IMPROVEMENT IN MEDICAL EXPERTISE INDEPENDENT OF STABLE KNOWLEDGE

G.R. NORMAN, L.R. BROOKS, S.W. ALLEN, D. ROSENTHAL
McMASTER UNIVERSITY, HAMILTON, ONTARIO

An intuitively plausible position about the acquisition of expertise is what we will call the Independent Cues interpretation: learners gain expertise mainly by acquiring knowledge about the specific features (signs or symptoms) which characterize a disease or condition and those features which are best able to differentiate among diseases. (The term independent cues follows Smith & Medin, 1981 in their classification of concept theories). This model of learning is the implicit, if not explicit, goal of most instruction in clinical diagnosis. This assumption also underlies models, such as Bayesian or regression decision models that capture increasing expertise with changes in weights of features.

One consequence of an independent cues model is that performance should improve more rapidly on typical than atypical cases. Since typical cases possess more of the features which are characteristic of a category, these should be mastered with relative ease. Conversely, atypical cases have few features in common with a category, hence would require a high level of expertise to differentiate from other conditions. In addition, cases which are empirically relatively easy for neophytes should be mastered with perfect accuracy by experts.

In this paper we will concentrate on the relative performance on typical and atypical, easy and difficult cases of clinicians at three different levels of expertise in dermatology. The task we studied was the diagnosis of common skin disorders on the evidence provided by color slides, some of which were judged typical of the represented disorder and others were atypical.

EXPERIMENT 1

Method

Six subjects were chosen at each of three levels of expertise in dermatology: first year residents in family medicine, general practitioners, and practicing dermatologists.

The stimulus materials were 100 slides chosen from the slide collection of an academic dermatologist. Five slides were chosen from each of 20 common skin conditions, with two judged by the dermatologist to be typical presentations and three atypical presentations. A brief history, consisting of 1 to 4 lines of typed text and intended to be typical of the disorder was created by the dermatologist for each slide. Since this variable is irrelevant to the focus of the current paper, all error analyses were collapsed across history. The slides were presented in a randomized series, using four different starting positions to

balance for order. For half the items, balanced across subjects, the subject first read the history and then viewed the slide. Subjects were asked to diagnose each case as rapidly as possible or to indicate "don't know."

Results

Mean Errors. The average error rate for the groups were Residents = 44%, General Practitioners = 33%, Dermatologists = 14.5%, resulting in a highly significant effect of expertise on mean errors (chi-squared = 1337, p=.0001). The use of the "Don't Know" option was minimal, ranging from 7% for residents to 0.3% for dermatologists.

Response time

Response times were separately calculated for correct, incorrect, and 'don't know ' slides. The mean response time for correct identifications declined slightly with expertise, from 9.0 sec. for residents to 7.7 sec. for dermatologists. By contrast, errors for all groups were significantly slower than corrects, and increased significantly with expertise (residents 12.2 sec.; general practitioners 15.3 sec.; and dermatologists 17.5 sec.; $F=10.9$, $p<.001$). The positive association with expertise was even more pronounced for the 'don't know' slides; residents took an average of 19.3 sec., general practitioners 24.4 sec., and dermatologists 26.3 seconds ($F=3.31$, $p<.05$). These results suggest that errors do not apparently result from carelessness or speed and lack of thoroughness; if anything, the converse appears to be true.

Typicality.

To assess the effect of typicality, the number of errors was first corrected for frequency (there were 2 typical and 3 atypical cases per disorder) and then formed into the ratio of errors on typical items divided by total errors. An equal tendency to make errors on typical and atypical items would result in a .5 value for this ratio, and learning to deal effectively with typical items before atypical items would result in declining values with expertise. In fact, the proportion of errors made on items designated as typical by the dermatologist were approximately constant over the three levels of expertise, despite the threefold decrease in overall errors (R=.40, GP=.42, D=.40).

Average Item Difficulty.

An independent cues model implies that i) the difficulty of a case is related to the degree to which the cues present in the item support a single diagnosis, and ii) expertise is related to the knowledge of the appropriate combinations and weightings of

these cues. Thus if we consider items which are empirically easy or difficult for residents, most improvement with expertise should arise on easier items. 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 models, we characterized the difficulty of each slide on the basis of the errors committed by residents, thus an easy slide had 0/6 errors by residents, a difficult slide had 6/6 errors by residents, and there were 5 intermediate levels of difficulty. We then examined the proportion of errors at each level of expertise committed on slides at each level of difficulty.

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 distributions move to the right with expertise. Conversely, if errors were a result of random processes such as inattention, the likelihood of an error by G.P. or dermatologist should be unrelated to the resident item difficulty, and the distribution should be flat.

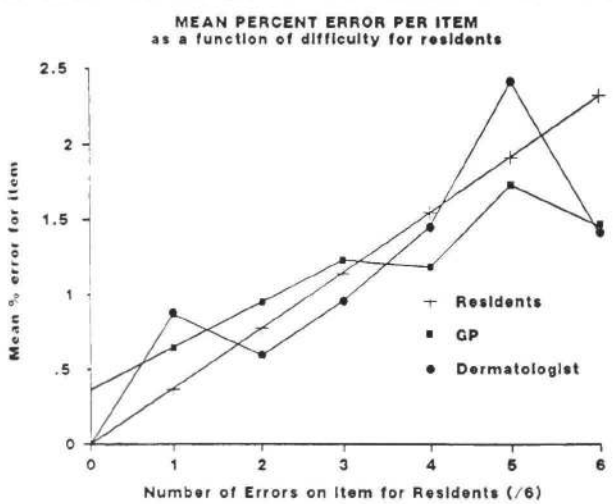

FIGURE 1
ERROR RATE OF RESIDENTS, G.P.'S, AND DERMATOLOGISTS RELATED TO ITEM DIFFICULTY

The results are shown in Figure 1. It is apparent that the distributions for general practitioners and dermatologists are similar to those of residents, i.e. 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.

Thus, although the absolute error rate declined by about a factor of three from resident to dermatologist, we found no evidence that expertise resulted in improvement on relatively easy or typical items.

EXPERIMENT 2

The analysis of typicality from the first study is subject to possible idiosyncracies in the ratings of typicality since the categorization was done by only one dermatologist and consequently might be subject to some unreliability. Also, perceived typicality itself might vary across expertise groups as well as having an uncertain relation to the basis of diagnostic performance. Therefore, it is critical to determine if the results of this study held up when the typicality ratings of another group were substituted.

The second study addressed some of these issues. In the context of another study examining the effect of prior exposure on ratings of plausibility (the hindsight effect), we obtained ratings of typicality on a total of 69 of the slides used in the present study from a varying number of general practitioners ranging from 3 to 6. Following the initial presentation, a second session was arranged with each subject a minimum of 4 weeks later. At this followup session, subjects were shown a total of 32 slides, of which 16 had been used in the first session and 16 were new (counterbalanced across subjects), and were asked to diagnose the conditions. The second session permitted an examination of the effect of a single prior exposure on diagnostic performance.

To conduct this analysis, the proportion of the general practitioners rating each slide as typical was determined. Five levels of categorization were created: typical by total agreement (22 slides), typical by majority agreement (25 slides), equally divided (3 slides), atypical by majority agreement (13 slides), and atypical by total agreement (6 slides). Although it is evident that G.P.'s were more inclined to rate slides as typical than the original classification of the dermatologist would indicate. nevertheless there was reasonable agreement between the two sources. Of the 23 slides originally rated as typical by the dermatologist, 21 (91%) were rated typical by a majority of G.P.'s. However, only 16 of the 44 slides (36%)initially classed as atypical were so rated by a majority of G.P.'s.

An analysis was then conducted as before, calculating the proportion of errors made on typical and atypical slides, considering both slides on which there was total agreement and slides where a majority, or all, G.P.'s agreed. The results are shown in Table 2 below:

Table 2
Proportional Error Rate by Expertise

	Resident	General Practitioner	Dermatologist
Total Agree	.339	.345	.343
Majority Agree	.338	.395	.339

It is evident from this table that the constant proportionality of errors on typical slides as a function of expertise is also evident with the G.P. ratings, thus it does not appear to be a result of the idiosyncratic categorization of the dermatologist.

As one final converging evidence on the topic, we examined the performance of G.P.'s in diagnosing slides which they themselves had previously rated as typical or atypical. Average error rate on self-rated typical slides was 29%, and on atypical slides was 54%, for a ratio of .34, consistent with the ratios shown in the previous table.

Thus, it is apparent that although performance of all groups was better on slides judged as typical by a variety of approaches, there was no association between improvement in performance related to expertise and typicality. Put another way, slides judged as typical presented just as much diagnostic difficulty (proportionately) to dermatologists as to residents.

Prior Examples and Diagnosis

It is apparent that the independent cues model and a model which views errors as a random event fare poorly in accounting for the improvement in expertise observed in the present study. An alternative model of concept formation (Brooks, 1987) postulates a central role of prior instances in recognition.

The second study provides a partial test of this model. In addition to examining relative error rates in the second session on those slides rated by each G.P. as typical or atypical, we also examined the error rate on slides which had been seen previously in the context of the typicality-rating task and a balanced set of new slides. The results are as shown below:

Table 3

Error Rates on Old and New Slides

	Previously Seen	New
Typical	31%	46%
Atypical	48%	60%

Thus, a single prior exposure to the slide, a minimum of four weeks previously, resulted in a 20-30% reduction of error rates. These data suggest that the diagnostic task may be strongly influenced by recall of prior instances of a category.

DISCUSSION

In these data the traditional indices of category structure - typicality and average item difficulty, are roughly constant over a large range of accuracy. We conclude that the improvement over the range of expertise observed in this study is not a matter of learning items in order of difficulty or learning more appropriate weights for the essential symptoms and signs. In other words, these data are incompatible with an independent cues interpretation of acquisition of expertise. In fact, the observed constant proportionality rules out any model that determines typicality, average item difficulty, and improvable error by the same information.

This finding also provides difficulty for stable instance-based models of categorization (e.g. Hintzman 1986, Brooks, 1978), which would hypothesize constant availability of prior instances of a category, since in this view the expert has available a relatively stable array of prior instances of a category, and diagnosis is conducted by a comparison of similarity to available instances. The difficulty of this model is that there should be many more typical instances available to the expert for similarity judgement, thus expert performance should be proportionately higher on typical and easy slides.

By contrast, although the data suggest a role of prior instances in expertise, 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. Previous items in encountered in the same context are more available then items processed in a different context (Godden and Baddeley, 1975). Similarly,items treated in a similar manner are more available than those processed differently (Cermak and Craik, 1979).

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 and atypical, easy and 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. A second possibility is that certain contextual factors, such as the location of the lesion or the physical appearance of the patient might result in bias in favour of a particular diagnosis and result in errors which are unrelated to objective categorizations such as difficulty or typicality. The contribution of such error factors could be expected to decline with increasing expertise.

Obviously, we are not claiming that there is no such thing as "independent cues" knowledge of the features of particular diseases, or that such knowledge is irrelevant to expertise. Rather, these data constitute a case in which there is massive improvement that is independent of such knowledge. It is entirely possible that the acquisition of the basic definitional, "independent cues" type of knowledge is restricted to the initial stages of learning, and the substantial change while obtaining practical experience is due to the retrieval factors just discussed. We further conjecture that these findings might be a common feature of any field in which the challenge is to recognize any of a large number of disorders occurring in a mixed series. Perhaps in a mixed series the effect of difficulty of an item is small by comparison to the effect of processing variations induced by the series itself. If so, then the emphasis in training in such areas should be on providing practice with mixed series and with emphasizing to the learners the importance of mixed practice for developing expertise.

REFERENCES

Brooks L.R. Non analytic concept formation and memory for instances. In E.Rosch and B.LLoyd (ed), Cognition and Categorization. Hillsdale N.J., Lawrence S. Erlbaum, 1978.

Brooks L.R. Decentralized control of categorization: The role of prior processing episodes. In U. Neisser (ed), Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization, New York, Cambridge University Press, 1978.

Cermak L.S. and Craik F.I.M. Levels of Processing in Human Memory. Hillside N.J., Lawrence S. Erlbaum, 1979.

Godden D.R. and Baddeley A.D. Context dependent memory in two natural environments on land and underwater. Brit. J. Psychol. 66, 325-332, 1975.

Smith E., Medin D.L. Categories and Concepts.Cambridge MA, Harvard University Press, 1981.