The Development of Expertise in Dermatology

Geoffrey R. Norman, PhD; Donald Rosenthal, MD;
Lee R. Brooks, PhD; Scott W. Allen, MS; Linda J. Muzzin, PhD

• To examine the development of expertise in dermatology, accuracy of diagnosis and response times of subjects at five levels of expertise were assessed. A total of 100 slides, 2 typical and 3 atypical slides from each of 20 common skin disorders, were presented to six subjects at each of the following levels: second-year preclinical medical students, final year medical students, residents in family medicine, general practitioners, and dermatologists. Accuracy of diagnosis rose from 21% for medical students to 87% for dermatologists. Correct diagnosis was associated with a decrease in response time with expertise, whereas errors were associated with a dramatic increase in response time, and was slower than correct response times at all levels, suggesting that errors do not result predominantly from carelessness or speed. Typical slides accounted for a constant proportion of diagnostic errors at all higher levels of expertise, and experts continued to make a significant proportion of errors on slides shown to be relatively easy for residents. The results are shown to be at variance with any model that equates expertise with the mastery of complex rules, but they are consistent with models of expertise that propose that expertise is equated with a rapid "pattern-recognition" process, and errors result from unintended confusion with previous similar examples.

(*Arch Dermatol.* 1989;125:1063-1068)

The practice of dermatology requires mastery of a unique body of knowledge and skills. Although systemic disease may manifest itself cutaneously, most of the common skin disorders have little systemic involvement and, therefore, are not dealt with in detail by the traditional disciplines of medicine. Furthermore, the diagnostic process of the dermatologist is unique; dermatologists rely more extensively on visual perception—the ability to recognize and classify lesions dependent on their visual appearance—than other physicians. In this aspect, dermatology may have more in common with laboratory disciplines such as histology or radiology than other disciplines of clinical medicine.

A few recent studies have examined the accuracy of dermatologists and primary care physicians in recognizing common skin disorders[1,2] or in distinguishing melanoma or precancerous lesions from benign lesions.[3] The conclusions of all these studies are similar; dermatologists were consistently more accurate than the primary care physicians at diagnosing skin diseases. In one study,[1] dermatologists had an overall accuracy of 98% vs 60% for primary care physicians; in the second study,[2] the comparable rates were 96% and 54%, respectively.

Nevertheless, it is unclear from any of the reported studies why dermatologists are superior at recognizing skin disorders nor what might be the essential components of an effective educational intervention. It may seem self-evident that their additional experience with skin disorders is sufficient explanation for the superiority of the specialists. Still, this leaves unanswered the question of how this additional experience is integrated into a superior problem-solving process.

This study was designed to examine the evolution of expertise in dermatology. The basic task for the clinicians we studied was to diagnose common skin disorders from color slides chosen to be typical of the disorder and atypical or able to be confused with other disorders. In addition, response time to diagnosis was recorded to provide another index of difficul-

Accepted for publication March 1, 1989.

From McMaster University, Hamilton, Canada.

Reprint requests to Department of Clinical Epidemiology and Biostatistics, Health Sciences Centre, McMaster University, 1200 Main St W, Hamilton, Ontario, Canada L8N 3Z5 (Dr Norman).

ty experienced in the diagnostic task. The study included a variety of expertise, ranging from medical students to dermatologists, to examine the acquisition of skill as students progress through undergraduate and postgraduate education.

A MODEL OF EXPERTISE

An intuitively plausible position about the acquisition of expertise is what we will call the *Independent Cues* interpretation[4]: learners gain expertise mainly by acquiring knowledge about the specific features (signs or symptoms) that characterize a disease or condition and those features that are best able to differentiate among diseases. This model of learning is the implicit, if not explicit, goal of most instruction in clinical diagnosis. For example, students are often given lists of signs that typify a disorder, assuming that any patient having a large portion of such cues is likely to be an example of the disorder. With practice, learners are supposed to acquire increasingly appropriate strategies for applying these rules to specific cases. This model also underlies statistical methods, such as bayesian or regression decision models, based on the weighting of features.

One implication of this model is that the difficulty of a particular case should be related to the degree to which it embodies features that are predictive of a single diagnosis and does not contain features that might suggest more than one diagnosis. To the extent that expertise consists of learning the combinations of features that distinguish among diagnostic categories, a direct consequence of an independent cues model is that performance should improve rapidly on typical cases, since they possess most of the distinguishing features of a category. Conversely, atypical cases have a minimum of features characteristic of a category and a greater degree of ambiguity; hence, they would require a high level of expertise to differentiate from other conditions. It is of particular interest, therefore, to examine performance on typical and atypical slides as a function of expertise.

SUBJECTS AND EXPERIMENTAL METHODS

Six subjects were chosen at each of five levels of expertise: second-year preclinical medical students, final-year medical students, five-year residents in family medicine, practicing family physicians, and practicing dermatologists. Medical students were selected from the undergraduate program at McMaster University, Hamilton, Canada. In common with many other programs in North America, there is limited formal instruction in dermatology at McMaster University; however, students are free to seek out electives in dermatology. Residents in family medicine were volunteers from the residency program at McMaster University. Family physicians were practicing in the Hamilton area; one of the six subjects was an academic physician and the remainder were in private practice. All of the dermatologists were in private practice in Hamilton or Toronto, Canada. All subjects were initially approached by mail and were followed up by telephone. Students were seen in the researcher's office at McMaster University; the practicing physicians were visited at their offices by the research assistant.

Table 1.—Accuracy of Subjects at Each of Five Levels for Each of 20 Dermatoses

	Level of Expertise*				
Dermatosis	1	2	3	4	5
Herpes zoster	17	10	63	67	87
Herpes simplex	30	30	57	73	100
Acne	60	53	60	80	93
Psoriasis	13	40	67	80	97
Dyshidrosis	0	7	23	63	83
Contact dermatitis	27	33	57	43	77
Atopic dermatitis	13	20	47	73	83
Impetigo	20	13	50	77	87
Basal cell carcinoma	20	37	83	60	93
Nevus	50	67	70	80	87
Warts	37	53	37	53	67
Lichen planus	3	3	27	30	70
Seborrheic dermatitis	3	27	73	73	90
Alopecia areata	50	20	70	73	90
Seborrheic keratosis	7	43	73	80	100
Actinic keratosis	7	17	57	30	73
Pityriasis rosea	0	7	57	73	87
Urticaria	33	27	50	77	77
Tinea corporis	10	33	63	60	70
Tinea versicolor	0	30	40	73	90
Average	21	31	55	66	87

*Mean of six subjects at each level. Levels are as follows: 1, second-year medical student; 2, final-year clinical clerk; 3, second-year resident in family medicine; 4, general practitioner; and 5, dermatologist.

The stimulus materials were 100 slides chosen from the slide library of an academic dermatologist (D.R.). Five slides were chosen from each of 20 common skin conditions (Table 1) with two judged by the dermatologist to be typical presentations and three judged to be atypical presentations. A brief history consisting of one to four lines of typed text and intended to be typical of the disorder was then created by the dermatologist for each slide.

The slides were placed in random order, alternating with black slides, in a carousel projector. To balance for order effects, four different starting positions were used. To investigate the effect of the brief history, for half the slides the subject first read the history, then viewed the slide. The presence or absence of history for each slide was balanced across subjects.

Each subject was asked to identify each lesion as rapidly as possible or to inform the experimenter that he could not arrive at a diagnosis. At this point, the subject pressed a large stop button that advanced the projector to a black slide and stopped the timer. Times were recorded to the nearest 10th of a second and diagnoses and other comments were recorded on audiotape. No feedback was provided.

Data analysis focused on accuracy of identification, categorized in three levels—correct, incorrect, or "don't know"—and on response times. Log-linear analysis was used on the proportion correct to examine the effect of

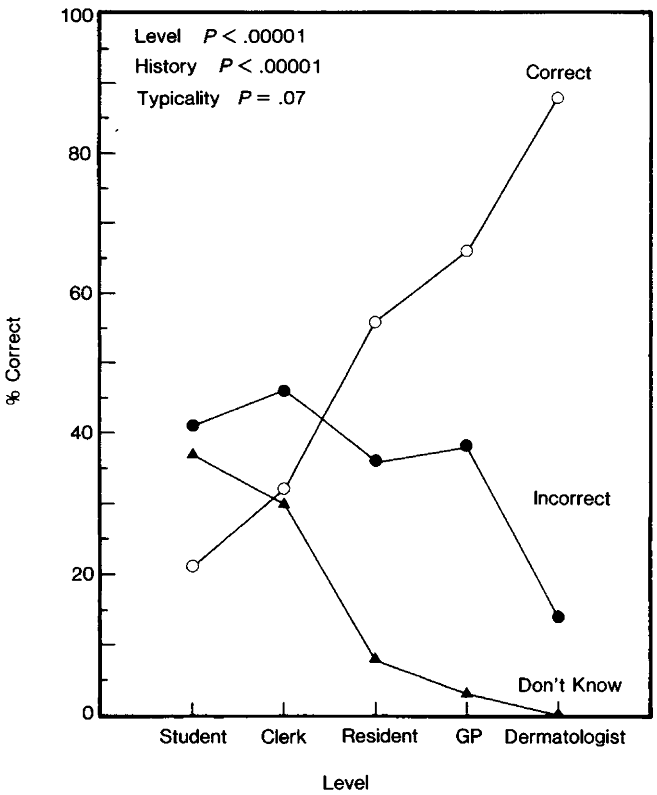

Fig 1.—Mean percent of slides with correct diagnosis, incorrect diagnosis, and "don't know" response by level of expertise. GP indicates general practitioner.

educational level, typicality, and presence or absence of history. Response times were analyzed using repeated-measures analysis of variance to examine the effect of expertise, correctness, typicality, and presence or absence of history. Secondary analysis explored in more detail the relative error rates by residents, general practitioners, and dermatologists on typical-atypical and easy-difficult slides.

RESULTS

As shown in Fig 1, a strong and approximately linear relationship existed between correct diagnosis rates and expertise, ranging from a low of 21% correct for second-year students to a high of 87% for experts. Also as expected, the proportion of slides labeled don't know decreased approximately linearly from 38% for students to near 0% for experts. However, the distribution of errors did not show a linear trend with expertise. The highest error rates occurred at intermediate levels of expertise, reaching a high of 52% errors for final-year students. The presence of a written history increased accuracy at all levels ($\chi^2 = 74.4$; $P < .001$), amounting to an average increase in accuracy of about 11%.

The accuracy of subjects in the five groups for each condition is shown in Table 1. Although a clear trend with increasing expertise is evident, there are some interesting anomalies. For normal variants, eg, nevus, and very common disorders (acne), the second-year students performed at about the level of residents. Conversely, some conditions (pityriasis rosea, dyshidrosis, lichen planus) showed nearly 0% accuracy prior to residency education. The accuracy of the dermatologists in this study is somewhat lower than that reported in the two previous studies[1,2] (86% in this study vs 98% and 96%, respectively), and the accuracy of the primary care physicians is moderately higher (66% in this study vs. 54% and 60%, respectively). The discrepancies may reflect the particular choice of stimulus slides, differences in the samples, or the particular features of the present experimental conditions.

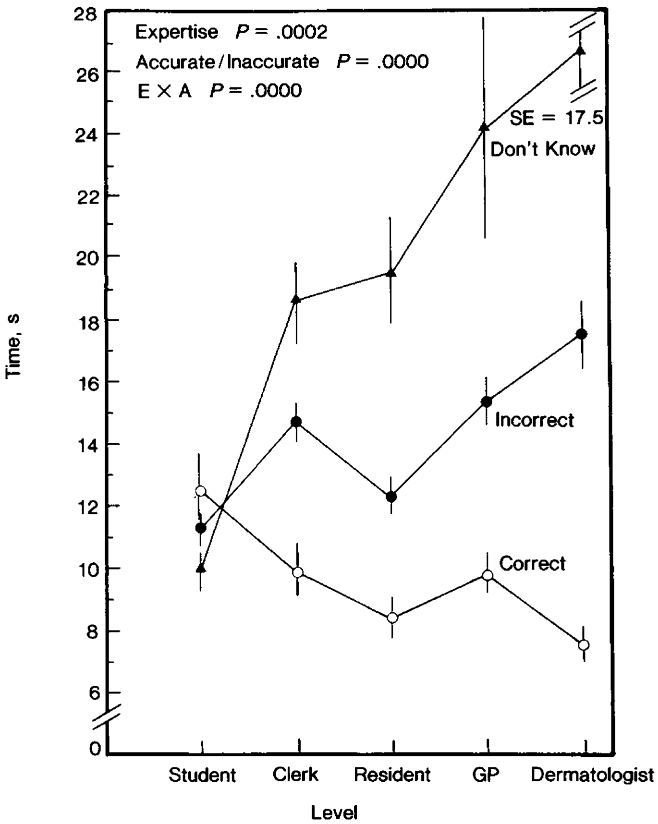

Fig 2.—Mean response time (with SE) per slide, in seconds, by level of expertise. GP indicates general practitioner. E X A indicates expertise times accuracy interaction term.

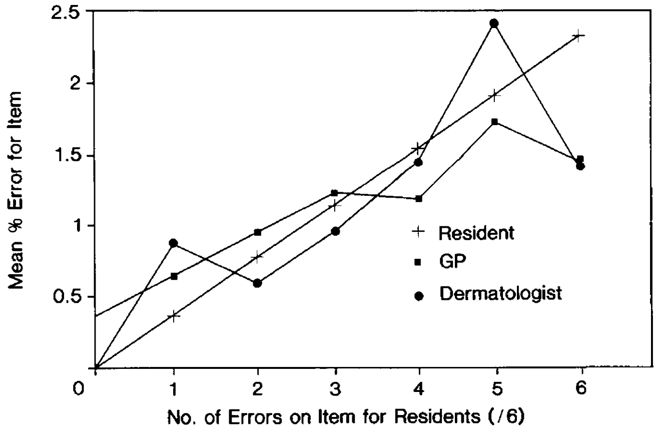

Fig 3.—Mean error rate per slide (expressed as a proportion of all errors) for residents, general practitioners (GPs), and dermatologists related to difficulty for residents (0/6 to 6/6).

Certainly, one possible source of error is carelessness or lack of attention to detail resulting from the demand for speed as well as accuracy imposed by the experimental task. If this were the case, we should find that errors are associated with short response times. Mean response times by expertise for correct, incorrect, and don't know responses are shown in Fig 2. There was an evident decrease in response time with expertise for correctly identified slides, ranging from a mean of 12.0 seconds for students to 7.5 seconds for the dermatologists. Conversely, there was a striking increase in response times for incorrectly labeled slides related to expertise, ranging from 11.6 seconds for students to 18.3 seconds for dermatologists and an even stronger positive relationship between response time and expertise was observed for the don't know alternative, ranging from a mean time of 9.6 seconds for students to 24 seconds for general practitioners and 26 seconds for dermatologists. Moreover, errors always took longer than correct diagnoses, and don't knows took longer still. The interaction among correct/incorrect/don't know and expertise was significant (F = 11.36; $P < .0001$) as were the main effect of educational level (F = 9.47; $P < .0001$) and the main effect of correct/incorrect/don't know (F = 51.2; $P < .0001$). No other main effects or interactions were significant with the exception of a small interaction between educational level and presence or absence of written history (F = 3.69; $P = .05$), suggesting that the presence of a history resulted in slightly lower response times at lower levels of expertise.

To explore the predictions of the independent cues model, the errors of three groups—residents, general practitioners, and dermatologists—on typical and atypical slides were examined in more detail. This secondary analysis was restricted to these three groups because, although error rates ranged from 45% for residents to 13% for dermatologists, the nature of the errors committed by the three groups was similar, ie, they were using similar diagnostic categories. By contrast, students tended to use categories like "tumor" or "infection."

Errors were examined to determine whether these occurred in typical or atypical slides. As we indicated, an independent cues model predicts that errors are more likely in atypical slides, since these possess relatively fewer of the essential features of the diagnostic category. Second, this should interact with expertise: relatively fewer errors should be committed by experts on typical slides, but there may remain a residual set of atypical slides where additional information, eg, from history or biopsy specimen, is required to rule out competing alternatives.

The analysis supported the first prediction. The ratio of typical errors to all errors by level of expertise is as follows: resident, 0.40; general practitioner, 0.42; and dermatologist, 0.40. There were proportionately about 50% more errors committed on atypical slides by subjects at all levels of expertise. However, the second prediction was not supported. The ratio of typical to all errors was constant for all three levels of expertise, despite the fact that the total number of errors had declined by about a factor of three. Thus, subjects at all three levels were continuing to make significant numbers of errors on slides that were considered typical.

Table 2.—Prediction of Diagnostic Errors by Three Dermatologists by Level of Expertise and Typical (Typ)/Atypical (Atyp) Slides

Dermatologist No.	Resident		General Practitioner		Dermatologist	
	Typ	Atyp	Typ	Atyp	Typ	Atyp
1						
First nomination	0.11	0.24	0.08	0.26	0.17	0.38
All nominations	0.28	0.47	0.38	0.51	0.30	0.60
2						
First nomination	0.16	0.24	0.10	0.34	0.29	0.27
All nominations	0.19	0.33	0.11	0.43	0.29	0.52
3						
First nomination	0.15	0.24	0.17	0.29	0.20	0.31
All nominations	0.18	0.27	0.19	0.37	0.24	0.42

One explanation is that errors on typical and atypical slides derive from different sources. Errors on typical slides may result from carelessness or inattention, whereas atypical errors may be related to inherent ambiguities in the slide. If this were the case, one might expect that response times for errors on typical slides would be short and response times for atypical errors would be relatively long. Analysis of response times demonstrated that errors on typical slides were not faster than those on atypical errors; the average difference in response times between typical and atypical errors was about 2 seconds for three groups and was not statistically significant. These data again suggest that errors are not a result of carelessness or inattention to detail (which would be associated with short response times) nor are they dictated primarily by the typicality of the slide, which would result in a predominance of errors on the atypical slides at higher levels of expertise.

A similar argument can be applied to performance on relatively easy and difficult slides. An independent cues model would predict that performance should improve fastest on relatively easy slides, so that slides on which residents make proportionately fewer errors should present no difficulty to experts. Thus, most improvement with expertise should arise on slides that are easiest for residents. Conversely, some ambiguous slides are likely to contain insufficient information for accurate diagnosis, and these should show little improvement with expertise.

An alternative position is that errors of clinicians are a result of carelessness or inattention. If this were the case, there should be no association between the difficulty of an item, based on the performance of residents, and errors committed by dermatologists, since errors result from a random process unrelated to any measure of item difficulty.

To explore these predictions, we characterized the difficulty of each slide on the basis of the errors by residents, thus an easy slide had zero of six errors by residents, a difficult slide had six of six errors by residents, and there were five intermediate levels of difficulty. We then examined the proportion of errors at each level of expertise committed on slides at each level of difficulty; ie, we compared the probability of error as a function of item difficulty for residents at three levels of expertise.

Because the difficulty of slides is based on the performance of residents, the plot of resident errors at each level of difficulty is a straight line through the origin. From the independent cues model, we would anticipate that as expertise is acquired, proportionately more errors will be committed on difficult slides, so that the curves move to the right with expertise. Conversely, if errors were a result of random processes such as inattention, the likelihood of an error by a general practitioner or dermatologist should be unrelated to the resident item difficulty, and the line should be flat.

The results are shown in Fig 3. It is apparent that the distributions for general practitioners and dermatologists are similar to those of residents, ie, an approximately straight line with positive slope. More important, although we would predict from an independent cues model that with increasing expertise the curve would shift to the right, the data provide no evidence of this shift. In addition, the curves for general practitioners and dermatologists have about the same slope as the residents, which would not be expected if errors resulted at random. Thus, although the absolute error rate declined by about a factor of three from resident to dermatologist, there is no evidence that expertise resulted in relatively greater improvement on easy items.

One other form of converging evidence is derived from the ability of experts to predict the diagnostic errors of others. An independent cues model would suggest that, since disorders are predicted from the association between specific features and diagnostic categories, errors result from the absence of critical features or the presence of ambiguous features, thus the nature of the diagnostic errors should be predictable. We could refine this hypothesis to suggest that errors of students may indeed be less predictable, since their knowledge base may be unstable and idiosyncratic, but errors of other experts should be highly predictable, since they are operating on a common knowledge base that underlies their shared expertise.

To test this assumption, three expert dermatologists (D.R. and two other dermatologists) were shown each slide successively in a single session and they were asked to predict the likely errors. All were asked to indicate the most plausible error and all were encouraged to provide multiple alternatives. There were an average of 2.5 alternatives per item suggested by the first dermatologist, 1.8 by the second, and 1.7 by the third. Of the 100 slides on which predictions were made, only 10 resulted in total agreement on the most likely alternative, and there was no agreement at all among the experts on 49 of 100 slides.

Despite the multiple alternatives offered by each dermatologist, none were able to predict actual errors with any degree of accuracy (Table 2). Considering only the first alternative suggested, this matched the errors only 8% to 38% of the time. Even considering all alternatives mentioned, the experts were only able to predict from 11% to 60% of the errors. Dermatologists were best able to predict the errors of other dermatologists and least able to predict resident errors. Turning the analysis around, of the alternatives predicted by the two experts, 22% to 66% were never mentioned by residents, 13% to 52% were never mentioned by general practitioners, and 21% to 64% were never mentioned by any dermatologist. Thus, the task of predicting likely errors made by others appears far more difficult than might be anticipated if errors result from features within the slide.

COMMENT

The results of the study provide some confirmation of previous studies but extend our understanding of the basis of expertise in dermatology. We have demonstrated a positive relationship between accuracy in diagnostic judgment and experience that extends approximately linearly from medical students to specialists. However, the greatest gain in accuracy occurred during residency training and practice, highlighting perhaps the lack of dermatologic training in the undergraduate years. It was surprising that error rates did not show a monotonic relationship with training, but rather were curvilinear, with maximum error rates at intermediate levels of training, suggesting that confidence in diagnosis rises more rapidly than accuracy. The data on response times tend to discount any explanation of errors resulting from inattention. In particular, when the expert commits errors, these are associated with significantly longer, rather than shorter, response times.

The analysis of error rates on typical-atypical, easy-difficult lesions demonstrated that, although acquisition of expertise is associated with a dramatic reduction of error rates, a constant proportion of errors continues to be committed by experts on slides judged typical by their peers and demonstrated to be empirically easy. Furthermore, many of these errors were not predicted by other experts examining the slides.

These findings suggest that diagnostic errors are not predictable on the basis of stable characteristics or features of the lesion. In particular, the extreme rapidity with which experts make correct diagnostic judgments and conversely the slow response times associated with errors suggest that different processes may be operative. Correct judgment may result from an automatic "pattern-recognition" process in which the lesion is considered as a whole, without regard for individual features. Conversely, the slow judgment times for errors may represent an analytical, feature-by-feature analysis that represents a failure of the pattern recognition process. If the pattern recognition process associated with rapid and correct responses is independent of individual features, then there is no reason to assume a strong association between typicality based on the presence of characteristic features and improvable diagnostic skill. As a result, one would anticipate that experts would continue to make errors on typical slides. Some support for this view arises in the radiological literature. Swenssen and colleagues[5,6] and Blesser and Ozonoff[7] have advanced the view that the reasoning process in diagnostic radiology is a two-step process with the first stage based on a rapid, largely unconscious pattern recognition process and the second stage based on a conscious analytical process; however, this view remains controversial.[8,9]

An alternative perspective is that errors result from processing variability, ie, the same stimulus materials may be handled differently on successive occasions. One potential source of this variability may be the influence of prior similar-appearing lesions. In this view, the judgment of diagnostic category is based, to large degree, on similarity to prior examples rather than on an analysis of the features of the presenting lesion consistent with each diagnostic class. Since this similarity match is not based on a feature-by-feature analysis, what makes a previous instance available is not just a fixed set of features but contextual information relating to how a prior instance was processed and in what setting. As in many other areas of memory research, previous items in the same context are more available than items processed in a different context.[10] Similarly, items treated in a similar manner are more available than those processed differently.[11]

How could this kind of variability in the availability of prior instances provide an explanation of the observed data? If access to prior instances is context dependent, this could affect overall performance without necessarily changing the relative difficulty of typical-atypical, easy-hard items. For example, the previous occurrence of particular diagnoses in a series may result in the increased availability of that category, hence a diagnostic bias. This bias could be manifested in either the initial consideration of an incorrect hypothesis or by a decreased probability of catching the error once the hypothesis is being evaluated. A second possibility is that certain contextual factors, such as the location of the lesion or the physical appearance of the patient, might create a bias in favor of a particular diagnosis and result in errors that are unrelated to objective categorizations such as difficulty or typicality. The contribution of such error factors could be expected to decline with increasing expertise.

Since the errors of an individual physician are then dependent, at least in part, on those prior instances that are similar, both in salient and contextual features, this theory explains why experts are unable to predict the errors committed by others. Support for this perspective can be found in psychological research on concept formation that has demonstrated the considerable influence of a single prior exposure to a similar-appearing example.[12] This hypothesis is being pursued through current research at this laboratory.

Whatever the appropriate explanation, the results of the present studies are at variance with the accepted view that expertise is related to mastery of the "rules" of medicine; the results are consistent with the notion that specific prior experience in the form of prior examples has a central role in the diagnostic task. The diagnostician may often be in the position of concluding that a particular presentation is an example of a particular diagnosis simply because it bears a strong resemblance to a presentation by a previous patient with the diagnosis. To the extent that such mechanisms are operating, it will be important to investigate further the influence, both positive and negative, of prior examples on the diagnostic task.

This investigation was supported by the Natural Sciences and Engineering Research Council of Canada, Ottawa.

References

1. Ramsay DL, Fox AB. The ability of primary care physicians to recognize the common dermatoses. *Arch Dermatol.* 1981; 117:620-622.

2. Clark RA, Rietschel RL. The cost of initiating appropriate therapy for skin diseases: a comparison of dermatologists and family physicians. *J Am Acad Dermatol.* 1983;9:787-796.

3. Cassileth BR, Clark WH, Lusk EJ, Frederick BE, Thompson C, Walsh WP. How well do physicians recognize melanoma and other problem lesions? *J Am Acad Dermatol.* 1986;14:555-560.

4. Smith E, Medin DL. *Categories and Concepts.* Cambridge, Mass: Harvard University Press; 1981.

5. Swenssen RG, Hessel SJ, Herman PG. The value of searching films without preconceptions. *Invest Radiol.* 1985;20:100-114.

6. Swenssen RG, Hessel SJ, Herman PG. Radiographic interpretation with and without search: visual search aids the recognition of chest pathology. *Invest Radiol.* 1982;17:145-151.

7. Blesser B, Ozonoff D. A model for the radiologic process. *Radiology.* 1972;103:515-521.

8. Swets JA. Editorial comments and response. *Invest Radiol.* 1985;20:108-109.

9. Kundel H. Editorial comments and response. *Invest Radiol.* 1985;20:110-111.

10. Godden DR, Baddeley AD. Context dependent memory in two natural environments: on land and underwater. *Br J Psychol.* 1975;66:325-332.

11. Cermak LS, Craik FIM. *Levels of Processing in Human Memory.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc Publishers; 1979.

12. Brooks LR. Decentralized control of categorization: the role of prior processing episodes. In: Neisser U, ed. *Concepts and Conceptual Development.* New York, NY: Cambridge University Press; 1987:141-175.