing the eight MI test items (i.e., for a given case vignette test item, the greater it's typicality/logit value, the greater the percentage of PGY1 subjects who will correctly diagnose the case). Hypothesis # 2: The simulated PAT will closely mirror the actual performance of PGY1 subjects.

Phase One: Generation of test case pool. All board certified emergency medicine physicians who were members of the Texas chapter of the American College of Emergency Physicians (117 members) were requested via a questionnaire to produce conditional probability estimates for 67 signs-symptoms (previously published[16]) as they related to nine common or important diseases known to cause acute chest pain (myocardial infarction, angina/coronary ischemia, dissecting thoracic aortic aneurysm, pericarditis, upper gastrointestinal disorders, pneumonia, pneumothorax, musculoskeletal disorders and pulmonary embolus). Each subject therefore was asked to submit a total of 603 conditional probability estimates. Following three mailings, thirty-four members returned their questionnaire.

For each subject, the Monte Carlo procedure then used their estimates to create 100 different test cases representing sign/symptom variations in the disease class known as myocardial infarction. The rule underlying this procedure was as follows: if a subject determined that 60% of patients with MI had sign-symptom '5', then 60% of the MI cases generated were assigned a positive sign-symptom '5'. Given 34 subjects, a total of 3,400 potential MI test cases were constructed.

Phase Two: Creation of case typicality estimates and case selection. KBIT subsequently utilized the 603 conditional probability estimates (produced by each subject) to simulate each subject's DDX performance against all 3,400 MI test cases. For each simulated subject, their diagnosis, measure of within-class (WC) and measure of between-class (BC) estimates were recorded for each test case. Test cases with less than 40% of the KBIT simulated subjects correctly diagnosed were dropped from further consideration.

Of those remaining cases in the test pool, each test case's 'typicality' (T) was defined as sum of the logit of each case's WC and BC measure (T = Logit WC + logit BC). An averaged T estimate for each case was then determined from all simulated subjects correctly diagnosing the case. It was assumed that the averaged T estimates for all items could be arranged along a unidimensional linear continuum. A plot revealing the range and distribution of cases rank-ordered in terms of their averaged T was then produced. This plot represented the MI disease class 'case typicality gradient' (i.e., level of test item difficulty).

Eight test cases were selected to represent points along the typicality/logit gradient. Typicality/logits for the eight cases (from easy to hard) were as follows: #1, -.735; #2, -.432; #3, -.035; #4, .000; #5, +.075; #6, +.160; #7, +.280; #8, +.327. The positive case findings associated with the selected test cases were subsequently transformed into brief case vignettes for which nine diagnoses served as possible answers (See Table 1 for listing of each case's positive signs and symptoms). Only one diagnosis counted as a correct answer.

Phase Three: Examination procedures and correlation of typicality/logit values with PGY1 performance. The selected test items were distributed among the part three licensure examination administered by the National Board of Osteopathic Medical Examiners in February, 1994 to 628 postgraduate year one (PGY1) residents-in-training. The percent of correct subject responses to each MI test item were as follows: #1, 91%; #2, 57%; #3, 59%; #4, 26%; #5, 58%; #6, 12%; #7, 22%; #8, 6%. PGY1 performance was subsequently correlated with the typicality/logit values.

Phase Four: Simulated practical adaptive testing (PAT). A schema utilizing a test item starting and stopping procedure similar to one advocated by Wright[17] was employed to simulate the PGY1 subjects' performance in the PAT. More specifically, the simulated PAT began with MI test item number four. The PAT simulation would next move to item six if the subject's actual performance on the paper-based examination demonstrated that he/she correctly had diagnosed item four. If the subject correctly diagnosed item number six, the PAT stepped up to item number eight. If item number eight was correctly diagnosed then the PAT stopped and the student was ranked as if eight items were correctly diagnosed. If item number eight was not correctly diagnosed, then the subject was ranked as if seven items were correctly diagnosed. This ranking procedure was mirrored in reverse if the subject incorrectly diagnosed item number 4 and so on. (For Wright's PAT algorithm for administering a set of items with a Rasch model see appendix.)

Analysis. Hypothesis # 1. Pearson correlation coefficient relating the typicality/logit value for each of the eight test items and the percentage of PGY1 subjects correctly diagnosing each of the eight test items was performed. The correlation was 0.83, $p < .01$, df=7. Hypothesis # 2. The degree to which the PAT mirrored the actual performance rankings of PGY1 subjects can be seen in Figure # 1.

DISCUSSION

A positive and statistically significant correlation exists between typicality/logit values produced via KBIT-derived pattern match (WC) and pattern discrimination (BC) estimates, and, the performance of PGY1 residents-in-training.

It is important to recall that KBIT was designed to function in a manner consistent with an abstraction-based theory of how clinicians perform DDX. That is, similar to KBIT, clinicians may actually employ a knowledge base comprised of disease class concepts, which in turn is comprised of generalizations or abstractions derived form knowledge of disease/sign-symptom conditional probabilities. Clinicians may also employ an inferencing (decision making) mechanism similar to KBIT's pattern matching (WC) and pattern discrimination (BC) in an effort to perform DDX.

Given as sound theoretical basis with which to assert that similar knowledge base structures and inferencing mechanisms may be operative in practicing clinicians, then the authors suggest that the use of typicality/logit values to draw inferences regarding the robustness of a subject's disease class DDX concepts seems plausible.

The degree to which the simulated PAT mirrored the ranking of actual subject performance on the paper-based examination enables the authors to draw the following inferences. First, PAT appears to have a definite logistical advantage in terms of testing time. Specifically, the CAT arrived at it's ranking of a given subject utilizing three test items while the paper-based format required all eight test items to arrive a final ranking. Second, Like the paper-based testing format, a PAT containing disease class-specific, typicality/logit value based test items may provide educators with an opportunity to feasibly assess the robustness of a subject's disease class DDX concepts.

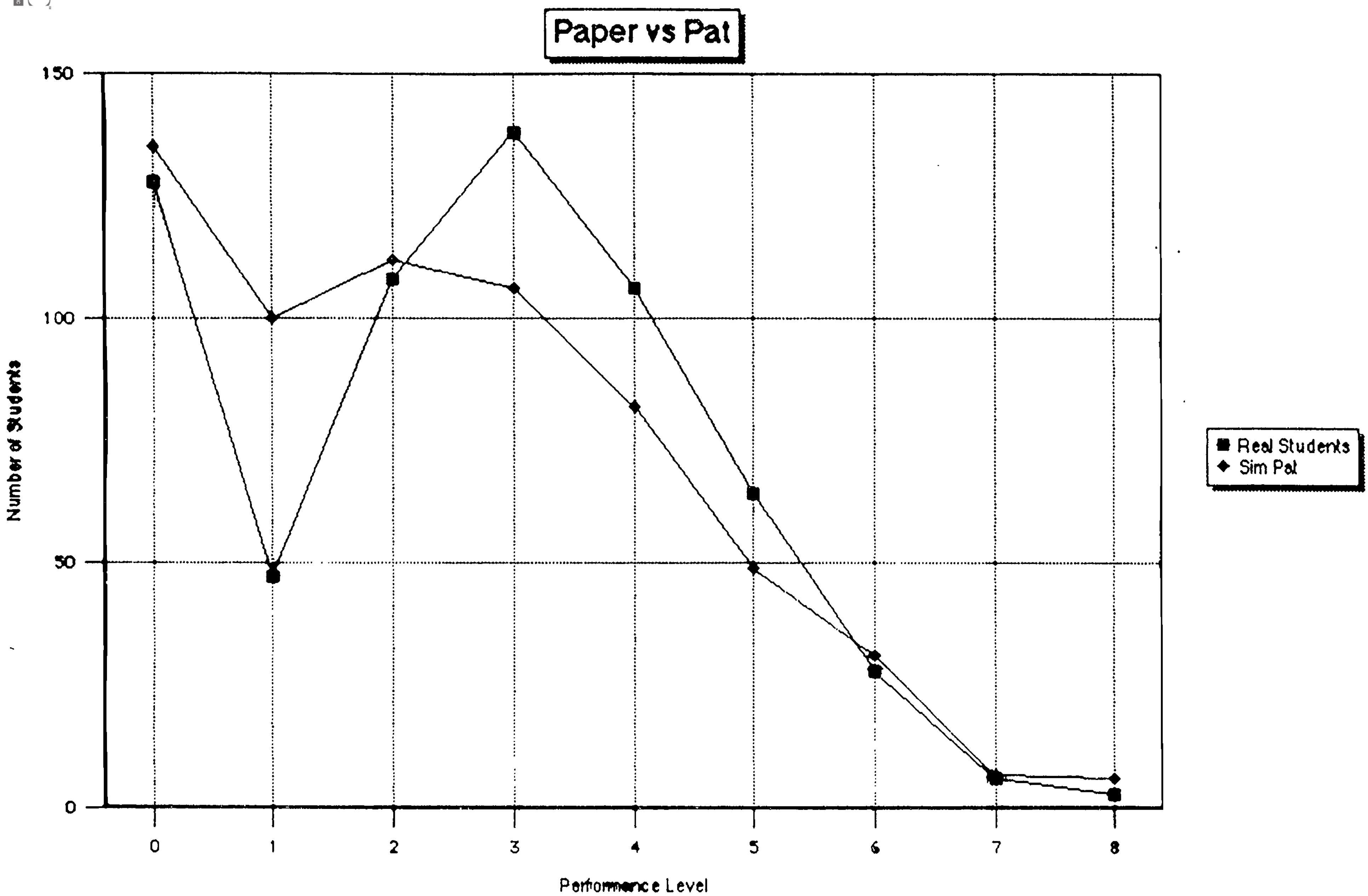
CONCLUSION

Classification theories have long demonstrated that performance is a function of an instance's typicality. The utilization of abstraction-derived classification theories and artificial intelligence tools to construct typicality/logit value based medical test case vignettes enabled the authors to successfully anticipate the DDX performance of a group of PGY1 subjects. The authors suggest that these theories, tools, and findings give reason to believe that it may now be possible to draw inferences regarding the robustness of the disease class concepts which clinician's utilize to perform DDX.

PAT formats utilizing test case items selected because of their typicality/logit values appear to be a useful vehicle for deriving logistically feasible inferences regarding the robustness of a clinician's disease class DDX concepts. Such inferences may prove to be the basis for the development of efficient and effective medical instructional and curricular reforms. Clearly, artificial intelligence tools, PAT formats and the formalized application of classification theories may prove to be the foundation for new and more meaningful assessment procedures in medical education.

TABLE 1. Positive signs and symptoms associated with each of eight selected MI cases.

CASE NUMBER SIGNS/SYMPTOMS	#1	#2	#3	#4	#5	#6	#7	#8
AGE GREATER THAN FORTY	•	•	•		•			
MALE	•	•	•	•	•	•	•	•
SUDDEN PAIN ONSET	•	•	•			•		
DURATION 1 - 3 HOURS	•	•	•	•		•	•	•
PAIN PRESENT NOW	•	•	•	•	•	•	•	•
DULL/PRESSURE/SQUEEZING	•	•	•	•	•	•		•
BURNING								•
SUBSTERNAL/LEFT PRECORDIAL	•	•		•	•		•	•
POSTERIOR CHEST								•
RADIATION TO NECK/JAW/ARM	•				•		•	
RADIATION TO BACK							•	
SUDDEN DYSPNEA	•	•		•	•	•	•	
SYNCOPE	•		•					
PALPITATIONS		•	•				•	•
PAIN MODIFIED BY MEDICATION	•	•	•	•	•			
PAIN MODIFIED BY REST	•		•					•
PAIN MODIFIED BY EXERTION	•	•			•		•	
RECENT IMMOBILITY			•					
RECENT PRIOR SURGERY					•		•	
WHEEZES								•
RHONCHI							•	
TACHYPNEA	•					•		
TACHYCARDIA							•	
S4 GALLOPS	•				•			
PRECORDIAL FRICTION RUB					•			
NECK VEIN DISTENTION						•		
MILD HYPERTENSION	•			•				
DIAPHORESIS	•	•	•	•	•			
COOL/PALE/MOIST SKIN	•	•	•	•	•		•	
CYANOSIS	•						•	



Paper vs Pat
150
100
50
0
Number of Students
0
1
2
3
4
5
6
7
8
Performance Level
Real Students
Sim Pat

APPENDIX

0. Request next candidate.
   Set D=O, L=O, H=O, and R=O.
1. Find next item near difficulty (D).
2. Set D at the actual calibration of that item.
3. Administer that item.
4. Obtain a response.
5. Score that response.
6. Count the items taken. L = L + 1
7. Add the difficulties used. H = H + D
   If response not correct,
8. Reset item difficulty. D = D - 2/L
   If response is correct,
9. Count right answers R = R + 1
10. Reset item difficulty. D = D + 2/L
    If not ready to decide to pass/fail,
11. Go to step 1.
    If ready to decide pass/fail
12. Calculate wrong answers. W = L - R
13. Estimate measure. B = H/L + log(R/W)
14. Estimate Error. S = sqrt[L/(R*W)]
15. Compare measure B with pass/fail standard T.
16. If (T - S) < B < (T + S), go to step 1.
17. If (B - S) > T, then pass.
18. If (B + S) < T, then fail.
19. If all candidates administered test- stop, else
20. Go to step 0.

BIBLIOGRAPHY

1 Medin, D.L. & Smith E.E. (1981). Strategies and Classification Learning. Journal of Experimental Psychology: Human Learning and Memory. 1, 241-253.

2 Medin, D.L. & Schaffer, M.M. (1978). Context Theory of Classification Learning. Psychology Review. 85, 207-238.

3 Elstein, A., Schulman, L. and Sprafka, S. (1978). Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press.

4 Allen, S.W., Brooks, L. R. and Norman, G. R. (1988). Effect of Prior Examples on Rule-based Diagnostic Performance. In Research in Medical Education. Proceedings of the Twenty-Fifth Annual Conference. Washington, D.C.: Association of American Medical Colleges.

5 Bordage, G. and Zacks, R. (1984). The Structure of Medical Knowledge in the Memories of Medical Students and General Practitioners: Categories and Prototypes. Medical Education, 18, 406-416.

6 Patel, V.L., Evans, D.A. and Groen, G.J. (1988). Biomedical Knowledge and Clinical Reasoning. In D.A. Evans & V.L. Patel (eds.), Cognitive Sciences in Medicine, Cambridge, MA, MIT Press.

7 Schmidt H.G. Norman, G.R. and Boshuizen, H.P.A. (1990). A Cognitive Perspective on Medical Expertise: Theory and Implications. Academic Medicine, 65, 611-621.

8 Norman, G.R., Rosenthal, D., Brooks, L.R., Allen, S.W. and Muzzin, L.J. (1989). The Development of Expertise in Dermatology. Archives of Dermatology, 125, 1063-1068.

9 Papa, F.J. & Meyer, S. (1987). An Expert Program Shell Designed for Extracting "Disease Prototypes" and Their Use as Models for Exploring the "Strong Problem Solving Methods" Employed in Clinical Reasoning. In Hart, IR (ed.), Further Developments in Assessing Clinical Competence (pp. 354-364). Quebec, Canada, Heal Publication

10 Papa, F.J., Shores, J.H. and Meyer, S. (1990). Effects of Pattern Matching, Pattern Discrimination, and Experience in the Development of Diagnostic Expertise. Academic Medicine, 65, s21-s22.

[11] Papa, F.J., Shores, J.H., Meyer, S. (1992). The Use of a Pattern Recognition-based, Prototype-Driven Research Tool to Study the Cognitive Constructs in Medical Decision Making. In Hart IR, Harden RM & Des Marchais J, (eds) Current Developments in Assessing Clinical Competence (Pp 210-213). Ottawa, Can-Heal Publications.

[12] Case SM, Swanson DB Stillman PL. Evaluating Diagnostic Pattern Recognition: the Psychometric Characteristics of a New Item Format. In *Research in Medical Education: 1988.* Proceedings of the Twenty-Seventh Annual Conference, p 3-8. Chicago, IL: Association of American Medical Colleges.

[13] Papa FJ, Elieson B. (1993). Diagnostic Accuracy as a Function of Case Prototyicality. Academic Medicine, 68, s58-s60.

[14] Papa, F.J., Shores, J.H., Meyer, S., O'Reilly, R. and Bourdage, R. The Role of Pattern Matching and Pattern Differentiation in Clinical Problem Solving. In Teaching and Assessing Clinical Competence, Bender W, Hiemstra RJ, Scherpbier, AJJA and Zwiestra RP eds., pp. 322-331. Groningen, The Netherlands, BoekWerk Publications, 1990.

[15] Kellogg, R.T. Simple Feature Frequency Versus Feature Validity Models of Formation of Prototypes. Perceptual and Motor Skills. 51(1980) 295-306.

[16] Papa, F.J. and Meyer S. A Computer-assisted Learning Tool Designed to Improve Problem Solving Skills. Annals of Emergency Medicine. 18(1989) 273-296.

[17] Wright BJ. (1988). Practical Adaptive Test. Rasch Measurement Transactions, 2(2), 7.