Role of Specific Similarity in a Medical Diagnostic Task

Lee R. Brooks
McMaster University
Hamilton, Ontario, Canada

Geoffrey R. Norman
Clinical Epidemiology and Biostatistics
McMaster University
Hamilton, Ontario, Canada

Scott W. Allen
University of Lethbridge
Lethbridge, Alberta, Canada

Three experiments are reported showing that diagnosis of skin disorders by medical residents and general practitioners was facilitated by similar cases previously seen in the same context. Diagnosis of similar cases was facilitated more than that of dissimilar cases in the same diagnostic category, demonstrating that facilitation was not solely due to activation of the diagnostic category as a whole. Because diagnosis was posed in a multiple-choice format that always included the correct diagnosis, the relative disadvantage of dissimilar items was not due to the unavailability of the category name. The similarity effect also occurred with 2-week delay between the initial case and the test cases. Variations in diagnostic procedure, ranging from giving a quick first impression to arguing for given alternative diagnoses before selection, did not interact with the effect of similarity. This result suggests that the similarity effect is not strongly dependent on a particular diagnostic strategy.

Medical diagnosis is primarily a categorization task. The goal of the clinician is to determine the likely diagnostic class to which the patient's illness belongs to guide treatment. An implicit assumption of much medical education is that skill in diagnosis is acquired mainly by learning about the separate features (signs or symptoms) that characterize and differentiate various diseases and conditions. Textbooks of clinical diagnosis describe explicitly the features of disorders, and students are taught to list the causes of particular complaints and the signs and symptoms of various diseases. With practice, learners are supposed to acquire increasingly appropriate weightings for combining and applying these features to specific cases. Following Smith and Medin (1981), we will refer to this class of approaches as independent cues models.

Independent cues assumptions underlie models, such as Bayesian (e.g., Fischhoff & Beyth-Marom, 1983) or regression decision models, that capture increasing expertise with changes in weights or conditional probabilities applied to individual features; these are most often the same cues that are listed in the canonical rules used in instruction. These models have been extensively applied to research on medical diagnostic tasks (see Slovic, Rorer, & Hoffman, 1971, for an early example of a multiple-regression—policy-capturing approach to medical diagnosis, and Wigton, 1988, for a recent review; see Schwartz, Gorry, Kassirer, & Essig, 1973, for an early argument for a Bayesian, decision analysis approach to medical treatment management, and Weinstein & Fineberg, 1980, for a more extensive treatment; finally, see Spiegelhalter & Knill-Jones, 1984, for a recent review of medical discussion of the statistical approaches and comparison with knowledge-based, expert systems approaches. A recent comment on research approaches to medical diagnosis and reasoning is in Elstein, Shulman, & Sprafka, 1990). Independent cues assumptions also underlie most prototype models that use a set of characteristic features to describe the graded structure of a concept (see Bordage & Zacks, 1984, for an application of a prototype approach to medicine).

The role of examples in a pure independent cues model of learning is limited to providing the material from which more abstract knowledge is developed. Similarly, when examples are used in instruction, the implicit assumption seems to be that the practice examples merely provide an opportunity to develop weights or rules for combination, with little lasting effect of the specific instances used. However, Medin, Altom, Edelson, and Frecko (1982) in a simulated medical diagnostic task have demonstrated that learners can be sensitive to correlations between features that are not a part of independent cues representations. They argue that the demonstrated sensitivity of their subjects to co-occurrence of features is well fit by a model that assumes comparison of test items with representations of previously experienced exemplars. One additional function of training examples, then, could be to provide representations that would store co-occurrence information that is not directly given in the independent cues form of information. In addition, Medin and Edelson (1988) have suggested that retrieval of individual cases plays a role in judgments of base rates.

There have been several proposals that people's performance on verification, identification, and categorization tasks is based on a hybrid of memory for prior cases and knowledge
of general regularities (e.g., Allen & Brooks, 1991; Brooks, 1987, 1990; Jacoby & Brooks, 1984; Malt, 1989; Medin, Dewey, & Murphy, 1983; Medin & Ross, 1989; Reber & Allen, 1978; Ross, 1987, 1989; Whittlesea & Brooks, 1988). However, it is still possible that for categorical tasks, generalization around old items is an unusual basis of generalization, occurring only when a sufficient analytic basis is not available. This “analytic default” position would seem to be that taken by Homa, Sterling, and Treppel (1981) for dot patterns when they suggested that instances are used for generalization only early in a training sequence, before the learner has had an opportunity to abstract a prototype. Similarly, Reber (1989) has suggested for artificial grammars that instances would likely be used only when there was special stress on memorizing the details of items, a circumstance he thought hinders the tacit abstraction of a grammar describing the whole set of items. If this position applies to medical domains, then reliance on generalization around known cases would be limited to early phases of training or to cases that do not clearly fit with general knowledge.

However, there are several lines of recent evidence that suggest that generalization around similar instances occurs despite having explicit, analytic knowledge or dealing with generally familiar categories. In a paradigm closely related to the current research, Allen and Brooks (1991) demonstrated a continued effect of training exemplars even when the subjects were given a simple and sufficient rule for classification of artificial animals. As in the Medin et al. (1982) result, subjects were sensitive to similarity between training and test instances that could not be accounted for by typicality or weightings of individual cues. Because subjects had been given explicit training in the application of a perfectly predictive rule, the importance of similarity-based responding could not have been restricted to a period of learning before the development of a more concise and reliable basis for responding. Furthermore, the information about old items appears to come without requiring deliberate effort, as shown by a similarity effect occurring even when it reduced accuracy and the subjects were encouraged to be accurate. Also important for present purposes is the fact that this result was dependent on distinctive perceptual representation: The evidence for similarity-based responding did not occur when the items were presented as written lists of features.

A series of experiments using line drawings (Jacoby, Baker, & Brooks, 1989) showed that similarity to prior exemplars experienced in the same context was important to identifying common natural objects. These drawings were seen through gradually clearing visual noise or through the gradual revelation of features produced by watching an object appear while walking around a corner. In Jacoby, Baker, and Brooks, it was shown that the degree of generalization to new objects was dependent on whether the previous similar objects had been seen in the clear or through noise, which was interpreted as varying the degree of attention to visual details that was required. When the training items had been seen through noise, later performance on those same items was improved, and performance on similar items was decreased relative to performance following training on items seen in the clear. That is, generalization depended on what was done with the prior exemplar, rather than just its decontextualized similarity to the target.

There are areas of medicine, such as dermatology, radiology, and histopathology, which also place a heavy reliance on visual arrays. The presenting stimuli are characterized by the existence of strong perceptual configurations among the features in addition to a more conceptual level of featural co-occurrence. In such areas, rules about separate characteristic features are useful, as evidenced by their prominence in instruction, but are regarded by no one as being sufficient. The identity of features is often ambiguous, and the co-occurrence and configurational interaction of features are useful forms of information. As a consequence, we could easily expect effects of similarity to prior examples in these medical domains. Given that physicians both have explicit instruction and gain considerable familiarity with these disorders, we could investigate whether reliance on similarity to prior cases is useful for knowledgeable medical personnel, or is limited to early stages of expertise or to individuals who do not have the best available analytic knowledge. The purpose of this article is to investigate this possibility.

These studies are directly aimed at investigating an effect of similarity to prior instances on diagnosis in a highly visual area of medicine. The guiding premise in developing these experiments is that some representation of individual cases is being laid down during training and is influential during transfer. However, these experiments will not provide evidence against a modified independent-features model. If we show an effect of prior instances on diagnoses made by experienced physicians or on diagnoses made after a delay, then we will have shown effects that have not routinely been taken as implications of an independent-features approach. However, if such a model were modified such that recent instances result in disproportionately large changes to prior weights, in contrast to being treated as an increase of one in the total known sample, then both of these results could be accounted for. The articles just referenced that were arguing evidence against a modified independent-features model. If we show an effect of prior instances on diagnoses made by experienced physicians or on diagnoses made after a delay, then we will have shown effects that have not routinely been taken as implications of an independent-features approach. However, if such a model were modified such that recent instances result in disproportionately large changes to prior weights, in contrast to being treated as an increase of one in the total known sample, then both of these results could be accounted for. The articles just referenced that were arguing evidence against a modified independent-features model. If we show an effect of prior instances on diagnoses made by experienced physicians or on diagnoses made after a delay, then we will have shown effects that have not routinely been taken as implications of an independent-features approach. However, if such a model were modified such that recent instances result in disproportionately large changes to prior weights, in contrast to being treated as an increase of one in the total known sample, then both of these results could be accounted for. The articles just referenced that were arguing more strongly for effects of specific prior episodes also provided evidence of the processing and contextual specificity that is normally associated with memory for prior episodes. This type of evidence poses more of a challenge for independent-cues models and would accordingly require greater modification of them. In principle, of course, a feature model could explain any result in episodic memory, but would do so with the addition of assumptions that were not associated with the independent-cues models previously applied in medicine.

The specific-domain chosen for study was dermatology. Although history information (e.g., pain, itchiness, duration, prior occurrences, and alleviating factors) is often important, diagnosis is primarily based on the appearance of the lesion, although appearances are rarely the sole determinant of treatment. Textbooks of dermatologic diagnosis describe the various categories in terms of predominantly additive rules: lists of features that are characteristic, but not necessarily definitional, of a particular disorder. Individual color photographs are commonly used in textbooks, publications, and student examinations, although rarely if ever as the sole basis for diagnosis. Dermatologists routinely gather slide collections of
their more interesting patients for teaching purposes and for continuing education with colleagues; these collections are a useful resource for the present purposes. Dermatologic problems are relatively common, in one study accounting for some 6% of visits to family doctors (McFarlane, Norman, & Spitzer, 1971) and in another about 7% (Ramsay & Fox, 1981); thus, family physicians can be viewed as relatively experienced in dealing with common classes of skin diseases.

The diagnostic categories used in the experiment were deliberately chosen to avoid rare conditions; all conditions were among the 20 most common disorders seen in family practice (Ramsay & Fox, 1981). The slides were chosen to allow us to compare test items that are similar to previous cases with cases that are dissimilar but in the same diagnostic category. This allows us to determine whether any facilitation from prior cases is solely due to activation of the diagnostic category as a whole or to specific similarity. The diagnostic question was posed in a multiple-choice format to see whether the major effect of similarity is to improve the chance that a given diagnostic possibility would come to mind. Because the correct diagnosis was always one of the alternatives provided, any effect of similarity must be a result of something other than just increasing the availability of the category name. To determine whether any similarity effect is short-lived, we introduce, in Experiment 1, a 2-week delay between exposure to the initial cases and presentation of the test cases. The subjects in Experiment 1 were family practice residents. In Experiment 2, we repeat the same design with general practitioners who have had an average of 15 years experience to see whether similarity to recent cases in context are influential for diagnosticians with considerable experience. Finally, in Experiment 4, we investigate whether an effect of similarity is strongly dependent on the diagnostic procedure adopted. It is possible that if the diagnostician was "really careful" or systematically considered alternatives by arguing for them before giving a diagnosis, then an effect of similarity would be much smaller.

Experiment 1

The major purpose of Experiment 1 is to determine whether similarity to items previously seen in the current context influences subsequent diagnoses in subjects who have acquired expertise in medicine. In the initial training, first-year family medicine residents are asked to rate the typicality of a series of photographs of dermatological lesions. The purpose of this exercise is to have the subjects attend to correctly diagnosed cases. The test phase followed either immediately or after a 2-week delay. The critical comparison is between items that are visually similar to items seen in the study phase and dissimilar items that are in the same diagnostic category.

Method

Subjects. The subjects were 16 volunteer first-year family medicine residents from McMaster University who were paid for their participation.

Materials. The materials were color photographic slides of dermatological lesions chosen from the slide libraries of two practicing and teaching dermatologists. For the purposes of counterbalancing, the slides were chosen in quartets. Each quartet consisted of two pairs of slides, chosen to produce high within-pairs similarity and low between-pairs similarity. This allowed for the manipulation of similarity independent of category membership, that is, same diagnosis similar items (SD/sim) and same diagnosis different items (SD/diff). The quartet design used was originally used for the same purpose with line drawings. Criterion for similarity in the current material was a relatively unsophisticated judgment of overall appearance made by the experimenters (further discussion on this judgment will be given later), matching as far as possible on appearance and distribution of the lesions; area of body affected; angle of photograph; and age, sex, and race of the patient.

Sixteen diagnoses were chosen for inclusion in the study, all among the 20 most common disorders occurring in North American practice. Four diagnoses contained 4 quartets each, four with 3 quartets, four with 2 quartets, and four with 1 quartet, for a total of 4 Diagnoses × (4 + 3 + 2 + 1) Quartets × 4 Slides (each, or 160 slides). For balancing purposes, these diagnostic categories were combined into four groups so that each group (A, B, C, and D) contained one 4-quartet diagnosis, one 3-quartet diagnosis, one 2-quartet diagnosis, and one 1-quartet diagnosis.

The 30 items used in the Study Phase consisted of 10 items from each of three of the counterbalancing sets. The set of items used in the Test Phase, consisted of 60 items: 10 old items, identical to those seen in the typicality rating phase; 10 same diagnosis similar (SD/sim) items, the other member of the similar pair for those seen in the Study Phase; 10 same diagnosis different (SD/diff) items, items chosen from the dissimilar pair in the quartet of the items used in the Study Phase; and 10 new category items, items from the diagnostic categories in the fourth counterbalancing set of items, diagnoses that were not seen in the Study Phase. In addition, there was a second set of 10 SD/sim and 10 SD/diff items (Test Phase extras) that were balanced only across these two categories. This second set was included to increase the power on the critical comparison between the SD/sim and SD/diff item and still stay within reasonable time limits. Each group of four categories was rotated through the four conditions (old, SD/sim, SD/diff, and new) in a Latin-square design, resulting in a counterbalanced set of materials. One effect of this balancing is that no training item served as both an old test item and as the analog to a SD-sim item for the same subject.

Three diagnoses, consisting of the correct diagnosis and two plausible alternatives, were presented with each slide used in the test phase. In many cases, the alternatives were incorrect diagnoses given by dermatologists or general practitioners in a previous study. Additional diagnoses for slides for which we did not have empirical alternatives were generated by our informant dermatologist, Donald Rosenthal. He also checked the overall plausibility of the alternative sets for the entire list of slides.

Procedure. In the Study Phase, the subjects were told they would see 30 slides of dermatological lesions, each correctly labeled with the diagnosis. They were asked to rate, on a percentage scale, how typical each slide was of the disorder it represented and to provide a couple of sentences of justification. The purpose of this phase was primarily to have the subjects attend to correctly diagnosed cases. The slides were presented using an individual slide viewer with a viewing area of approximately 70 × 70 mm. The label for each slide was presented on the screen of an Epson PX-8 portable computer into which the subjects typed their ratings.

The Test Phase was a probability rating task. Subjects were told they would see 60 slides of lesions, some of which they had seen in the first phase. Their task would be to rate the plausibility of each of three different diagnoses presented with each slide on a percentage scale, anchored with the descriptions absolutely sure at 100% and not in a million years at 0%. They were told that there was no necessity
to make the ratings add to any particular total, with the restriction that there would be a clear first choice (no two- or three-way ties). If there was a diagnosis that they did not recognize, they were to tell the experimenter. Ratings for these alternatives were omitted from subsequent analysis; this omission affected the correct diagnosis for one category for each of 2 subjects. The presentation of the slides and entry of the answers was the same as in the Study Phase. No feedback regarding the correct diagnosis was given until the entire set had been completed. In the immediate condition, the Test Phase immediately followed the Study Phase with a break only for instructions. In the delay condition, there was a 2-week (+1 day) delay between the two phases. In the debriefing session, the correct diagnoses of all slides was discussed, based on the diagnoses and descriptions given by a dermatologist, fulfilling a major motivation of the residents for participating in the study.

Results and Discussion

The purpose of the typicality ratings in the first phase was simply to familiarize the subjects with the slides. Therefore, although the typicality ratings were recorded for use in subsequent studies, they were not analyzed. Because of the counterbalancing of items used, there are two sets of comparisons of interest for the probability ratings collected in the Test Phase. One involves all four types of test items (old, SD-sim, SD-diff, and new diagnosis) and contains, for each subject, 10 items per group. The slides used in this comparison were fully counterbalanced across the four conditions. The second comparison involves only the two same-diagnosis groups (SD-sim and SD-diff) and contains 20 items per group (including the 10 also included in the four group analysis) for each subject. As previously mentioned, this additional group of SD-sim and SD-diff items was included so that all comparisons between conditions would be with material that had been rotated through the conditions being compared and would increase power on the critical comparison without exceeding limits of cooperation on session time. The diagnosis given the highest probability rating was taken to be the decision and was scored as correct or incorrect. The resulting data are displayed in Table 1.

For both sets of comparisons, the data were analyzed in a two-factor analysis of variance (ANOVA) with item type within subjects and delay between subjects. Significance level was set at .05 throughout. Considering first the comparisons involving the four item types (old, SD/sim, SD-diff, new), there was a significant effect of item type, \( F(3, 42) = 5.62 \), but no significant interaction with delay, \( F(3, 42) = .80, MSe = .0239 \). The effect of delay was not significant, \( F(1, 14) = 3.76, MSe = .176 \). The approach to significance of the delay variable did not affect the critical SD-sim and SD-diff items; comparison of the larger appearing difference, SD-diff(immediate) versus SD-diff(delay), yielded a nonsignificant \( t = .40 \). The second analysis concentrated on the critical similarity comparison, using all the data from the same diagnosis similar and same diagnosis different cells. This was a 2 × 2 ANOVA with delay as a between-subjects factor and similarity as a within-subjects factor. This analysis showed more correct answers for SD-sim items than for SD-diff items, \( F(1, 14) = 12.81, MSe = .00928 \), and no effect of delay or delay by similarity interaction (both \( F < 1 \)).

Table 1

<table>
<thead>
<tr>
<th>Item type</th>
<th>Training-test interval</th>
<th>Immediate</th>
<th>Delayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>85</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>SD-sim</td>
<td>67</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>SD-diff</td>
<td>62</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>68</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>All similars</td>
<td>65</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>All different</td>
<td>54</td>
<td>53</td>
<td></td>
</tr>
</tbody>
</table>

Note. SD-sim refers to items that are in the same diagnostic class as old items and are visually similar to old items. SD-diff refers to visually different items in the same diagnostic class. All similars refers to the combined scores for the SD-sim items rotated through all four item conditions and the SD-sim items rotated only through the SD-sim and the SD-diff conditions.

Thus, there is a strong and persistent effect of similarity to prior episodes. This effect of prior cases apparently does not reliably generalize to slides from the same category that appear different; in fact, different appearing lesions are given slightly lower probability ratings than the new items in the immediate condition. Thus, the prior exposure is not leading to category wide priming, but its influence extends beyond a presentation of the original material. Finally, whatever the effect of similarity is, it is not solely due to availability of the diagnostic category, because the correct diagnostic label was always one of the three alternatives presented for rating.

Experiment 2: Effect of Diagnostic Experience

The purpose of Experiment 2 is to see whether the effect of similarity to prior cases is attenuated or eliminated in higher levels of experience with diagnosing dermatological disorders. It would be entirely plausible that physicians who have seen many cases of each of these disorders are less subject to the effect of the slides that were seen in this particular context. In addition, it is possible that the cuing function of the multiple-choice alternatives would be more effective for the general practitioners because, with their more extensive knowledge, they might be able to deal with alternatives that initially seemed unlikely better than would the residents.

Method

Subjects. The subjects were 16 practicing family physicians and general practitioners recruited from the Hamilton, Ontario, Canada, area. Many of them were recruited in conjunction with a continuing medical education day devoted to dermatology. The time from graduation averaged 15 years, ranging from 5 to 49 years.

Materials and procedure. The materials and counterbalancing were identical to those used in Experiment 1. The procedure was identical to that used in the immediate condition of Experiment 1 with the exception that the responses were given orally by the physicians. Each physician was tested individually in his or her own office.
Results and Discussion

The data are displayed in Table 2 and were treated identically to those in Experiment 1. Again, there were two analyses, the first using the slides that were counterbalanced across old, SD-sim, SD-diff, and new conditions and the second using all slides in the SD-sim and the SD-diff conditions. The first ANOVA, examining four levels of similarity, treated item type as a within-subjects variable and a comparison with the residents of Experiment 1 as a between-subjects variable. There was no significant difference between the residents and the general practitioners ($F < 1$). There was a significant effect of item type, $F(3, 90) = 19.76$, but the interaction between experience and item type did not approach significance, $F(3, 90) = 1.49, MS_{e} = .0219$. The second analysis of similar and different items again treated experience as a between-subjects variable and item type as a within-subjects variable. Again, there was no significant difference between the residents and the general practitioners, $F(1, 30) = 1.04, MS_{e} = .0078$. There was a significant effect of item type, $F(1, 30) = 44.47$, but the interaction between experience and item type did not approach significance, $F(1, 30) = 1.56, MS_{e} = .0081$.

There is a strong effect of similarity to prior episodes for general practitioners as well as residents. As with residents, this effect of prior cases for general practitioners apparently does not reliably generalize to slides from the same diagnostic class that appear different. There is no apparent effect of experience and no interaction between similarity and experience. In a further analysis, there was not a significant correlation between overall accuracy on the SD-sim and SD-diff items and the number of years of practice. The apparent direction of the correlation was negative ($r = -.40$), an observation not uncommon in studies of diagnostic expertise (e.g., Evans et al., 1984). In addition, there was not a significant correlation between years of practice and the difference between SD-sim and SD-diff, which might have been a measure of a tendency to change reliance on similarity with increasing experience. Thus, there is no encouragement for the idea that increased experience leads to less reliance on similarity to contextually available cases.

Table 2
Percentage Correct Diagnoses for the General Practitioners in Experiment 2 and the Residents From Experiment 1, Pooled Across Delay

<table>
<thead>
<tr>
<th>Item type</th>
<th>General practitioner</th>
<th>Resident</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>85</td>
<td>80</td>
</tr>
<tr>
<td>SD-sim</td>
<td>75</td>
<td>68</td>
</tr>
<tr>
<td>SD-diff</td>
<td>58</td>
<td>60</td>
</tr>
<tr>
<td>New</td>
<td>54</td>
<td>61</td>
</tr>
<tr>
<td>All similars</td>
<td>70</td>
<td>65</td>
</tr>
<tr>
<td>All different</td>
<td>52</td>
<td>53</td>
</tr>
</tbody>
</table>

Note. SD-sim refers to items that are in the same diagnostic class as old items and are visually similar to old items. SD-diff refers to visually different items in the same diagnostic class. All similars refers to the combined scores for the SD-sim items rotated through all four item conditions and the SD-sim items rotated only through the SD-sim and the SD-diff conditions.

Considering the implications of the multiple-choice format, general practitioners did not seem to be any better than the residents in "filling out the hint" provided by the correct diagnostic label appearing among the alternatives. However, one could argue that the difference between the general practitioners and the residents is being underestimated by the multiple-choice procedure. That is, the general practitioners might be better at thinking of possibilities than are the residents, a skill that would be eliminated by providing the correct diagnosis among the alternatives given to both groups. If this is so, then the residents should show markedly worse performance when alternatives were not provided. Evidence relevant to this possibility is provided in the next experiment.

Experiment 3: Effect of Variations in Diagnostic Procedure

In Experiment 3, a group of family practice residents were encouraged to engage in a variety of different diagnostic procedures. It is possible that if they were "really careful," or systematically considered alternatives by arguing for them before giving a diagnosis, an effect of similarity would be much smaller. That is, a noticeable effect of similarity might be restricted to circumstances in which the subjects are not systematically deploying the analytic knowledge that they have. The training phase was the same as in the two preceding experiments. In the first portion of the Test Phase, subjects were asked to give a quick first impression of what the diagnosis was, after which, on the same trial, the alternatives were given orally and rated as in Experiment 2. In the second part of the Test Phase, for different slides, the subjects were asked to argue in turn for each of the three diagnoses before making their ratings. The different procedures were rationalized to the subjects in terms of a contrast frequently discussed in this medical school: "pattern recognition" versus considering the alternatives. Because the initial impressions are given without providing alternatives, this procedure makes two contributions to the present argument. First, it provides a high estimate of the potential effect of instance-based familiarity, in that it also includes an effect of prior instances on the availability of the diagnostic alternative. Second, it gives a low extreme from which to estimate the benefit to accuracy derived from the provision of diagnostic alternatives.

Method

Subjects. The subjects were 14 volunteer first-year family medicine residents from McMaster University who were paid for their participation.

Materials. The materials were adapted from the materials used in the first two experiments. Modifications were made to accommodate the extra time required by the argument condition and still stay within approximately an hour of running time. The only item types tested were old, SD-sim, and SD-diff. In addition, the diagnoses represented by only one quartet were dropped, leaving 12 diagnostic categories. Slides were divided into two sets to accommodate the two test phases; slides were rotated across item-type conditions within sets, and the assignment of sets to test phases was balanced across subjects.
Procedure. The Training Phase was the same as in Experiments 1 and 2. The Test Phase was divided into two portions. In the first portion of the Test Phase, subjects were asked to give a quick first impression of the diagnosis, after which the alternatives were provided as in the preceding experiments. The first-impression procedure was rationalized as encouraging the immediate recognition that is referred to as pattern recognition in medicine. For this first impression, the subjects were encouraged to respond quickly and to not worry if nothing came immediately to mind. The rating phase of each trial was run at whatever pace was comfortable for the subject.

The second Test Phase was explained to the subjects as an attempt to avoid premature closure. "Pattern recognition often gets bad press because it may lead to premature closure. In this phase, in order to protect against misleading first impressions, we will go to the opposite extreme. I'd like you to argue for each of the three alternatives before you give your final ratings. Some of the alternatives may be hard to argue for, but give each of them your best shot. This is meant to be the logically extreme form of considering differential diagnoses." For this phase, then, the alternatives were given one at a time in a random order. The Test Phase using the argument procedure was always run after the test phase using the first-impression procedure because we were concerned that if the argument procedure came first, it might reduce a potential difference between rating of the alternatives with and without arguments. As it turned out, the difference was not significant.

Results and Discussion

The data are displayed in Table 3 and were treated identically to those in Experiment 1. Inspection suggested that the two sets of slides used to balance slides across test phases might have had different effects; consequently, balancing set was included as a factor in all subsequent analyses. Again, there were two basic analyses, the first using the slides that were counterbalanced across old, SD-sim, and SD-diff conditions, and the second using all slides in the SD-sim and the SD-diff conditions. The first ANOVA treated item type and diagnostic procedure as within-subjects factors and balancing set as a between-subjects factor. There was neither a significant effect nor an interaction involving balancing set. There was a significant effect of both diagnostic procedure, $F(2, 24) = 4.31, M_{Ss} = .0246$, and of item type, $F(2, 24) = 4.10, M_{Ss} = .74$. The interaction did not approach significance, $F(4, 48) = 1.08, M_{Ss} = .0245$. The analysis using only the SD-sim and SD-diff data also showed no significant effect or interaction involving balancing condition. Again, there were significant effects of both diagnostic procedure, $F(2, 24) = 5.82, M_{Ss} = .0192$, and item type, $F(1, 12) = 7.70, M_{Ss} = .0524$, with no significant interaction, $F(2, 24) < 1, M_{Ss} = .0244$. Finally, subanalyses indicated a significant main effect difference between the first-impression (without alternatives) and rate-alternatives (with alternatives provided) conditions for both the three-item type data, $F(2, 24) = 10.7, M_{Ss} = .0160$, and the two-item type data, $F(1, 12) = 38.47, M_{Ss} = .0048$. Comparable analyses for the comparison between the rate-alternatives and the argument conditions (both with provided alternatives) showed no comparable effects ($F < 1$, in both cases), a lack of difference that might only mean that subjects were already effectively considering the alternatives in the rate-alternatives condition.

Thus, the effect of similarity was found in this study, with no evidence suggesting that it is independent on the diagnostic procedure used. The difference between giving first impressions without alternatives being provided and giving a diagnosis only after having argued for each of three alternatives is an extreme contrast, suggesting that similarity to contextually available cases is likely to be a normal part of diagnosis.

The fact that there was a difference in this study between the first-impression condition, in which no alternatives were provided, and the rate-alternatives condition, in which alternatives were provided, opens the possibility that Experiment 2 underestimated the difference between residents and general practitioners. Perhaps a difference would have appeared if the general practitioners had also been run under conditions in which alternatives were generated by the diagnosticians themselves. However, the fact that there was no interaction with similarity in either study indicates that this factor does not attenuate the major conclusion: An effect of similarity seems robust across both expertise level and diagnostic procedure.

In the future, we could attempt a more clinically plausible condition by asking the subjects to generate alternatives, rate them, and then rate any of the alternatives that we provided that had not already been generated, thus permitting the determination of the effect of similarity on the generation of alternatives. Another possibility for future investigation is provided by the observation that performance on the old items in all three experiments is considerably below 100%. If the diagnosticians could remember the diagnosis of the old items seen a short time earlier, then it is paradoxical that performance is so low on old test slides. One possibility is that the diagnosticians remembered the diagnosis that was given during training but did not use it; possibly they simply did not believe the diagnosis that we gave during the Training Phase or thought that it would be inappropriate for them to change from their original diagnosis. This problem probably could be remedied by instructions. A second possibility that is potentially more interesting is that the diagnosticians are primarily trying to diagnose the materials de novo and are not trying to retrieve the prior diagnosis. Our interest in this alternative is that if they were instructed to try to retrieve the diagnosis of old items, the diagnosis of similar items might also be facilitated.
General Discussion

It is apparent from these experiments that prior examples can and do influence diagnosis in dermatology. The differences between slides similar to those seen previously in the same context and slides that were different from prior cases in the same context varied from 10 to 20%, a clinically important and statistically significant effect. The effect is not a short-term effect, as it appears to be undiminished by a 2-week delay between presentation and test. It is not due to context-specific activation of a category as a whole, because it extends mainly to examples that are similar in appearance to the test slide. It is also found in general practitioners with considerable experience in the diagnosis of skin diseases, so it does not appear that the effect of examples is restricted to diagnosticians who only have a small stock of experience on which to draw. Finally, the variations in diagnostic procedure, ranging from giving a quick first impression to arguing for given alternative diagnoses, did not interact with the effect of similarity. The diagnosis of similar cases, then, does not seem to be dependent on failing to deliberately deploy the analytic knowledge that the diagnostician possesses.

Limitations of Canonical Rules

In general, the results are consistent with the previous work by Allen and Brooks (1991) using artificial materials. The subjects in both sets of experiments had access to explicit rules, as evidenced by their justification of their decisions, when asked, by naming the applicable features from the recommended rules. However, the impact of similarity to specific prior examples was still relatively large. To this extent, neither set of data can be handled by an independent-cues model that restricts itself to stable weightings of the features named in the rules. But there are informative differences between the results of Allen and Brooks and those of the present experiments. With the artificial materials, subjects spontaneously corrected virtually all of their errors within seconds. Thus, the primary effect of examples, when corrections are considered, was on response times, rather than on underlying competence. Because of the construction of the material, the rules were guaranteed to be sufficient for correct classification, and the features were evidently sufficiently unambiguous to allow nearly perfect corrected responding. The similarity effects with the artificial materials apparently resulted from the convenience, when categorizing new perceptual targets, of using memory for prior exemplars, rather than restricting response to relevant features. Such was not the case in the present series; the facilitative effects of appropriate matches on accuracy were not easily duplicated by deliberate, analytic effort.

One way of describing the difference in the materials is that the dermatological rules used here, as with most canonical medical rules, were insufficient. Without special experience with the materials, learners are not able to immediately assess the features specified by the rules and attain complete success. One obvious source of this insufficiency is that there is relevant diagnostic information that is not included in the rule. Academic dermatologists commonly specify further information verbally when they are discussing a case with a student. These additional descriptions are not normally included in either survey texts or initial instruction, undoubtedly partly because of their sheer volume. Despite this, the rules as given in texts, and known by our subjects, seem to have some privilege, both because they are commonly cited with an implication of sufficiency and because they include an overwhelming majority of the descriptions given when dermatologists are asked to justify their diagnoses.

Some of the insufficiency of the canonical rules seems traceable to limitations of the verbal medium. For example, feature weightings are seldom given beyond a necessarily vague statement such as "I pay special attention to mica-like scale in the diagnosis of Psoriasis." There is also incomplete specification of the features: For a beginner who has never seen the range of possible values that features can take, how dark is "intense pigmentation"? These matters of incomplete specification could in principle be explained by assuming that the learner accumulated separable information on the range of appearances that a given verbally specified feature could take and their contingencies with the diagnosis. As such, this effect of experience could be handled by an independent-cues regression model.

At least two considerations suggest that more than this specification of independent cues is operative. First, considerations within dermatology suggest there are obvious interactions of relevant and irrelevant features that, by their sheer volume, are infeasible to mention early in the course of instruction: The color "red-brown" means something different for Black and White skin; "pink or purplish" color is obviously different for people with or without a sunburn; "lichenification" (yellow, shiny hardening) appears different on young and old, male and female skin; and prior treatment (possibly done by the patient or another physician) can substantially alter visual appearance of individual features. These interactions are sometimes commented on, apparently in an effort to reconcile initially discrepant information. At a minimum, such attempts at reconciliation would require learning interactions between the specific relevant and nominally irrelevant features.

A second consideration indicating that a "learning-of-feature-appearance" version of an independent-cues model is insufficient is provided by a recent series of experiments in which the variety of forms that named features can take was varied (Regehr & Brooks, 1991). Using a feature and rule structure similar to that used in Allen and Brooks (1991), some animals were constructed from "unimorphic" features, that is, features that were identical from animal to animal. Other animals were constructed from "multimorphic" features, that is, features that were unique in appearance to each animal, although they might be categorically the same. For example, although two multimorphic animals might have had six legs, the appearance of the six legs themselves was unique to that animal; the six legs on two unimorphic animals were indistinguishable. As a group, the unimorphic animals gave the subjective impression of all looking the same, although clearly some had two legs, some six, some had rounded bodies, some angular, and so on. In contrast, the individuality of the multimorphic animals was immediately apparent. A simple,
perfectly sufficient classification rule was given for both sets of animals. After experience in classifying a small number of unimorphic animals, classification time and accuracy was the same for familiar animals and new, for good analogies and bad. With the multimorphic animals, however, classification was considerably slower and less accurate on animals that looked like a familiar animal from another category, analogously to the Allen and Brooks results. In other words, despite knowledge of a simple rule, classification of the animals with an individuated appearance was affected by similarity to old exemplars. These materials should have been perfect candidates for a “learning-of-feature-appearance” version of an independent-cues model, yet they provided results that demonstrated the insufficiency of such a model. Obviously, our suggestion is that dermatological cases are multimorphic, rather than unimorphic, stimuli, which leads us to expect that learning of prior cases as well as individual-feature appearances is required to produce competent performance from knowledge of verbally presented rules.

Status of Similarity

We selected the SD-sim and SD-diff cases on the basis of immediate visual appearance and without any personal expertise in dermatology. Yet the categories constituted this way were sufficient to produce the similarity effects in this article. We take this as a demonstration that judgments of close similarity are not necessarily dependent on special expertise. However, these results cannot be interpreted as a demonstrating that similarity judgments in this area are theory-free. It has been pointed out that judgments of similarity depend on the perceptual apparatus and knowledge of the judge and should not be taken as theory-neutral (e.g., Murphy & Medin, 1985). Such comments are relevant to positivist claims about warrants for truth being provided by sense experience, but do not constitute an argument that judged similarity has no force independent of the best analytic theory that the judge has about the domain. There is every reason to believe that both we and our least experienced subjects already had ideas about dermatological similarity. It is self-evident that most people have considerable experience with skin and can construct theories from this experience. For example, we all know what healthy skin is; red and swollen will probably be taken as importantly different from dry and scaly; hands, faces, and scalps will probably be presumed to be importantly different areas. By contrast, judged similarity in other medical disciplines, such as histological pathology, is surely more dependent on specialized expertise than is dermatology; most people will not have had experience with stained and mounted specimens in everyday life. However, even in such an area, it might be true that a relatively shallow level of expertise would be sufficient to exploit close similarity.

Regardless of how much special expertise is required to make judged similarity useful in medicine, we have every reason to take close similarity as a special case. When we think about a problem such as “is a whale more like a bear or a shark,” it is easy to be impressed with the theory dependence of the weighting of particular features in a judgment of similarity. However, in this example, one is impressed with the number of differences as well as the number of similarities among these three animals. If a person is viewing a whale for the first time and the closest physical matches in memory are sharks and bears, then we could expect variable and slow classification performance. But on viewing a second whale, the person has available whatever was learned on the first encounter. On the first encounter, whether a bear or a shark was retrieved as an analogy would depend critically on how the various stimulus attributes were weighted; on the second encounter, the challenge would be to find a weighting scheme in which the processing of the first whale would not be retrieved. The extreme of judged near identity is a special case that should be respected in discussions of similarity, although still not taken as being completely theory-free. There are circumstances, including the ones in the current experiments, under which judgments of close similarity on many physical attributes are likely to be less variable and less dependent on special expertise than are judgments that two items are “merely conceptually the same.” As such, judgments of close similarity must be considered in accounts of expertise and instruction in areas utilizing perceptually rich stimuli.

Scope of Results

The influence of prior close matches in other areas of medical diagnosis or other domains of expertise remains to be demonstrated. There are two obvious possibilities that might limit the generality of the current demonstration to a special case. The first possibility, unlikely in our view, is that there is something uniquely literal about dermatology. Certainly, dermatology would appear to have some characteristics that lend themselves to effects of close surface similarity: The stimuli are perceptually rich and highly configural, and the features are potentially ambiguous. Nevertheless, discussions with dermatologists show interpretations of the surface stimuli that are reminiscent of the processing to deep representation that is well-documented in other visual domains of medicine, such as radiology (Lesgold et al., 1988). Dermatologists routinely turn immediate appearance into an interpretation that includes causal inferences (e.g., causation by viral, bacterial, contact, or allergic agents), inferences regarding time progression (e.g., a herald patch in early stage pityriasis rosea, fissuring in late stage dysidrosis), and discounting of variations in surface appearance caused by prior treatment (e.g., cortisone cream increases redness, ASA (aspirin) paste alters the appearance of warts). Any influence of close similarity occurs despite ample evidence that experts are not treating the stimuli in a uniquely literal fashion. Dermatology is clearly at an extreme in the availability of immediate, highly informative perceptual information. But there is nothing immediately obvious to suggest that it is unique in either medicine (e.g., radiology and histology) or everyday areas of perceptual categorization.

A second factor that might severely limit the scope of these findings is the possibility that the cases we used as positive and negative matches might represent a degree of match that would have little chance of occurring naturally. We currently have no convincing answer to this criticism. Our impression is that positive matches as convincing as the ones used here
are not uncommon in the slide collections that we examined, despite a presumed motive for not photographing too many cases having nearly identical appearance. What is clear is that we did not need to engage in extravagant search times for quartets, even when, for current research, we were trying to control typicality. These impressions were shared by our consulting dermatologist.

Regardless of the incomplete nature of the current evidence, we think that the influence of prior diagnostic episodes is worth pursuing. Most particularly, we are interested in variability in episodic encoding and retrieval as a source of variability in diagnosis. Diagnostic models based solely on stable knowledge structures have few intrinsic resources for dealing with the unreliability that is of concern in medicine. Episodic influences have been documented as producing interesting and surprising effects on other categorical identification tasks (e.g., Jacoby, Kelley, Brown, & Jasechko, 1989; Whittlesea, 1987), effects that seem likely to have parallels in dealing with diagnostic unreliability. In addition, exciting new possibilities for assessing the diagnostic process are the procedures recently developed by Jacoby (in press) for probing awareness of knowledge source. None of the procedures in this article demonstrated whether the diagnosticians were aware at the time of diagnosis of an influence of prior instances. Jacoby is currently adapting these techniques to classification problems, which should provide an important tool for assessing and controlling sources of variability in diagnosis. Marrying the techniques that originally came from research on episodic memory to the problems of classification should provide interesting perspectives on the process of medical diagnosis.

References


Call for Nominations for *Neuropsychology*

The APA Publications and Communications (P&C) Board has opened nominations for the editorship of *Neuropsychology* for the years 1993–1998. B. P. Uzzell is the incumbent editor of this newly acquired APA journal in the area of experimental and applied neuropsychology, which will begin publication under APA in 1993.

Candidates must be members of APA and should be available to start receiving manuscripts in January 1992 to prepare for issues published in 1993. Please note that the P&C Board encourages more participation by members of underrepresented groups in the publication process and would particularly welcome such nominees.

To nominate candidates, prepare a statement of one page or less in support of each candidate. Submit nominations to

Martha A. Storandt  
Psychology Department  
Washington University  
1 Brookings Drive  
St. Louis, Missouri 63130

Other members of the search committee are Sandra P. Koffler, Charles G. Matthews, and Michael I. Posner.

Nominations will be reviewed individually as received to ascertain nominees' interest in being considered. The search committee will begin systematic review of all nominations sometime after August 15, 1991, and it is expected that a slate of possible nominees will be presented to the P&C Board at its October 25–26, 1991 meeting.