Knowledge Acquisition (1990) **2,** 167–178

The efficacy of knowledge elicitation techniques: a comparison across domains and levels of expertise

A. M. BURTON, N. R. SHADBOLT, G. RUGG AND A. P. HEDGECOCK
Department of Psychology, University of Nottingham, Nottingham NG7 2RD, UK

(Received 17 November 1989)

Despite an increased interest in knowledge elicitation, there is still very little formal evidence evaluating the relative efficiency of the techniques available. In this paper we compare four KE techniques: structured interview, protocol analysis, card sort and laddered grid. Studies are reported across two classification domains, using eight experts in each. Despite its common usage, protocol analysis is shown to be the least efficient technique. The implications of this finding are reviewed. Finally, a study is reported in which non-experts are subjected to "knowledge elicitation". Subjects entirely ignorant of a domain are able to construct plausible knowledge bases from common sense alone. The ramifications of these findings for knowledge engineers is discussed.

Introduction

Over the past two or three years, interest in knowledge acquisition has increased radically. It is generally accepted in the expert systems community that knowledge acquisition is a considerable problem, and that research is needed to develop efficient techniques for acquiring codifiable knowledge. In response to this, there is now a large literature describing a very large number of techniques in detail (e.g. Gaines & Boose, 1988; Boose & Gaines, 1988). Furthermore, considerable effort has been invested in automating some of these techniques (Boose, 1989), and in tailoring them to specific domains (Marcus, McDermott & Wang, 1985; Klinker, Bentolila, Genetet, Grimes & McDermott, 1987).

Despite the large-scale investment in development of techniques for knowledge acquisition, there has been surprisingly little effort invested in evaluating the effectiveness of the techniques. It is often asserted that different techniques should suit different domains, different types of knowledge engineer and expert, and different types of expert system architecture. However, this assertion is very seldom backed with empirical evidence. If the techniques under development are to be useful to the expert systems community, then developers must address themselves to the question: "under what circumstances will particular techniques be most useful for knowledge acquisition?"

In this paper we present a comparison of four knowledge elicitation techniques across two domains. The work presented here forms a development of previously published work in this area, and so we will briefly outline our previous results, before describing the present study in detail.

In previous experimental work, we have compared four knowledge elicitation techniques: structured interview, protocol analysis, laddered grid and card sorts. The first two of these techniques are commonly used in expert system development,

1042-8143/90/020167 + 12$03.00/0

and we will label them *traditional* in what follows. The latter two techniques are intended to reveal an expert's "conceptual map" of a domain, but do so in ways which are likely to be unfamiliar to the expert. For this reason we label laddering and card sort *contrived* techniques. For the purpose of previous experiments, and of the new work presented here, we have developed standardized versions of each of these techniques (Shadbolt & Burton, 1990). The standardized versions have been used in construction of commercial systems, as well as in laboratory experiments. In summary, the structured interview comprises a planned session in which the interviewer is restricted to a small number of prompts. The protocol analysis consists of the expert solving a problem in front of the elicitor, and being asked to "think aloud" during this process. The laddered grid is a technique in which the elicitor constructs a well-defined structured representation of the domain in collaboration with the expert. This can be drawn as a graphical representation of the domain. The card sort involves the expert repeatedly sorting into meaningful piles a deck of cards, on each of which is marked the name of a domain element. Shadbolt and Burton (1990) describe each of these techniques in full, with examples from the standardized versions.

For the purposes of formal comparison, it is necessary to compare these techniques across *many* experts. There are clearly idiosyncratic features associated with any individual KE session with any individual expert. In order to provide *generalizable* results, it is necessary to perform multi-expert studies which transcend these idiosyncrasies. For this reason, two large-scale experimental comparisons of techniques were performed, each on 32 experts, within the domains of igneous rock identification (Burton, Shadbolt, Hedgecock & Rugg, 1987) and glacial feature identification (Burton, Shadbolt, Rugg & Hedgecock, 1988). In order to render these large-scale studies feasible, the domains were severely restricted. For each study, an experienced professor constructed a sub-set of the domain which would be well-known by a subject pool of students. Hence the subjects in these experiments were experts only in domains of very limited scope.

The results of these studies provided a number of statistically reliable results. The protocol analysis was the least effective of all techniques studied. The contrived techniques took less time to administer and code than the interview, but provided about the same amount of information. However, protocol analysis took longer, and provided a smaller coverage of the domain than any other technique tested. The importance of this result is highlighted by the very high frequency with which protocol analysis is used, and the very low frequency with which contrived techniques are used by knowledge engineers (Cullen & Bryman, 1988).

The subjects used in these experiments were indeed expert in the small domains under study. There were virtually no errors made in any of the KE sessions used in the experiments. However, it is possible that they have different characteristics to a population of experts typically used in knowledge acquisition. Our subjects were young, articulate, and possibly quite used to being placed in "odd" situations in the course of their education. It is necessary then, to establish whether these results apply to a population of genuine experts. In this paper we describe similar studies to those described above, but this time performed on "adult", real experts. Of course, it is not possible to take the time of 32 real-world experts in a homogeneous domain. In each domain then, we have studied just eight experts. This does not allow us to

make statistical comparisons of the kind possible in previous studies. However, it is rather more reliable than the single-expert case study common in the literature (e.g. Smith & Baker, 1983; Butler & Corter, 1986; Schweickert, Burton, Taylor, Corlett, Shadbolt & Hedgecock, 1987). In short, our aim in this study is to discover whether the results from previous laboratory-based studies are likely to apply when dealing with fully fledged experts in the context of "real world" knowledge engineering. In effect we are trying to establish the ecological validity of our previous studies.

Experiment 1

DESCRIPTION OF DOMAINS

The two domains of interest in this study were both archaeological: the identification of flint artefacts from Stone Age tool production, and the identification of pottery sherds from the mediaeval period of English history (roughly 1066 to 1485). In common with previous studies, these are both domains rich in classification knowledge.

Flint domain: Analysis of flint artefacts is a well-established field in archaeology. For the purposes of this study, we constructed a set of sixteen artefacts with the assistance of several experts. These were chosen to be a representative sample of flints. Each of these artefacts is either a tool, or a by-product of tool production, which has a generic classification in the appropriate literature. These items cover the palaeolithic, mesolithic and neolithic periods, the three periods into which the Stone Age is usually divided. All were British, and non-regional, to avoid biases between experts familiar with particular regional specialities. Any expert in this field would be expected to be familiar with all these artefacts.

Pottery domain: analysis of pottery sherds often takes place as part of an archaeological excavation. Experts use information from these analyses to help them establish information about the site under study (e.g. trading practices). Diagnostic dimensions include such factors as the colour of the fabric, the glaze used, the thickness of the pot wall, and so on. A set of 16 sherds was assembled with the help of experts in pottery identification. These were chosen as a representative sample of mediaeval pot types. So, for example, half were glazed, there was a mixture of coarsewares and finewares, and there was a spread of examples with different surface decorations and material inclusions.

SUBJECTS

Eight flint experts and eight pottery experts agreed to take part in the experiment. The 16 experts were all professional archaeologists engaged in academic, museum or similar work, and specializing in one of the domains under study.

METHOD

The four KE techniques under study were: structured interview; protocol analysis; laddered grid; and card sort. As already indicated, we label the first two of these techniques "traditional", and the last two "contrived". Each expert took part in two knowledge elicitation sessions, one using a traditional, and one using a contrived technique. For each domain then, two experts were subject to each of the four

possible combinations of techniques. This design provides four elicitation sessions of each type in each of the two domains. Experts were randomly allocated to these conditions, and the order of testing was counter-balanced.

In the Structured Interview condition, experts were subject to the standardized interview referred to above. The protocol analysis consisted of experts being given a randon subset of the domain items (chosen independently for each session), and asked to "think aloud" while deciding on its classification. Similarly, in the Laddered Grid condition, experts were asked a series of questions while the elicitor systematically explored the domain space. The Sorting Condition was different for the two groups of experts. The flint experts were asked repeatedly to sort cards each bearing the name of one of the sixteen domain elements. However, the pottery experts were asked to sort not *cards* but *sherds*, the artefacts themselves. This allows us to compare the two variants of the sorting task within the experiment.

The choice of dependent variables for this type of study is always problematic. One needs to establish measures of "effort" taken in each session, and measures of "gain". As with previous studies (Burton *et al.*, 1987; 1988), effort was measured by the time taken for each KE session and the time taken to transcribe and formulate each session into pseudo-English production rules. This form of intermediate representation (Young, 1987) has an IF, AND, THEN structure, but is not of a directly implementable form. We describe elsewhere the method by which raw data is coded into this intermediate representation (Shadbolt & Burton, 1990). The measure of gain was the number of clauses present in the transcribed rules. A clause is defined as one conditional statement in a pseudo-English production rule. This is clearly a very broad measure of the amount of information elicited in each session. An attempt to refine this measure is described below in Experiment 2.

RESULTS AND DISCUSSION

The effects on the main dependent variables are shown in Table 1. We will consider the results from the flints domain first. It is clear that in this domain, the contrived techniques take considerably less time to administer than the traditional techniques. This disparity is also present in the time taken to transcribe the KE sessions into the intermediate representation. However, for this latter variable, protocol analysis is seen to take much longer than its comparable technique, the interview. Considering

TABLE 1
Mean scores for subjects in each domain. n = *4 per group*

Domain	Flints				Pottery			
Technique	Interview	Prot An	L Grid	Sort	Interview	Prot An	L Grid	Sort
Time to KE (min)	59	57	36	25	47	50	38	54
Transcription time (min)	159	294	109	74	193	126	139	91
Total time	217	351	145	98	240	176	177	145
No. clauses	270	269	188	123	317	184	278	216
Clauses min^{-1}	1·2	0·8	1·3	1·3	1·3	1·0	1·6	1·5

the *gain* measure, we can see that the traditional techniques yield much more information than the contrived techniques. In terms of *effort for gain*, it can be seen that protocol analysis is the least efficient technique, while the remaining techniques perform comparably.

The results for the pottery domain show a rather different pattern. The time taken for a knowledge elicitation session is roughly equal for interviews, protocol analyses and sorts, with laddered grids taking a shorter period. The difference in the sorting results between the two domains is almost certainly due to the different tasks presented in each domain (e.g. card sort versus sherd sort). In this domain, the smallest number of clauses is delivered by the protocol analysis. The efficiency scores, as with the flint domain, show that protocol analysis is the least efficient technique, while the contrived techniques are the most efficient.

In summary, protocol analysis performs the least efficiently in both these domains. This the same pattern of results as reported in the large scale experimental studies described above (Burton *et al.*, 1987; 1988).

In addition to these measures, it is also possible to examine the *type* of knowledge elicited by techniques. Results from both domains show a very small overlap between the two sessions with the expert. In the flint domain, there was an average of only 10% overlap between knowledge gained from a traditional and a contrived technique. In the pottery domain, five experts provided no overlap between sessions, and the remaining three experts provided only a very small amount. This suggests that the techniques provide different types of knowledge. Although further studies would be necessary to confirm this result, it suggests that knowledge engineers should consider using these contrived techniques as a supplement to the traditional techniques. In these studies, not only are the contrived techniques comparatively efficient, they also add to knowledge gained from traditional techniques.

A problem with the present study is the lack of a rigorous measure of information gained. In closed-world domains such as those used for our previous experimental studies, one can take a measure of coverage of a known "gold-standard" rule base. In the present case this is not possible. It is quite possible that the coding of information by "clauses" is misleading as a measure of gain. For this reason, a further study was performed in one of the domains. Information coded as English rules was passed back to subjects who had taken part in Experiment 1, and rated by them on a number of criteria.

Experiment 2

METHOD

All the experts from one of the domains used in Experiment 1 (flints) were contacted and asked if they would be willing to take part in a follow up. Five of the original eight subjects agreed to proceed. Each of these experts was presented with the codified rule set generated from each of their two knowledge elicitation sessions. They were asked to read through these rule sets and to categorize each clause as *true, trivial, garbled* or *false*. The *true* and *false* categories are self-explanatory. The definition of *trivial* given is that the clause is not relevant to the diagnosis. *Garbled*

was defined as being partially true, i.e. the clause would become *true* if some modification was made. Although these definitions are rather informal, the experts were perfectly able to understand and use them consistently.

After this coding had been made, subjects were asked to make one final contribution. Having seen the rule set derived from their original KE sessions, the experts understood the form of the intermediate representation device used in this study. They were now asked to construct such a rule base from scratch. They were instructed to present rules which would cover the domain under study. These new rule sets were then compared with the original rule sets generated from KE sessions and measures of overlap were computed.

RESULTS AND DISCUSSION

The results of the rule-set coding were averaged by the technique used to elicit the initial rule-set, and the means of this analysis are presented in Table 2. The results show that very few clauses were rated as false for any of the techniques. The laddered grid and interview techniques provided similar distributions of codes for clauses, with around 60% of the originals being rated as *true*. The protocol analysis and card sort techniques both produced fewer *true* clauses, though the distribution of codes for remaining rules was different in the two cases. For protocol analysis, the majority of clauses not rated *true* were classified as *garbled* (32%), whereas for card sort, most clauses not rated *true* were classified as *trivial* (44%). The garbled nature of the rules from the protocol analysis is probably due to the fact that raw data from this technique is difficult to interpret. The transcripts are often rambling, ungrammatical, and give the appearance of being "garbled". It is possible that the inefficiency of protocol analysis is due not to features of the session itself, but to the difficulty of formulating data from these sessions into a suitable representation. In short, there may be more in a KE session than a knowledge engineer can extract. This result has important implications for the treatment of raw data from KE sessions. Work on automated "concept editors" (e.g. Anjewierden, 1987) is currently aimed towards making this part of the knowledge engineering process more efficient.

Table 3 shows the overlap scores between the elicited rule base and the rule base constructed directly by experts, averaged by the techniques of the original elicitation. Two overlap scores are necessary as the knowledge bases are of differing sizes. We therefore take the proportion of rules from the KE session (1st pass) which also appear in the direct transcription (2nd pass), and also the converse. The

TABLE 2

Experts' coding of the rule set generated from their own first pass elicitation session. Mean percentages (n *between 2 and 4 per group*)

Code(%)	True	Trivial	Garbled	False
Interview	63	22	8	6
Protocol analysis	46	17	32	5
Card sort	43	44	7	6
Laddered grid	55	22	17	6

TABLE 3
Comparison of experts' first and second pass rule sets. Mean percentages (n *between 2 and 4 per group*)

Overlap (%)	1st in 2nd pass	2nd in 1st pass
Interview	54	33
Protocol analysis	25	11
Card sort	29	33
Laddered grid	45	33

table clearly shows that protocol analysis provides the smallest overlap with subsequently elicited knowledge, card sort the next lowest, while laddered grid and interview perform roughly equivalently.

There are two possible reasons for the relatively small overlap between information gained from the protocol analysis and from the direct construction of rules. First, it is possible that the protocol analysis is simply a poor technique in this domain, and that a relatively small amount of information is picked up. Second, it is possible that that protocol analysis picks up information which is not of the type suitable for elicitation by direct elicitation. We favour the first of these explanations, as it seems to be in keeping with the results gained in Experiment 1.

Both the manipulations made in Experiment 2 suggest results consistent with previous experiments. Once again, we have demonstrated the relative inefficiency of protocol analysis and the relative efficacy of the two contrived techniques under study. We will return to a general discussion of this finding at the end of the paper. However, we will first consider another factor contributing to the problem of KE: the level of expertise.

Experts and novices

The studies described in Experiments 1 and 2 were performed on subjects specially chosen to be experts in their domains. From these studies we have been able to draw conclusions about the relative efficacy of KE techniques. In this section we consider the effect of expertise itself on these techniques. We will describe results from a study in which the same experiments were repeated on complete novices in the same domain. The purpose of the study was twofold, firstly to discover the baseline performance of each of the techniques, and secondly to discover whether the techniques offer any reliable way of distinguishing between experts and novices. Before describing the study, we will first discuss the background behind these issues.

It is possible that what we have discovered in Experiments 1 and 2, and in previous experiments, is a pattern of the dependent measures inherent in the techniques themselves. It may be, for example, that protocol analysis will *always* deliver information in a less efficient way than other techniques, no matter whether it is used for knowledge elicitation, for task analysis or whatever. Alternatively, it could be that we have discovered a pattern of efficacy of techniques which is specific to this class of domain, with this class of expert. In order to distinguish between these two hypotheses it is necessary to establish a *baseline* efficiency for these techniques, the equivalent of *chance performance* in other experimental studies.

The establishment of baseline performance will also be useful to knowledge elicitors in new domains. It is possible that the techniques will give *nothing* when the subject knows nothing about the domain under study. Alternatively, it is possible that an intelligent subject could make plausible guesses about the structure of the domain from common sense alone. In the survey published by Cullen and Bryman (1988), by far the most commonly reported problem associated with knowledge elicitation was *quality of expertise* (47% of respondents reported this problem). In a sense this is not surprising, and tallies with our own experience. Management involved in supervising the construction of an expert system are often unwilling to donate the time of the most efficient expert. It is often the experience of knowledge engineers that they are faced with an expert who has been promoted above the job which forms the basis for the system.

For this reason, it is important to discover whether the baseline performance on these techniques is very low, or whether baseline performance would actually appear to contain useful knowledge to someone new to the domain. Furthermore, if indeed there is a high baseline associated with the techniques, it would be useful to know whether any of them could function as an "expert spotter" (or perhaps non-expert spotter). If there were a standardizable way of discriminating on purely syntactic grounds between genuine experts, and those who could merely construct a plausible story around the domain, this would be of benefit to the knowledge engineering community.

With both these considerations in mind, we repeated the study described above in the flint domain with complete novices in this domain. Although this may sound bizarre, it is in fact not a hopeless task. For example, one of the domain elements here was *plano-convex knife*. While novices would almost certainly not know the technical meaning of this term, most would be able to construct a plausible description of the artefact. Similarly, two of the elements were *burin* and *microburin*. While most subjects did not know these terms, they were often able to hypothesize a relationship between them. In this case, the particular discrimination is diagnostic of expertise, as a *burin* is something quite different to a *microburin*, and they do not bear the common sense relationship implied by their names. We now describe the experimental details of this study.

Experiment 3

METHOD

Eight subjects were recruited who had no experience or interest in archaeology. The subjects were all University researchers in different disciplines, and so it can be assumed that they were generally matched, on for example general intelligence, to the sample used in Experiment 1.

Subjects were told that they were to take part in a "Psychological Experiment", in which their task was to try to make sense of the flint domain. The various KE techniques were explained to them, though the context of knowledge elicitation for expert systems was not explained at the start of the experiment. Subjects were properly de-briefed after completing both sessions.

The design of the experiment was exactly as in Experiment 1. Two subjects were

allocated at random to each of the four possible combinations of KE techniques (one traditional and one contrived technique). This provides data for four KE sessions for each technique. The same "effort" metrics as above were used. However, to measure gain, only "clauses gained" was used, as there is no sense in which the information gained from these subjects could be compared with knowledge from genuine experts.

RESULTS AND DISCUSSION

Despite reporting that they found the task difficult, no subject withdrew from the experiment, and all appeared to make a genuine effort to carry out the instructions. Analysis of the transcripts of KE sessions shows that all subjects were able to construct plausible accounts of some parts of the domain at hand. We will discuss this further below.

Table 4 shows the mean dependent measures taken by KE technique. Comparison with Table 1 shows that these measures give a relatively *flat* pattern across all techniques. Unlike experts, these subjects do not show a characteristic pattern of efficiency across techniques. Although there are some small differences in the "effort" measures, these are not very marked, by comparison to those evidenced in the expert studies reported above. The "gain" measures are flat across all techniques. In short, it appears from these results that any technique is as good as any other to the non-expert. This is evidence that the results gained in previous experiments are a consequence of the use of these techniques specifically as KE techniques. It does not seem that we have merely being measuring aspects of the techniques themselves.

The base-line performances are interesting here. Looking at the "gain" measure, it can be seen that the techniques *appear* to provide information. In fact, subjects are able to offer a set of quite plausible "rules" which an engineer new to the domain would have difficulty in faulting. Interestingly, the same information is repeated across all combinations of techniques. Even though the subjects know nothing of the domain, they conclude the same in both the KE sessions they attend.

This is rather a worrying result for knowledge engineers. For someone approaching the domain for the first time, it is necessary to discriminate between "genuine" knowledge, and knowledge which is plausible but unfounded. In detailed analysis of the transcripts from these sessions, we were unable to discover a predictive *reliable* discrimination metric between sets of rules generated by experts or non-experts. A

TABLE 4
Mean scores for non-expert subjects in the flint domain (n = *4 per group*)

Technique	Interview	Prot An	L Grid	Sort
Time to KE (min)	15	29	17	15
Transcription time (min)	43	48	49	43
Total time	58	77	66	58
No. clauses	75	85	69	77
Clauses min^{-1}	1·29	1·10	1·05	1·32

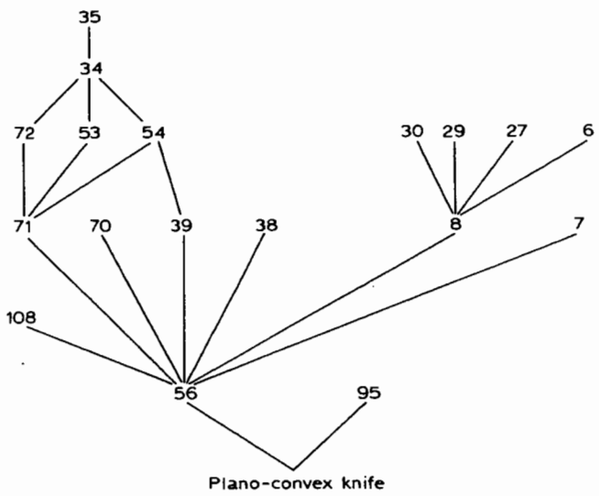

FIGURE 1. The connectedness of rules for a single decision elicited from an expert in the flint domain. Numbers refer to rule numbers.

number of metrics suggested themselves, though these were mostly circular, i.e. one would need to know about the domain in order to discover whether one was talking to an expert. This is clearly not the case in most knowledge elicitation sessions. Two possible discrimination functions are suggested by the data here. Although the data are not strong enough to make reliable recommendations, we will pursue these possibilities further. First, the results from non-experts are "flat", in that all elicitation techniques provide roughly the same amount of information (and usually the *same* information). This is not the case with genuine experts. If one were prepared to ask experts to take part in several types of elicitation session this might serve as a possible discriminator. The second possible discriminator relies on the *connectedness* of the rule set. Figure 1 shows a representation of rules leading to a single decision, gained from one expert in the flint domain. This can be read as an implication tree, where the numbers refer to rules. So, for example, rule 35 has a consequent which is used as an antecedent in rule 34. This shows complex interconnectedness of knowledge in the expert's transcript. This degree of complexity is *never* present in the non-experts' rule sets. Decision trees are all very shallow (i.e. very few rules chain to reach a particular goal).

While this result is interesting, it is not conclusive. The degree of interconnectedness is, of course, confounded with the amount of information gleaned from the two classes of expert. We plan further experiments (in new domains) to disentangle these two factors, and we hope to be able to use the interconnectedness and other features of the "topology" of rule sets as possible discriminators in the future.

Conclusions

In this paper we have described an experiment comparing the efficacy of four knowledge elicitation techniques, across two classification domains. As with

previous experiments, we have shown that protocol analysis performs poorly in this type of domain. Furthermore, we have shown that two contrived techniques perform very well, providing *complementary* knowledge to the standard interview, and doing so very efficiently. We have then described further experiments on novices which suggests an underlying problem in knowledge elicitation, that of inadequate expertise. In this final section, we will offer some concluding remarks on the nature of knowledge elictation, and on the choice of KE technique.

Given the evidence against protocol analysis in classification domains, we should ask why it continues to be so popular. We suggest that the answer lies in the comfort of use of this technique. It is by now clear that experts typically prefer KE sessions which seem familiar to them in some way (e.g. Schweickert *et al.*, 1986). Furthermore, they are prone to object to the use of contrived techniques on such grounds as "I don't think that way". We have here provided further evidence supporting a lack of correspondence between an expert's view of the session, and its utility for a knowledge engineer.

Given the many social constraints on a KE session (e.g. Hart, 1986), it is not surprising that elicitors try to maximize the comfort of the session. However, we suggest that they do so at the expense of efficient elicitation. Clearly there are cases where one is so constrained by the opinion of the expert, that is would be foolish to risk losing his or her co-operation. However, we suggest that in most circumstances, it is worth bearing a little discomfort in order to improve the efficiency of elicitation.

Inadequate expertise is likely to continue to be a problem for those working in applied settings. Given the difficulty in finding a robust and reliable syntactic discriminator between transcripts from experts and non-experts, elicitors should be on their guard against this type of error in their future work. Although it is not always possible, we suggest that considerable time be put into the original selection of an expert. External validation of an expert's suitability will save considerable time and wasted effort in future sessions. Prerau (1987) has provided an account of how the selection process may proceed. Finally, we conclude that more empirical studies of the effectiveness of KE methods are needed. It is important that progress in developing new techniques is made in parallel with appropriate evaluation of these techniques. This is equally true of automated knowledge acquisition and traditional knowledge acquisition. Without proper evaluation, research in knowledge elicitation becomes an end in itself, rather than an attempt to solve an applied problem faced by those constructing knowledge based systems.

This work was completed under an award funded by the Alvey Committee (Grant number IKBS 134). We are grateful to the many experts who contributed to the experiments reported here.

References

ANJEWIERDEN, A. (1987). Knowledge acquisition tools. *AI Communications*, **0**, 29–39.

BOOSE, J. H. (1989). A survey of knowledge acquisition techniques and tools. *Knowledge Acquisition*, **1**, 3–37.

BOOSE, J. H. & GAINES, B. R., Eds. (1988). *Knowledge Acquisition for Knowledge-Based Systems, Volume 2*. London: Academic Press.

BURTON, A. M., SHADBOLT, N. R., HEDGECOCK, A. D. & RUGG, G. (1987). A formal evaluation of knowledge elicitation techniques for expert systems: Domain 1. In D. S. MORALEE, Ed., *Research and Development in Expert Systems IV*. Cambridge: Cambridge University Press.

BURTON, A. M., SHADBOLT, N. R., RUGG, G. & HEDGECOCK, A. D. (1988). Knowledge elicitation techniques in classification domains. *ECAI-88: Proceedings of the 8th European Conference on Artifical Intelligence*, pp. 85–90.

BUTLER, K. E. & CORTER, J. E. (1986). Use of psychometric tools for knowledge acquisition: A case study. In W. A. GALE, Ed., *Artifical Intelligence and Statistics*. Cambridge, MA: Addison-Wesley.

CULLEN, J. & BRYMAN, A. (1988). The knowledge acquisition bottleneck: Time for reassessment? *Expert Systems*, **5**, 216–225.

GAINES, B. R. & BOOSE, J. H. (1988). *Knowledge Acquisition for Knowledge-Based Systems, Volume 1*. London: Academic Press.

HART, A. (1986). *Knowledge Acquisition for Expert Systems*. London: Kogan Page.

KLINKER, G., BENTOLILA, J., GENETET, S., GRIMES, M. & MCDERMOTT, J. (1987). KNACK:
Report-driven knowledge acquisition. *International Journal of Man–Machine Studies*, **26**, 65–79.

MARCUS, S., MCDERMOTT, J. & WANG, T. (1985). Knowledge acquisition for constructive systems. *IJCAI-85*.

PRERAU, D. S. (1987). Knowledge acquistion in expert system development. *AI Magazine*, **8**, 43–51.

SCHWEICKERT, R., BURTON, A. M., TAYLOR, N. K., CORLETT, E. N., SHADBOLT, N. R. & HEDGECOCK, A. P. (1987). Comparing knowledge elicitation techniques: A case study. *Artifical Intelligence Review*, **1**, 245–253.

SHADBOLT, N. R. & BURTON, M. (1990). Knowledge elicitation. In J. R. WILSON & E. N. CORLETT, Eds., *Evaluation of Human Work: Practical Ergonomics Methodology*. Basingstoke: Taylor and Francis.

SMITH, R. G. & BAKER, J. D. (1983). The Dipmeter Advisory System: A case study in commercial expert system development. *IJCAI-83: Proceedings of the 8th International Joint Conference on Artifical Intelligence*.

YOUNG, R. M. (1987). *The Role of Intermediate Representation in Knowledge Elicitation*. Keynote Address to "Expert Systems 87", Brighton, UK.