Physicians' Use of Deep Features: Expertise Differences in Patient Categorization

Sarah L. Devantier (sdevanti@uwo.ca)
Department of Psychology, University of Western Ontario
London, ON N6A 5B8 CANADA

John Paul Minda (jpminda@uwo.ca)
Department of Psychology, University of Western Ontario
London, ON N6A 5B8 CANADA

Wael Hadarra (Wael.Haddara@LHSC.ON.CA)
Schulich School of Medicine & Dentistry, University of Western Ontario
London, ON N6A 5C1 CANADA

Mark Goldszmidt (mgoldszm@uwo.ca)
Schulich School of Medicine & Dentistry, University of Western Ontario
London, ON N6A 5C1 CANADA

Abstract

Medical doctors make many decisions when interacting with patients, including diagnosing the problem and determining an appropriate treatment of the problem given the patient's circumstances. While diagnostic reasoning has been studied extensively, reasoning about patient management has not. Using a forced-choice triad task we investigated the use of deep structures related to diagnostic and management reasoning in novices (medical school students), intermediates (medical residents), and experts (endocrinologists). We found that expert participants are generally more likely to choose deep feature matches than intermediate and novice participants. Specifically, experts and intermediates are more likely than novices to choose deep matches on management trials while experts are more likely to choose deep matches than either intermediates or novices on diagnostic trials. These results reveal a difference in performance for intermediate subjects on management, versus diagnostic, trials. We suggest that expertise in management and expertise in diagnosis develop along different trajectories as physicians complete their medical training.

Keywords: medical reasoning; expertise; diabetes; problem solving; medicine; forced-choice triad task

Introduction

Medical doctors make many kinds of decisions when they interact with patients. Even once the diagnosis is completed, the physician must determine what the most appropriate treatment will be, and help the patient manage his disease. For example, when considering a diabetic patient who may need to go on insulin, a doctor must take into account both the biomedical aspects of the patient, and also issues related to the patient's lifestyle. A patient with severe arthritis might not be a good candidate for any drug that requires self-administration unless he or she has assistance.

While the issue of diagnostic reasoning has been well studied (e.g., Bordage, 1994; Bordage, 1999; Norman & Brooks, 1997), medical reasoning around patient management has received little attention in the literature. *Patient management* can be defined as treating the patient's disease or disorder in the context of the patient's life. The choice of what treatment to use for a particular disorder is not only based on the disease itself, but also on other factors such as medication tolerance, lifestyle, and patient compliance with a treatment regime. If a patient cannot or will not comply with a treatment plan, that treatment will not be an effective one. We believe this is something experienced physicians can take into account, while medical students are unlikely to consider anything other than disease-based or biomedical factors when determining patient treatments. We suggest that patient management is an important part of the physician-patient encounter, and we predict that physicians will demonstrate expertise effects with patient management that is analogous to the kinds of expertise effects that are found in other areas of medicine and other domains in general. Of particular interest to our investigation is the idea that experts can often ignore or suppress attention to the surface features of a problem and attend instead to the deep, solution-relevant features of a problem.

Expertise effects have been well studied in the cognitive literature. For example, Chi, Feltovich and Glaser (1981) asked physics Ph.D. students (experts) and undergraduate students (novices) to sort 24 physics problems based on similarities of solution and to explain the reasons for their groupings. Novices sorted the problems on the basis of surface features. That is, they grouped problems on the basis of the literal physics terms explicitly mentioned in the problem and the physical configuration described in the problem. Experts, on the other hand, sorted their problems on the basis of deep features that were related to the major

physics principles governing the solution of each problem. Similar effects have been found with experts in tree classification and fish classification (Medin, Lynch, Coley, & Atran, 1997; Shafto & Coley, 2003).

These studies have implications for the study of how physicians think about patient management. We argue that expert physicians should be able to perceive deep structures related not only to diagnosis but also to the management of their patients. Perceiving this deep structure could assist the expert physician in making decisions about how to treat the patient, how to interact with the patient, whether or not to follow up with the patient for compliance, etc. We argue that all of these things are central to being a good physician, yet this kind of decision making is not typically the focus of investigation and research, and this kind of thinking process is not typically taught in medical school.

An initial goal for our research was to develop a task that was sensitive to expertise differences in diagnostic reasoning as well as reasoning about patient management. To do this, we turned to a forced-choice triad task that is commonly used in cognitive psychology (Johnson & Mervis, 1997; Lin & Murphy, 2001; Rabinowitz, & Hogan, 2002; Rips, 1989; Smith & Sloman, 1994), which involves choosing one of two items that best matches a target. Rabinowitz and Hogan (2002) successfully used this task to investigate the effects of expertise in statistics: subjects were presented with three statistics problems in a forced-choice triad task. The target matched one problem in terms of surface features (similar story characters and similar kinds of dependent/independent variables) and another in terms of deep features (solution-related features like the kind of statistical tests needed to solve the problem). Rabinowitz and Hogan found a positive correlation between the number of statistics courses taken (expertise) and the tendency to choose pairs that were related in terms of deep features, indicating that more expertise in statistics generally resulted in more attention to deep, solution-relevant features of the problem.

In the present research we asked practicing physicians, residents, and medical students to complete a series of forced-choice triads. We predicted that expert physicians would be better able to perceive and react to deep features and would be able to appreciate the similarity between patients who might require similar management approaches, as well as patients who have similar diagnoses. To the extent that experts perform well on classification in which the deep-feature match deals with management, we argue that this ability develops with relevant clinical experience rather than with explicit medical training. Put another way, we argue that physicians become experts in patient management by managing patients, not by being in medical school. This implies several other predictions. First, we predict that novices should perform poorly on all triads regardless of whether the deep-feature match is related to diagnosis or management because they do not have either the clinical experience or the medical training to recognize deep matches. Second, because they have clinical experience, intermediates should be sensitive to deep-feature relations about patient management. However, since many of our deep diagnostic matches are based on detailed biomedical information, and given that medical residents have not completed their endocrinology fellowship training, it is unlikely they will have the capacity to recognize deep diagnostic matches.

Method

Subjects

Three groups of subjects were tested. *Novice* subjects were 15 second-year and third-year medical school students who had not yet completed an endocrinology rotation at the Schulich School of Medicine and Dentistry. Novices were recruited via mass email sent by the Department of Medicine to all students. *Intermediate* subjects were 8 medical residents (postgraduate years 1 and 2) at the Schulich School of Medicine and Dentistry recruited via email from the third and fourth authors. *Expert* subjects were 13 endocrinologists from across Canada recruited via email from the third author.

Materials

The forced-choice task consisted of a series of ten items, each with three hypothetical patient profiles: a target and two possible matches. All of the patient profiles were designed by the third and fourth authors, and each one was rated by a third physician for understandability and

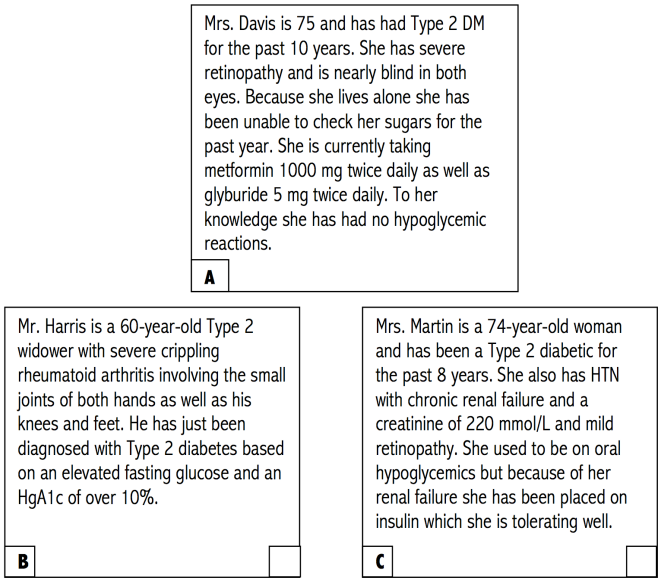

Figure 1: Sample Triad

Table 1: Re-scoring information

	Overall		Surface to Deep		Deep to Surface	
	Number Changed	Proportion Changed	Number Changed	Proportion Changed	Number Changed	Proportion Changed
Expert	13	.036	3	.008	10	.028
Intermediate	16	.044	2	.006	14	.039
Novice	21	.058	0	0.0	21	.058

readability. The rater used a scale of 1 through 7, 1 being very difficult to understand/read, and 7 being very easy to understand/read. All profiles that were used in the study were rated at least 6 for each of understandability and readability.

Half of the triads were designed so that the deep match was one that related primarily to patient management. One of the triads used in this study is shown in Figure 1. The target is shown first (Profile A), followed by two potential matches (Profile B and Profile C). In this example, Profile C is the surface match because both patients are older females with similar disease duration. Profile B is the deep match because both patients cannot be treated with injectable insulin: neither can check their blood sugars by themselves due to blindness and arthritis. The other half of the triads were designed so that to recognize the deep match, subjects had to pay attention primarily to diagnostic information (e.g., diagnosing concurrent symptoms or underlying causes of diabetes).

Procedure

Subjects completed the triad task online. As described above, subjects were primarily recruited via email, which contained a link to the survey site. Survey software was provided by the University of Western Ontario's ITS department. Demographic information, including level of schooling, number of years in practice, and proportion of patients with diabetes, was completed first.

Subjects were instructed to choose the profile that made the "best" pair with the target profile; they were not informed that we were investigating deep matches, or that we had described some triads as "diagnostic" and others as "management" related. The ten triad questions were presented in a fixed random order. On each triad, subjects made a selection and provided a short explanation to justify their choice: this justification allowed subjects' choices to be verified. Subjects were required to submit an answer before they could move on to the next question; once submitted, answers could not be changed. The entire process took approximately 30 minutes to complete, though there was no time limit and participants could view questions as long as they liked. In addition, subjects were permitted to log out and return to the task if they wished.

Re-scoring[1]

Before the primary analyses were carried out, we verified each subject's responses (deep or surface) by examining the justification for each choice. A deep choice should be accompanied by a justification that indicates recognition of the relevant features or aspects of the case. If the justification did not match the response, it could be due to a guess, an error, or a response bias: in this case, the choice may not represent the subject's ability and the response should be rescored. A small number of responses (fourteen percent) were changed after scoring by two independent raters who were blind to the experience level of each participant. Raters separately decided whether each participant's text on each question indicated they were using surface or deep features to inform their match. Item responses were changed if the participant's text answer clearly indicated they were relying on surface (or deep) structures to make their match when they chose the deep (or surface) answer; answers lacking enough information to determine if they used surface or deep criteria were not changed. Raters then met and compared their scoring response-by-response. If a consensus was reached, the answer was changed; however, if consensus could not be reached, the answer was not changed.

Fifty of a total of 360 responses (36 subjects x 10 answers per subject) were recoded (14%). See Table 1 for details of the proportion of responses changed for each group (novices, intermediates, and experts). Of the 50 responses that were rescored, 74 percent (37 responses) were agreed on prior to consultation; 26 percent (13 responses) were changed through discussion. Seventeen responses were not changed after discussion, either because agreement could not be reached, or because through discussion the raters agreed the response should not be changed.

Results

The primary dependent variable in our study was the proportion of deep-feature choices by each subject. As such, we scored each item in terms of deep-feature responding.

[1] Analyses of Variance were also significant before rescoring for Overall performance ($F(2,33) = 6.87$, $p = .003$), performance on Management triads ($F(2,33) = 4.47$, $p = .019$), and performance on Diagnostic triads ($F(2,33) = 4.39$, $p = .020$).

Each participant's score on the task was recorded as a proportion ranging from 0 to 1, 0 indicating all surface responses, and 1 indicating all deep responses. Responses were examined both as an overall score, including both types of triads (a total of 10 items), as well as management and diagnostic triads separately (five items in each group). Figure 2 shows the average proportion of deep matches for each group of subjects.

The data show a general effect of expertise in which the experts chose the greatest proportion of deep responses, followed by the intermediates and novices. We entered the overall proportion-deep scores for each subject into an ANOVA with expertise (expert, intermediate, novice) as a between subjects factor. We found a significant effect of expertise, $F(2, 33) = 10.71$, $p < .01$. A post hoc Tukey HSD test indicated the performance by the experts exceeded that of the novices (M's = .48 and .17, $p < .01$). The performance of the intermediates ($M = .36$) also exceeded performance of the novices ($p < .05$). The performance of experts and intermediates did not differ.

We also examined the proportion of deep responses by subjects for each type of triad separately. Recall that half of the triads were designed so that the deep-feature match was related to patient management and the other half were designed so that the deep feature match was related to a diagnostic issue. An ANOVA with expertise as a between-subjects factor on the proportion-deep responses for the Management triads found a significant effect for expertise, $F(2, 33) = 4.94$, $p < .01$. A post hoc Tukey HSD test indicated the performance by the experts exceeded that of the novices (M's = .48 and .19, $p < .05$). The performance of the intermediates ($M = .45$) nearly exceeded performance of the novices ($p = .07$). The performance of experts and intermediates did not differ from each other.

An ANOVA with expertise as a between-subjects factor on the proportion-deep responses for the Diagnostic triads found a significant effect for expertise, $F(2, 33) = 9.29$, $p < .01$. A post hoc Tukey HSD test indicated performance by the experts exceeded that of the novices (M's = .48 and .16, $p < .01$). The performance of the experts nearly exceeded that of intermediates as well ($M = .29$, $p = .07$). The performance of intermediates and novices did not differ from each other. In short, for the management triads, experts and intermediates were equally likely to make deep matches and were also more likely to do so than the novices. For the diagnostic triads, the intermediates and novices were equally likely to make surface matches, and were more likely to do so than experts were.

Comparisons within groups

Although we were primarily interested in the effects between groups, we also analyzed the data within each group. That is, the data shown in Figure 2 indicate that

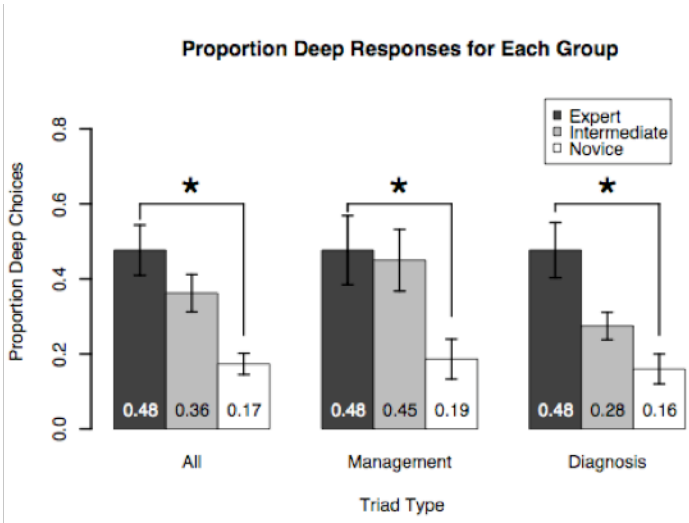

Figure 2: Average Novice, Intermediate, and Expert responses on the forced-choice triad task. Lower proportions indicate more surface responding. Error bars represent standard error of the mean (SEM). Significant differences (at $p < .05$) are flagged with an asterisk.

experts seem to perform similarly on both management and diagnostic triads, as do novices. The intermediates, however, seem to be more likely to make deep matches on the management triads than the diagnostic triads. In order to examine this effect, we conducted three paired t-tests. Each test compared the difference between diagnostic and management performance for each group of subjects. As expected, there was no difference between management and diagnostic triads for the experts, $t(12) = 0.0$, *ns,* or for the novices, $t(14) = 0.35$, *ns*. However, we did find a nearly significant difference between the management and diagnostic triads for the intermediate subjects, $t(7) = 2.20$, $p = .06$.

Discussion

The suggestion that experts and novices would differ in their ability to classify patients was investigated using a forced-choice triad task. We found an overall effect of expertise such that experts chose the deep feature match more often than did the novice subjects. This was true both for triads in which the primary deep feature match was diagnostic in nature and triads for which the deep feature match was related to patient management. Novice subjects, on the other hand, were unlikely to make deep matches on any triads and generally tended to choose matches on the basis of surface features. This suggests that our triad task was quite sensitive to the difference between novice and expert subjects. The tendency to make decisions on the basis of deep features is a hallmark of expert-level performance in many domains, (Chi et al., 1981; Rabinowitz & Hogan,

2002), and our results suggest that this general tendency is found in medicine as well.

Our data also suggest that intermediates tend to choose deep-feature matches on the management triads, producing expert-like performance, while they tend to choose surface-feature matches on the diagnostic triads, producing novice-like performance. This pattern was shown in our overall result in Figure 2 and also by the paired t-test analyses. This suggests that expertise in some aspects of clinical thinking may develop earlier in a physician's career than other aspects. More specifically, expertise effects in patient management might start to develop soon after a physician begins seeing patents during his or her residency. Clearly the sort of diagnostic sophistication needed to appreciate the deep features of the diagnostic triads is not quite present in the intermediates. Residents are responsible for much of the day-to-day patient care activities required in a hospital, which may give them a large amount of experience in these kinds of tasks in a short amount of time. The management issues studied in this task were likely more general in nature, applicable to a broad range of patients, while the diagnostic issues studied were specific to diabetes and/or endocrinology knowledge. It is likely that without specific training it would be difficult to recognize the deep diagnostic matches. It is more likely that intermediate subjects will have encountered management issues similar to those exemplified in this task than diagnostic ones, even if they have not yet completed an endocrinology fellowship. At this point our data do not distinguish between the possibilities that diagnostic expertise may depend on specific training received during an endocrinology rotation, or it may simply take longer to develop.

We suspect that both kinds of expertise, management and diagnosis, depend on some combination of direct experience with patients and training. Since the diagnostic triads depend on detailed knowledge about concurrent illnesses, secondary causes of diabetes, or specific complications of diabetes, it is likely that the novices, and even some of the intermediates do not have the required knowledge to diagnose the relevant conditions. Therefore, less experienced subjects likely do not have the diagnostic categories of patients that are required to make the deep feature match. Management categories of patients could come online much earlier in a physician's training: as soon as a student begins seeing patients, they may be creating these categories, recognizing that treatment for apparently disparate patients may be very similar. Though doctors may not be consciously aware of using these categories, our study shows that experienced physicians are more likely to use deep features to classify patients than novices are.

Our results provide predictions for other situations and other experts. For example, we expect that nurses, who manage patients, would be likely to make deep matches on management triads, perhaps even better than our expert sample, since much of their attention would be on management, rather than diagnostic, issues. However, since they do not have the specific training required, we expect they would often use surface features when matching on diagnostic triads.

It is also unclear at this point what physicians would do if management and diagnostic triads were pitted against each other. In our current task, there is only one deep match in each triad; however, it is possible to design a triad such that subjects would have to choose between a diagnostic match or a management match. This kind of design would allow us to determine which type of match is more salient to physicians of differing levels of experience (or different types of experience, such as nurses), or if there are certain situations that make management or diagnostic issues more salient. For example, when seeing a patient in the intensive care unit (ICU) management issues may be more relevant. Two patients could require immediate care for survival, even though they do not share a diagnosis. On the other hand, when seeing patients as a family doctor, it is possible that physicians could match patients based on either management or diagnostic issues. Further research will be required to determine how deep structures compete with each other.

Acknowledgements

This research was supported in part by a Discovery Grant Natural Sciences and Engineering Research Council of Canada (NSERC) and a grant from the Academic Development Fund of the University of Western Ontario, to John Paul Minda, and an Ontario Graduate Scholarship: Science and Technology (OGSST) to Sarah L. Devantier. We thank Wayne Weston, M.D. for rating the stimulus material, and Geoff Norman, M.D. for helpful comments on early versions of the project.

References

Bordage, G. (1994). Elaborated knowledge: A key to successful diagnostic thinking. *Academic Medicine, 69*(11), 883-885.

Bordage, G. (1999). Why did I miss the diagnosis? Some cognitive explanations and educational implications. *Academic Medicine, 74*(Suppl 10), S138-S143.

Chi, M. T. H., Feltovich, P. J., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science 5*(2), 121-152.

Johnson, K. E., and Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology*, *126*(3), 248-277.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General, 130*(1), 3-28.

Medin, D. L., Lynch, E. B. & Coley, J. D. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32*, 49-96.

Norman, G. R. & Brooks, L. R. (1997). The non-analytical basis of clinical reasoning. *Advances in Health Sciences Education,* 2, 173-184.

Rabinowitz, M., & Hogan, T.M. (2002). Using a triad judgment task to examine the effect of experience on problem representation in statistics. In J.D. Moore & K. Stenning, (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning*. New York, NY: Cambridge University Press.

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naive similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(4), 641-649.

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition, 22*(4), 377-386.