From deep to superficial categorization with increasing expertise

Thomas C. Ormerod
Lancaster University, UK.
T.Ormerod@lancaster.ac.uk

Catherine O. Fritz
Bolton Institute, UK.
cof1@bolton.ac.uk

James Ridgway
University of Durham, UK.
jimridgway@durham.ac.uk

Abstract

An experimental study of task design expertise is reported wherein a set of 12 mathematics tasks were sorted by specialist designers of mathematics tasks and by experienced mathematics teachers without specialist design experience. Contrary to the frequent finding of increasing conceptual depth with increasing expertise, conceptual depth did not differ between groups. Teachers sorted on the basis of mathematical content earlier than designers, and were more specific in their content-based categories. Designers produced more sorts than teachers and were more individualistic in their sorting. These findings suggest that domain expertise does not necessarily impair creative problem solving, as has been suggested in other studies. Instead, expertise includes the ability to shift perspectives with respect to the domain.

One of the basic phenomena of skilled performance is an increasing conceptual depth at which domain knowledge is mentally represented. Experts are able to call upon abstract and generalizable representations, such as schemata, which they subsequently adapt to meet current task demands. Typically, these representations embody fundamental principles that capture significant and useful commonalities among domain problems. In contrast, novices rely upon shallow representations that focus upon superficial features of the domain or task. For example, McKeithen, Reitman, Rueter & Hirtle (1981) investigated recall by intermediate and novice programmers. Intermediates recalled programming terms in an order that suggested organisation by algorithm or function, whereas novices' recall orders reflected superficial relations, nicely illustrated by the recall chunk of the terms "bits", "of" and "string". Similar expert/novice differences have been found in many domains, such as computing systems (Doane, Pellegrino & Klatzky, 1990), physics (Chi, Feltovitch & Glaser, 1981), geometry (Koedinger & Anderson, 1990) and experimental design (Schraagen, 1993). Where expert performance is based on deep representations, it is characterised by the rapid recognition of problem states and the structured development of solutions following a predetermined pattern.

While a deep conceptual representation confers many advantages, there may be situations where the re-use of established domain knowledge is insufficient or inappropriate. For example, Adelson (1984) presented expert and novice programmers with abstract (output-oriented) or concrete (step-by-step) program flowcharts prior to participants answering abstract or concrete questions about program code. She found that experts made fewer errors than novices when the level of abstraction of flowchart and question matched. However, where an abstract flowchart primed a concrete question, novices outperformed experts. The source of this effect appears to be the misapplication by experts of conceptual knowledge primed by the abstract flowcharts. More recently, Wiley (1998) has shown that priming of domain knowledge can impair performance in a remote associates task, in which participants are presented with three words, such as *plate*, *broken* and *shot* , and are required to generate a fourth word, such as *glass*, that forms a familiar phrase with each of the three presented words. Baseball primes impaired performance on trials where only the first word fits a baseball theme (e.g., *home plate*).

These studies, combined with other demonstrations of impaired expert performance, encourage a view that experts are unable to 'turn off' their deep domain knowledge when it is inappropriate for task performance. However, it might be argued that these demonstrations are artefactual. In these studies, experts' skills are systematically undermined, either by domain priming, as in Wiley's (1998) study, or by having them perform a task that typifies novice problem-solving behaviour, as in Adelson's (1984) study. One might argue that the message of these studies is simply that experts make poor novices. Whether there are cases of realistic domain activities where the presence of deep conceptual knowledge impairs expert performance remains to be demonstrated.

Of particular interest to the current research is Wiley's (1998) suggestion that domain knowledge can sometimes act to inhibit creative problem-solving. Design is a creative problem-solving activity where a case might be made for expertise involving more than re-use of conceptual knowledge. Design has been studied extensively (e.g., Goel & Pirrolli, 1989), and evidence has accumulated showing the same kinds of conceptual representation underlying expert design that are found in other domains of expertise (e.g., Visser, 1991). However, the application domains of these studies (architecture, engineering and software design) are constrained, either by technology or by the design brief or context, such that highly original solutions are the exception rather than the rule (see Goel, 1994, for a useful exposition on the nature of design constraints). When originality is the primary concern, prior knowledge may be less efficacious, perhaps leading to design fixation (Jansson & Smith, 1991).

The development of instructional tasks presents an interesting test case of creative design, and is the focus of the current study, conducted as part of a wider investigation into the nature of task design expertise funded by the UK Economic and Social Research Council. Changes in educational practice, such as increasing use of problem-based teaching, place an emphasis on creativity in task design. This is magnified by the need for tasks that are motivating for

To appear in Proc. 21st Annual Conference of the Cognitive Science Society, Vancouver, August 1999.

students, and that address curriculum and assessment goals without disenfranchising minority groups.

We have recently carried out a study (Ormerod and Fritz, 1999) in which we analysed verbal protocols of designers developing novel tasks to appear in English as a Foreign Language (EFL) textbooks The protocols of specialist designers, experienced authors of EFL textbooks, were typified by the early depth-first development of multiple task ideas, prior to a phase in which a single idea was developed in breadth. In contrast, the protocols of non-specialists, experienced EFL teachers without specialist task design experience, reveal an initial phase in which a single generic task was developed in breadth and subsequently instantiated in depth. The early depth-first work of specialists allowed them to generate and test alternative task ideas to a point where task feasibility could be evaluated before choosing one to develop more completely. The task generation of non-specialists appears constrained by their application of a pre-determined 'schema' embodying a generic task structure. Although it is difficult to assess objectively, the specialists' tasks appear more original than those of the non-specialists, whose tasks were strongly based on popular ESL textbooks. This is supported by protocols in which non-specialists refer to common ESL task types (e.g., 'information-gap').

What is the source of creativity in task design shown by specialists? The non-specialists' behavior appears consistent with Wiley's (1998) suggestion that conceptual knowledge can impair creative problem-solving. So, why does it not impair that of specialist task designers? It is possible, though unlikely, that specialists do not have the deep conceptual representations of tasks that ESL teachers develop. Alternatively, it may be that specialists acquire strategic knowledge that enables them to bypass the application of conceptual domain knowledge where necessary, or that they acquire alternative conceptual structures that take precedence over principles-based conceptual knowledge in design contexts. The study reported in the remainder of this paper set out to explore these hypotheses.

The study

One approach to exploring expertise is the sort method (e.g., Hoffman, Shadbolt, Burton & Klein, 1995). In this method, participants are given sets of domain item descriptions, and are required to sort these into categories according to one or more dimensions that are significant to them. By examining the nature of the sorts produced (e.g., the dimensions used to define categories, assignment to categories and the order in which dimensions are produced), one can infer something about participants' mental representations of conceptual knowledge. Chi et al (1981) used the sort method to explore expert/novice differences in conceptual representation of physics knowledge. They found that sorts produced reflected a deep/superficial distinction, with experts sorting according to underlying principles and novices sorting according to surface features of physics problems. Similarly, Schoenfield & Herman (1982) used the sort method to investigate mathematics expertise, again replicating the deep/superficial distinction. The sort method has been used to explore other forms of expertise such as programming (Davies, Gilmore and Green, 1995), Archeology (Burton, Shadbolt, Rugg & Hedgecock, 1990). and engineering design (Ormerod, Rummer & Ball, in press).

The present study used the sort method to investigate expertise in the design of mathematics tasks. Because we were not interested in studying mathematical expertise per se, which generally occurs alongside expertise in the design of mathematical tasks, it was important that our expert and non-expert groups be well matched with respect to their education and experience with mathematics. We selected specialist designers of assessment items for English exam boards as our expert group. Mathematics teachers, with equivalent educational and teaching backgrounds, served as our non-expert group. Because domain content varies considerably depending upon school year, we targeted GCSE-level mathematics (equivalent to the middle of high school). Both the designers and the teachers worked primarily at the GCSE level. The cards to be sorted each contained a task from the prior year's GCSE exams (e.g., Figure 1), so as to be realistic and familiar to both groups.

We were interested to see whether teachers and designers, being well matched on most dimensions other than actual design expertise, would perform the card sorts differently. If designers, more than teachers, are fixed in their concepts regarding tasks, then we would expect to find designers producing fewer sorts than teachers. Unless task design is idiosyncratic, designers might also be expected to be more similar to one another than teachers. On the other hand, if designers benefit from greater flexibility in their approach to tasks, then they should produce more sorts and be less fixed in their assignment of tasks to conceptual categories. Unlike other sort studies (e.g., Chi et al, 1981; Schoenfeld & Herrmann, 1982) we did not specify the sort dimensions or the pre-classify tasks according to conceptual level. Our interest lies in the sorts that participants produce spontaneously to reflect their own choice of dimensions.

Method

Participants

Participants included 20 GCSE-level math teachers from Northwest England and 14 GCSE task designers from 4 different exam boards. The designers were also experienced in teaching mathematics with 4-10 years experience in designing tasks for GCSE examinations.

Materials

Twelve GCSE tasks were selected from the MEG and NEAB 1996 exams, and were reduced to fit on A5 card (as in Figure 1). The tasks were selected to be representative of the exams as a whole, while still being reasonably related to other tasks in the set. Tasks were selected from all three exam levels (lower, intermediate, and higher).

Design and Procedure

Expertise was a between-participants factor with two levels, designers and teachers. The study was conducted in the form of an interview between experimenter and participant. Each participant was interviewed individually

for approximately one hour. All participants received the tasks approximately one week prior to their interviews so that they could look them over at their leisure. The interviews began with participants giving an account of their teaching or design training and experience. The sort and another related activity were counter-balanced with respect to order between participants. For the sort, the experimenter then demonstrated the sort activity, sorting a set of mammal names twice under example dimensions such as ferocity and attractiveness, while giving a verbal account of her reasons for choosing each dimension, category and assignment.

Participants were instructed "I'd like for you to organise these tasks into groups, more or less as I've just done with the animals. You can make as many or as few groups as you choose. Sort in ways that are useful and meaningful to you as a professional. All of the tasks are different, but sort them based upon the commonalities that you identify, that is, how they fall into different categories for a particular dimension of your choice. " Participants' verbalizations were recorded, along with a record of the task groupings derived from each sort. Where the participant's verbal report had not already revealed sufficient information concerning the dimensions and categories of a sort, the experimenter indicated each of the groups in turn, asking "What makes these form a group?". When all categories were described by the participant, the experimenter asked "Is there some overall theme or explanation to the way these have been grouped?" After each sort the tasks were shuffled, and the participant was asked to produce another sort under a different dimension. Participants were encouraged to continue with further sorts for as many dimensions as they could reasonably form.

Results

Designers produced reliably more sorts than teachers. (Designers mean = 4.2 , sd=0.77, teachers = 3.5, sd=1.05), $t(33)=2.17$, $p=.037$. A hierarchical cluster analysis of the participants was run using Euclidean distances and a complete linkage procedure. For each participant, each task pair was assigned a score calculated as the percentage of times that the pair was assigned to the same group by that participant. Thus, if two tasks were always grouped together by one participant, then the score for that pair for that participant would be 100%. If two tasks were never grouped together, the score would be zero. If the participant produced four sorts and the two tasks were grouped together in one sort but were not together in any of the others, the score would be 25%. The clusters that emerged from the analysis, using a reasonable cutoff, included six pairs of designers, 1 pair of teachers, and 4 larger groups (Ns=5,4,3,3) containing teachers with a single designer. There were no instances of designers forming clusters larger than a pair.

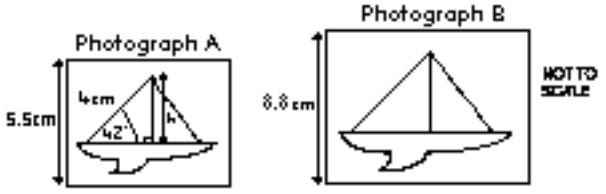
MEG Paper 5, 1996

For examiner's use only

10. Two photographs of a yacht are pictured to the right.

Photograph B is an enlargement of photograph A.

Photograph A has width 5.5 cm and photograph B has width 8.8 cm.



(a) (i) Find the scale factor of the enlargement. Give your answer in form $\frac{p}{q}$, where p and q are whole numbers.

Answer (a) (i) _____ (2)

In photograph A the sail of the yacht is a triangle with one side 4 cm and one angle 42 °.

(ii) Find the length of the corresponding side of the sail on photograph B.

Answer (a) (ii) _____ cm (3)

(iii) Write down the size of the corresponding angle on photograph B.

Answer (a) (iii) _____ (1)

(b) calculate the height, h, of the mast of the yacht in photograph A.

Answer (b) _____ cm (3)

Figure 1. An example of a GCSE Mathematics task used in the study (© MEG examination board, UK).

Designers and teachers used many of the same dimensions, but one interesting difference was that many teachers identified 'thinking' tasks as a category whereas designers identified 'open' tasks as a category. (See Table 1.) In addition, designers were more likely to sort on the basis of the level of the tasks than were teachers, and produced 'open', 'thinking' and 'level' sorts earlier than the teachers.

All participants except one teacher included at least one sort based on the mathematical content of the tasks. All 19 teachers who produced a content-based sort, produced it as

their first sort. Nine of the 15 designers led with a content-based sort, but many designers began by sorting on the basis of level and referred to math content in the later sorts. This difference is reliable, U=85.5, p=.048. The number of math content sorts did not differ reliably. Designers produced an average of 4.8 groups (sd=1.01) and teachers produced 5.3 groups on average (sd=1.06), t(32)=1.43, p>.05.

Nevertheless, it was apparent from cluster analysis that there were differences in the ways that designers and teachers sorted the tasks. (See Figure 2.) This cluster analysis, again using hierarchical clustering, Euclidean distances, and a complete linkage procedure, was run using only the participants' math content-based sort. (Four teachers provided more than one content-based sort; we used only the first content-based sort from each of these participants.) The linkage distances do not show a marked increase, used to suggest a cutoff for accepting clusters, until near the end of the run, at approximately 4.8. Whether using that cutoff, or no cutoff at all, it is evident that there are two main clusters forming, and that those clusters are specialist designers on the one hand, and teachers on the other. Designers and teachers were sorting the tasks differently.

Detailed examination of the categories and category members assigned by the two groups provides some explanation. Teachers often produced more specific categories, such as 'linear inequalities', 'fractions', and 'number patterns' as compared to the more general category of 'Algebra and number' which was more often adopted by designers. When more specific categories were collapsed to form the four Attainment Targets defined for the English mathematics curriculum (Applying & using math; Algebra & number; Shape & space; and Data handling), designers' and teachers' sorts were very similar with the only notable difference being in the use of the Data handling category, which designers used far more than did teachers. Otherwise, it was clear that teachers' and designers' perceptions of the tasks in terms of gross mathematical content were not distinguishable; differences were primarily due to the greater specificity on the part of the teachers.

Table 1. % of designers (D) and teachers (T) producing dimensions and mean position in which the dimension. Other dimensions produced by < 10% participants were Complexity, Exam board, Wordiness, Response type, and Mark.

Participant	Math topic	Level	Openness	Thinking vs rote	Difficulty	Context	Prefs / turn-offs	Graphics	Ramping
% D	100	60	50	14	14	33	29	14	21
% T	95	20	10	37	35	15	20	25	0
Position D	1.6	1.6	2.7	2.0	2.5	3.2	3.5	3.0	3.7
Position T	1.0	2.5	3.5	3.3	2.8	2.7	2.8	2.4	0

Discussion

Although designers produced reliably more sorts than teachers, it is possible that the difference was due to the different structures of their work days. Most teachers were scheduled to teach shortly after the interview whereas designers were less rigidly scheduled. However, teachers did not appear distracted and seemed to be as fully occupied by the task at hand as the designers. Furthermore, a related activity (not reported in this paper) was also scheduled during the hour; for half of the participants the sort occurred first and for the other half, the sort occurred last. If teachers limited their responses, then a difference between those who sorted first and second would be predicted. No difference was found (Teachers = 3.6 and 3.4 sorts, designers = 4.1 and 4.3 sorts first and second, respectively.)

The results suggest that different kinds of domain role invoke different kinds of conceptual representation. Teachers appear primarily to use the kinds of conceptual representations found in other studies of expert knowledge (e.g., Chi et al, 1981). This knowledge is precisely what is needed for the task of selecting appropriate Mathematics exercises for a particular stage of the curriculum. Designers, on the other hand, appear to use a wider range of conceptual representations, of which principles-based deep conceptual representations are not always primary. There are two potential explanations for this. The first is that designers have lost or under-rehearsed their principles-based representations. This seems unlikely given that all the designers used Math content for one of their sorts. The second is that design requires different kinds of knowledge to teaching. 'Superficial' dimensions may reflect the very things that make tasks interesting, original and practicable. A similar finding of distinct types of conceptual representation underlying different forms of expertise in the same domain was made by Weiser & Shertz (1983), again using a sort paradigm to explore conceptual representation. They found that expert computer programmers sorted problems by algorithm type while novices sorted by application area.

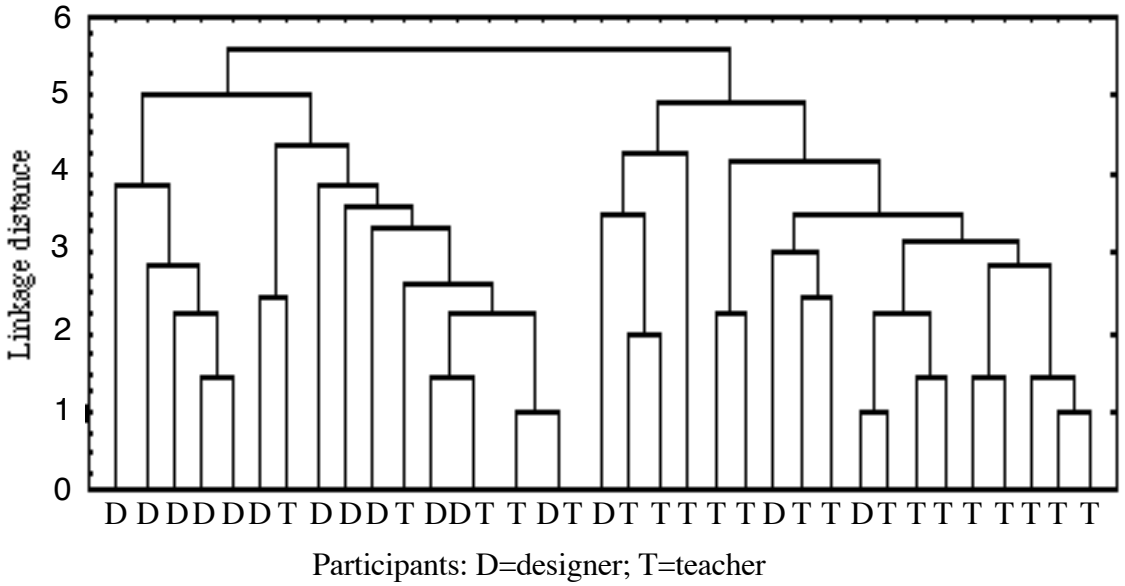


Figure 2. Cluster analysis of designer and teacher groupings based upon sorts under a Mathematics Topic dimension

In contrast, programming managers, all formerly experienced programmers, sorted by 'kinds of programmer' needed to solve each problem.

In much of the expertise research reported in the literature, there is an assumption that what is elicited through methods such as recall and sorting is relatively static. This assumption uynderlies reports of impaired performance resulting from inappropriate application of expert domain knowledge. We argue that to characterise experts' conceptual knowledge in this way is to miss an essential aspect: Experts have many layers of domain knowledge, and, when they are given realistic domain roles and contexts in which to perform, they know when and how to use it and when not to use it.

The notion of static conceptual representations is further challenged by Barsalou's (1985) distinction between taxonomic and goal-directed categories. In this view, goal-directed categories and their members are not fixed, but are determined by the task faced by the individual at any one time. The distinction between goal-directed and taxonomic categories has important methodological implications for the use of the sort method in studies of expertise. It has been suggested by some authors (e.g., Burton et al, 1991) that the sort method provides an equally informative but more cost-effective method for knowledge elicitation than traditional methods such as the analysis of verbal protocols. However, studies that restrict participants to a single sort or that impose pre-specified dimensions may limit elicitation to the kinds of taxonomic knowledge that underlie routine expertise, and fail to capture the sorts of goal-directed categories that may underlie highly skilled performance in non-routine activities.

Acknowledgements

The research reported in this paper was supported by a grant from the Economic and Social Research Council's (UK) Cognitive Engineering initiative, No. L127251031. We thank all the designers and teachers who took part in the study.

References

Adelson, B. (1984). When novices surpass experts : the difficulty of a task may increase with expertise. Journal of Experimental Psychology : L,M&C, 10, 483-495.

Barsalou, L. W. (1985). Ideas, central tendency, and frequency of instantiation as determinants of graded structure in categories. Journal of Experimental Psychology: L,M&C, 11, 629-654.

Burton, A. M., Shadbolt, N. R., Rugg, G. & Hedgecock,A.P. (1990). The efficacy of knowledge acquisition techniques, Knowledge Acquisition, 2, 167-178.

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

Davies, S. P., Gilmore, D. J., & Green, T. R. G. (1995). Are objects that important? Effects of expertise and familiarity on classification of object-oriented code. Human-Computer Interaction, 10, 227-248.

Doane, S. M., Pellegrino, J. W., & Klatzky, R. L. (1990). Expertise in a computer operating system: conceptualization and performance. Human-Computer Interaction, 5, 267-304.

Goel, V. (1994). A comparison of design and non-design problem spaces. AI in Engineering, 9, 53-72.

Hoffman, R. R., Shadbolt, N. R., Burton, A. M., & Klein, G. (1995). Eliciting knowledge from experts: A methodological analysis. Organizational behavior and human decision processes, 62, 129-158.

Jansson, D. G., & Smith, S. M. (1991). Design fixation. Design Studies, 12, 3-11.

Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. Cognitive Science, 14, 511-550.

McKeithen, K., Reitman, J.., Rueter, H., & Hirtle, S. (1981). Knowledge organization and skill differences in computer programmers. Canadian Jnl of Psychology, 13, 307-325.

Ormerod, T. C., Rummer, R., & Ball, L. J. (in press). An ecologically valid study of categorisation by designers. D.Harris (Ed.), Cognitive Ergonomics & Engineering Psychology Hampshire: Ashgate.

Ormerod, T. C. &. Fritz., C.O. (1999). Strategy changes across phase in problem solving: the case of task design Unpublished MS. Lancaster University.

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem-solvers. Journal of Experimental Psychology: L,M&C, 8, 484-494.

Schraagen, J. M. (1993). How experts solve a novel problem in experimental design. Cognitive Science, 17, 285-309.

Visser, W. (1990). More or less following a plan during design: opportunistic deviations in specification. Int. Journal of Man - Machine Studies, 33, 247-278.

Weiser, M., & Shertz, J. (1983). Programming problem representation in novice and expert programmers. Int. Journal of Man - Machine Studies, 19, 391-398.

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge increative problem-solving. Memory & Cognition, 26, 716-730.