Int. J. Man–Machine Studies (1988) **29**, 329–350

Validation in a knowledge support system: construing and consistency with multiple experts

MILDRED L. G. SHAW AND J. BRIAN WOODWARD

Department of Computer Science, University of Calgary, Calgary, Alberta, Canada T2N 1N4

(Based on a paper presented at the Second AAAI Workshop on Knowledge Acquisition for Knowledge-Based Systems, Banff, October 1987)

At the previous workshop on Knowledge Acquisition for Knowledge-Based Systems in 1986, criteria for a knowledge support system were discussed, and a preliminary version of KITTEN (Knowledge Initiation and Transfer Tools for Experts and Novices) was described and demonstrated on Apollo workstations. This study is a continuation of the validation studies done by Shaw & Gaines (1983), and investigates a framework for knowledge acquisition evaluation and validation. KITTEN has been evaluated against the first stage of the model and the results are reported in the two domains of spatial interpolation techniques to produce contour maps and in trouble-shooting and maintenance of valves for oil and gas pipelines. Some preliminary results are described on validation experiments to show the extent to which experts agree with each other, with themselves at a later date, and with the results of the processing of their knowledge. Some of the questions asked were:

(1) To what extent does an expert find the generated rules meaningful?
(2) Do experts agree on their terminology in talking about a topic?
(3) To what extend do experts agree among themselves about the topic?
(4) Does an expert always use the same terminology?
(5) To what extent does each experts agree with the knowledge at a different time?

Introduction

In the last Workshop on Knowledge Acquisition for Knowledge-Based Systems, it was clear that several researchers were considering the potential of the repertory grid (Kelly, 1955) as a tool for knowledge acquisition, and as a source of rules for a rule-based expert system. Shaw & Gaines (1983; 1986) have put great emphasis on the need to be able to validate any technique for knowledge engineering. It is not sufficient to show that reasonable expert systems can be developed. One must attempt to evaluate the accuracy and completeness of the knowledge transfer. This is not an easy task because there are few well-established domains of expertise where the accuracy of the elicitation can be tested. A study has been carried out using PLANET (Shaw, 1982) to determine whether PEGASUS can be used to elicit the Business Information Analysis Integration Technique (Carlson, 1979; Sowa, 1984) distinctions from those with some accounting or business knowledge (Shaw & Gaines, 1983). The experiments have been based on taking a spectrum of those with assumed expertise in business record keeping, from businessmen and accountants, to accounting students at various stages of their education, and using PEGASUS to elicit their constructs relating to business transactions. In the interests of an initial open-ended approach, the construing was not tightly constrained to relate to the record-keeping aspects of the transactions.

0020-7373/88/080329 + 22$03.00/0 © 1988 Academic Press Limited

Various forms of analysis were applied to the resulting grids. The constructs elicited were compared with the BIAIT constructs by analysing with ENTAIL (Gaines & Shaw, 1986; Shaw & Gaines, 1987) the compound grids obtained by merging the constructs elicited from each person with those corresponding to the BIAIT constructs. This showed that all the BIAIT constructs exist within those elicited. The set of grids obtained was also analysed with SOCIOGRIDS to see whether the variations of expertise across the group of subjects used shows up clearly. This is a good test of the basic validity of the technique and its foundations in the notion of construing. For example, in an earlier study (Shaw, 1980) of quality control for a garment company the SOCIOGRIDS analysis re-constructed the management hierarchy of the firm from the construct systems of staff involved in construing garment faults.

Evaluation of knowledge-based tools

Over 20 years ago, validation was labelled as the most elusive of all the unresolved problems associated with computer simulation techniques (Naylor, Balintfy, Burdick & Chu, 1966) and, despite work which has gone on in the intervening years, it is still an unresolved problem. The main problem with the evaluation of knowledge acquisition systems is that it is almost impossible to separate the use of them from the experts involved. A knowledge-base is produced from this interaction, and it is often the adequacy of this knowledge-base, or the number of rules produced which is measured. This, however, is still more of a test of the expert than the knowledge acquisition system. Is it a bad knowledge acquisition system if an inadequate knowledge-base is produced from an inadequate expert? Or a good one? These are all questions of *validity* in its general sense of the degree of homomorphism between a representation system and the system that it is supposed to represent (Vandierendonck, 1975).

Another factor which must be considered in the evaluation of a knowledge acquisition system is the impact of the user interface. It is very important to the quality of the resulting knowledge-base that the expert is motivated and finds it easy to use the system. It is difficult to evaluate the system without considering the user interface, hence this is an aspect which should also be studied.

Evaluation studies of knowledge-based systems (specifically expert systems) have been reported and methods for validation have been outlined. For example Siegel (1986) points out the need to evaluate the output of expert systems and outlines in detail the methods for accomplishing this task. He includes criterion variables and the methods for testing. These methods are based on the use of test cases as a developmental tool through successive improvements of the system. Test cases provide standard input to the system and output can be tested for consistency. O'Keefe, Balci & Smith (1987) discuss problems in the validation of expert systems and elaborate the distinction between validation and verification: *validation* means "building the right system" and *verification* means "building the system right".

Until the relatively recent development of knowledge acquisition tools, the evaluation of knowledge-based systems focused on the performance of the system on the problem it was designed to solve. Waterman (1986) views the iterative development of expert systems as a method to compensate for less than effective

knowledge acquisition procedures; the expert system, itself, acts as a check on the acquisition procedures. However, criticisms of this approach have been voiced (Gaines, 1987). Subjective validation of the expert system's knowledge-base and performance by the expert may provide inaccurate or biased criteria (Cleaves, 1987) or may point out only the components which do not "work" but may say little about what does work. Gaines suggests that objective performance standards be developed and knowledge-based systems and acquisition tools be evaluated against these standards. Knowledge relevancy, accuracy, consistency and completeness have been posited as objective performance criteria.

Reports of evaluation activity for automated knowledge acquisition tools usually appear only as tangential comments in papers outlining the design and development of a knowledge-based system. A notable exception is Michalski & Chilausky (1981) who compared the inductively derived rules from AQ11 with direct representations of the expert's knowledge. Other automated systems such as MORE (Kahn, Nowlan & McDermott, 1985), SALT (Marcus, 1987) and MOLE (Eshelman, Ehret, McDermott & Tan, 1987), which have been designed as automated interviewers for specialized knowledge-bases, approach knowledge-base criteria through a "successive refinement" method. Although this does not constitute a strict validation of the acquisition tool itself, the refinement process helps to build the integrity of the knowledge-base. This can be considered as one method of validating the knowledge acquisition tool.

Burton, Shadbolt, Hedgecock & Rugg (1987) have recently reported an evaluation study of a number of knowledge acquisition techniques: interview, protocol analysis, goal decomposition, and card sort in the domain of rock identification. They took 32 undergraduates as subjects and looked at time taken for elicitation, time taken to code the transcript into production rules, number of rules elicited (correct or incorrect), number of clauses constituting the elicited rules, and completeness of the rule set. They also measured the personality variables introvert or extravert, and field dependent or independent.

Some of the closest work to actual knowledge validation as well as validation of acquisition tools appears in the investigation of expertise. Johnson (1986; 1987) has established a research procedure which attempts to validate the expertise required to solve problems. Using structured interviews, verbal thinking-aloud protocols and protocol analysis based on systematic observations of problem-solving behaviour, Johnson models the operative knowledge of a task. His empirical testing of events includes the use of a panel of experts to validate results of knowledge acquisition as well as the use of test cases which are both similar and dissimilar to the original problem-solving task.

Although many knowledge acquisition systems have been designed, little has been reported on their evaluation. What is lacking is a useful framework to guide acquisition tool evaluation. The following section outlines a framework for the evaluation and validation of knowledge acquisition systems.

Framework for evaluation

In experimental psychology some "dependent" variable acts as the starting point and various factors are hypothesized which might predict that variable. In evaluation

research, the independent variable (e.g. the program, the system, the procedure) acts as the starting point and assessments are made against a set of criteria or goals (Bernstein, 1976). Evaluation research is expected to tell how well something "works": how well something does what it is supposed or conceived to do. Evaluation can also be structured to tell when, why and how something works.

In order to be effective, evaluation designs require clear explanation of components and requirements. The first is to determine the external criteria which act as comparison points for the system under evaluation. *A priori* specification of these criteria ensures rigor and clear measures of the system's performance. Establishing a useful evaluation framework also requires the recognition of various levels of operation and measurement in the external criteria. This represents a second requirement.

Specific determination and description of the input and output of the system under evaluation is a third requirement. The methods and procedures used to measure the inputs and outputs of the system require simplicity and accuracy. As well, the methods of processing inputs to outputs (e.g. the algorithms) require clarification, justification and theoretical support. This fourth requirement, together with the third, allows the developer or researcher to use evaluation results not only to establish the system's operative effectiveness but also to help guide future development and improvement of the system under evaluation.

A fifth requirement of evaluation research is the selection of populations for study. In many cases the target population has been considered in the design of the system. Samples drawn from this population for research purposes are then consistent with internal and external validity criteria (Anderson, 1975).

The final requirement reflects the number of measurements taken of the system during evaluation activity. A multiple measurement schedule built into the evaluation design ensures that replication can be strengthened and provides necessary information for development purposes and for assessing the degree to which the system meets performance criteria.

This section outlines an evaluation framework for knowledge support systems. Applying the above criteria to knowledge support systems, evaluation stages and specific evaluation tasks are outlined for the evaluation activity. General design considerations have been elucidated for knowledge support systems (Shaw & Gaines, 1987). These are also incorporated into the evaluation framework.

Figure 1 shows two distinct aspects of validation: those concerned with objective

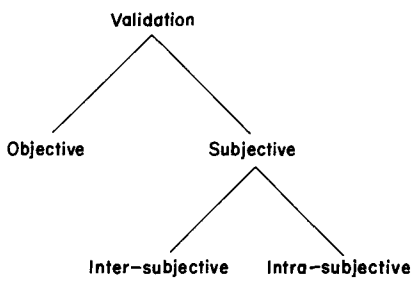

FIG. 1. A validation model.

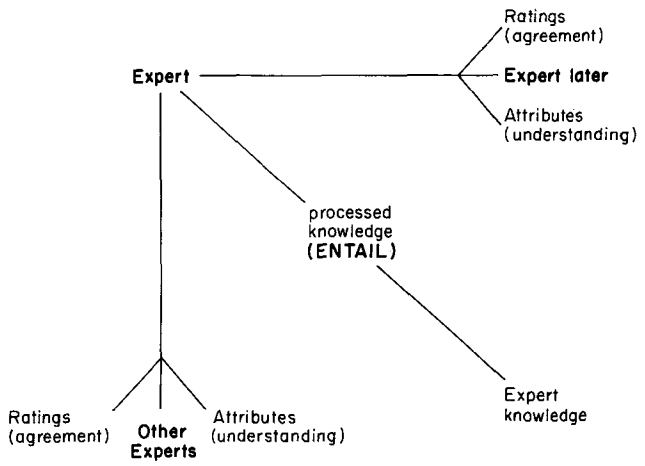

FIG. 2. The framework for the study.

aspects of the system and those concerned with subjective aspects of the system. These distinctions represent the various aspects of operation of knowledge support systems. The objective aspects are those concerning the performance of the expert system or knowledge base, elicited from the expert or experts, against objective criteria. A knowledge support system supports the knowledge processes of the expert, and allows her to develop a system which actually performs effectively in the real world, rather than merely confirming the opinion of the expert. The study done by O'Keefe *et al.* (1987) looks at objective aspects of a system.

The subjective aspects relate to the way in which the knowledge acquisition system acquires the knowledge that the expert has. It is clearly impossible to separate the knowledge acquisition tool from the expert, so the term *knowledge acquisition system* will be used to mean the knowledge acquisition tool, *and* the expert:

$$\textit{knowledge acquisition system} = \textit{knowledge acquisition tool} + \textit{expert}$$

Figure 2 shows the model developed for this study.

It seems essential to have a group of experts involved in building a knowledge base, so it was necessary to see if there was mutual consistency and/or construing across experts. Also, whether there was any consistency of an expert over a time interval, both in the terminology used, and how it was used. Finally, an investigation was made as to whether the analysis performed by the system on each expert's knowledge was seen as correct, or if any of it was incorrect.

SUBJECTIVE VALIDITY

A measure of diachronic reliability (Chow, 1987) appears most suitable for evaluating the *replication* of knowledge support systems. This measure reflects the degree of similarity, over time, of the calculated output. A knowledge support system reflects a high degree of replication in the sense that the output of the knowledge support system, the results of the algorithms, reflect a high degree of

similarity over time. Strictly speaking, this approach to replication is strongly influenced by the reliability of the user of the knowledge support system. However, the knowledge support system can be viewed as an aid to the user which helps the user maintain consistency over time given the same task demands. Also, knowledge support systems are not used in isolation but are used by individuals so any measure of a knowledge support system must be done with those who will use the system. If replication measures are low, the system cannot be considered consistent or helpful, unless this can be seen to be a product of the raw data from the experts.

Although *replication* is a possible requirement of acknowledge support system, other basic requirements need attention. The knowledge support system must also demonstrate various forms of *consistency*. In this sense, consistency refers to the extent to which the knowledge support system produces what it is expected to produce. At the subjective level of evaluation, consistency is defined by two measures. The first is *intra-subjective consistency*. Based on the concepts of content and face validity (Travers, 1969; Campbell & Stanley, 1963), this type of consistency is defined as the extent to which the knowledge support system output is correct as judged by the expert. Knowledge support systems produce output which has a high degree of similarity with what is apparently expected. The absence of this form of consistency indicates that the system may be performing its calculation incorrectly and/or that the output bears little resemblance to what would be expected by the designer, by the user of the tool or by the domain expert. This result has serious implications for any subsequent use of knowledge support system output.

Another, complementary, measure of consistency at this level of evaluation is that of *intersubjective consistency*. Based on the concept of *concurrent validity* (Chow, 1987; Travers, 1969) this type of consistency is a measure of prediction of events which occur at roughly the same time. If two or more individuals, with similar experience in an area, produce similar output using a knowledge support system, then the system can be said to afford a degree of intersubjective consistency. The extent to which two or more people produce the same output represents the degree of consistency.

However, inconsistency among experts is not unexpected as experts may well not agree. In many areas, each expert has a unique perspective on the topic and thinks that there is only one way that the topic should be considered. This is not necessarily an undesirable situation, as long as the system is able to reflect this. If a task is carefully defined, specifically outlined and agreed upon as important, then a certain degree of similarity can be expected in the outcome of the knowledge support system. This suggests the selection and development of a set of test cases with specific task requirements. These test cases can then be used to determine the degree of inter-subjective consistency demonstrated by various experts, while using the system. In the same way as for replication, the knowledge support system can be seen as an aid to users.

OBJECTIVE VALIDITY

Once the knowledge support system has demonstrated acceptable levels of replication and intra- and inter-subjective consistency, it can be evaluated against more objective criteria. *Objective validity* refers to the degree to which the output of

a knowledge support system agrees with well established facts in the domain. This can be accomplished by measuring the degree of convergence of knowledge support system output with established facts from reputable experts and or from reputable printed material. Also, a measure of the degree of divergence of knowledge support system output from known non-truths can be ascertained: the degree to which the knowledge support system output is dissimilar from known inaccuracies. Valid knowledge support system output demonstrates a high degree of similarity (convergence) with established domain facts and a high degree of dissimilarity (divergence) from incorrect statements.

Another form of validity at this level concerns the usefulness of the output of the knowledge support system. *Operative validity* is defined as that knowledge which is necessary for accomplishing a task or class of tasks (Johnson, 1986). By definition, operative knowledge is that body of information which is necessary and sufficient to meet task demands. It suggests that any knowledge which is unnecessary for the task is considered tangential and/or superfluous and should be eliminated from consideration. Only the knowledge which is used to solve the problem need be considered. The degree to which a knowledge support system is able to reduce the amount of unnecessary or superfluous output represents a measure of operative validity.

Once the knowledge support system has demonstrated acceptable degrees of objective and operative validity, its functional role attributes require assessment. In other words, the knowledge support system does what it is supposed to do but in order to be used effectively or for its use to be generalized, constraints and caveats on its use need elucidation and study. The *usability* (Gaines, 1987) of the system requires assessment. This refers to the ease of use and the understandability of the knowledge acquisition tool by the expert. The knowledge support system requires testing using a variety of task demands apart from those used in the last level of evaluation. Measures of the degree of knowledge support system replication and consistency on similar task demands and on dissimilar task demands begins to indicate the limits of the system's usefulness, effectiveness and areas of operation.

An outline of knowledge support system zero

Part of the interaction with one of the geographers will be used to give a flavour of Knowledge Support System Zero (KSS0). KSS0 is a Macintosh program with similar facilities to KITTEN which was described in some detail by Shaw & Gaines (1987).

Figure 3 shows the common entities agreed on by all the geographers, and is the start of expert 2's G3 as described in Fig. 10. She then selects *Triple,* and three techniques are shown as in Fig. 4, where she is asked to click the mouse on the one which is different from the other two.

She clicks on *Probability mapping* and elicits the attribute *local–global* which is shown with all the techniques on the scale in Fig. 5.

Figure 6 shows how the techniques are dragged on to the attribute bar from the left. Both the attribute labels and the techniques can be edited any time, and also the placing of the techniques.

Figure 7 shows two matched attributes, with both sets of ratings shown, so that

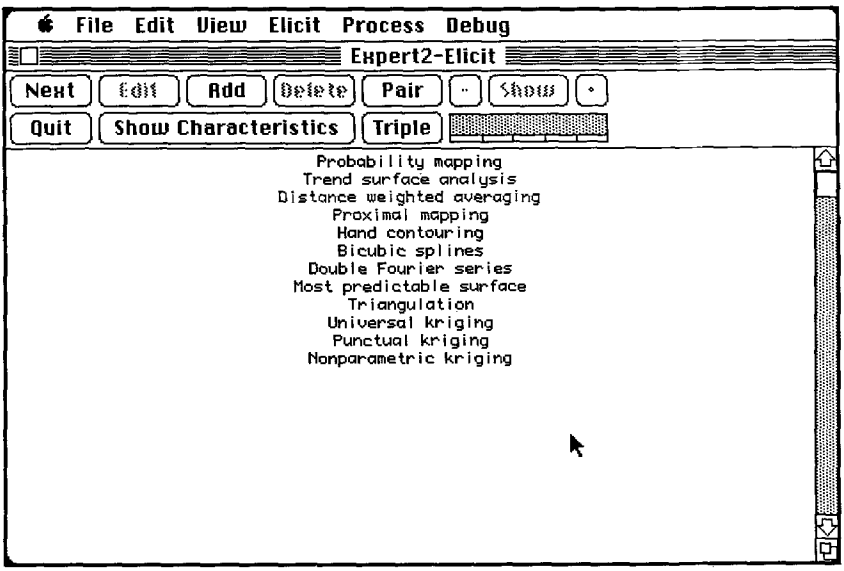

FIG. 3. KSS0 screen showing the entities used by the geographers.

they can be compared. Again, any one is moveable if the expert wants to adjust anything.

Techniques may also be matched, as shown in the screen in Fig. 8. The technique at the top has a marker on each attribute sticking up, and the one at the bottom a marker sticking down. Again, any marker is moveable if the expert wants to change anything.

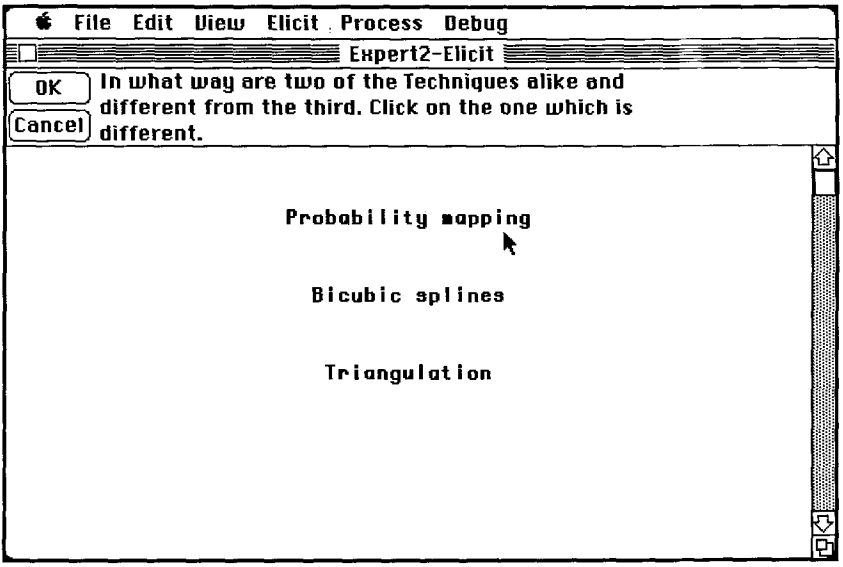

FIG. 4. KSS0 attribute elicitation from a triple.

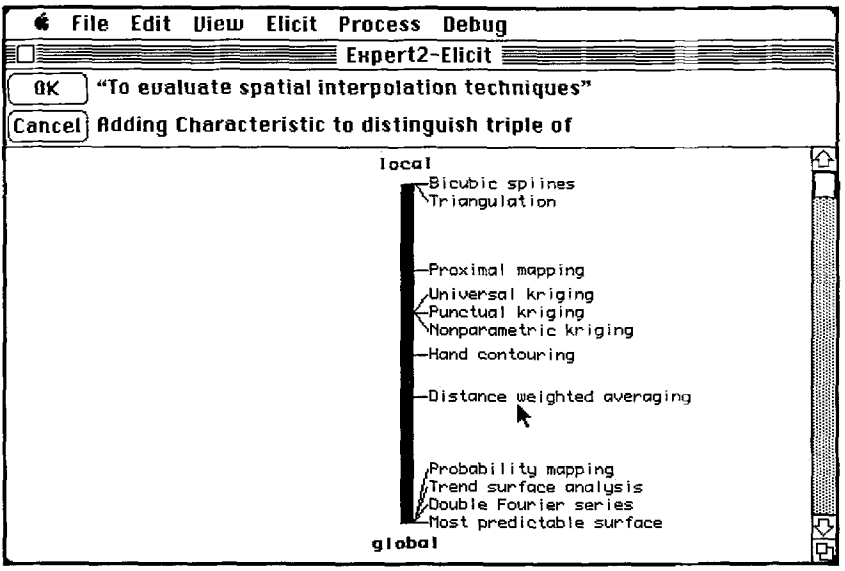

FIG. 5. KSS0 attribute screen.

Figure 9 shows the FOCUS clustering of the final grid. As can be seen on the attribute tree at the top right, there are three clusters at 75%, and one attribute which is not part of any cluster. From the bottom right tree of techniques, it can be seen that the one most different from the others is *hand contouring*. Many other conclusions can be drawn from this diagram, but it is sufficient here just to give a flavour of what the expert experienced in using KSS0.

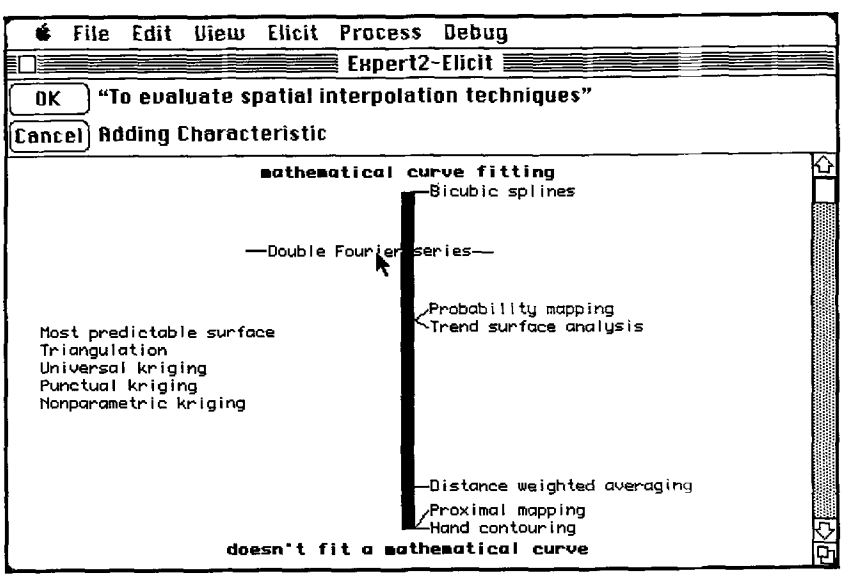

FIG. 6. KSS0 showing click and drag elicitation.

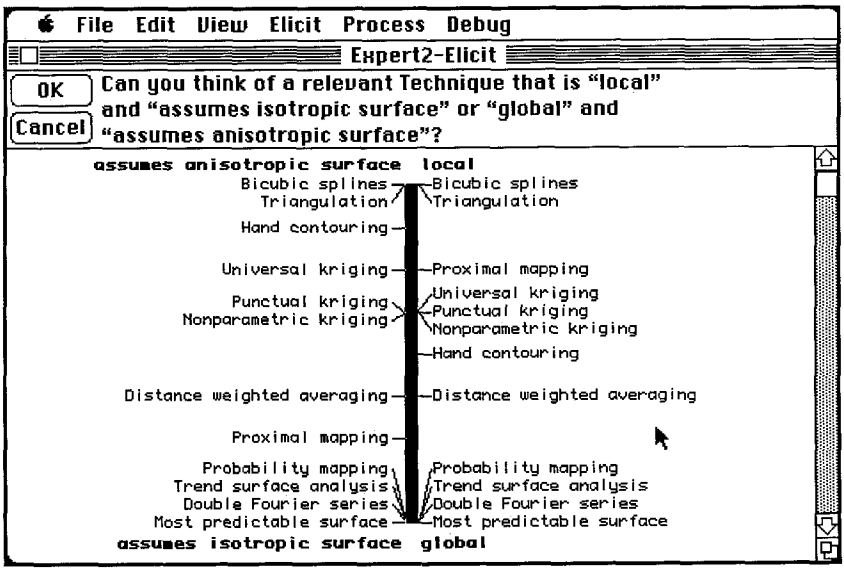

FIG. 7. KSS0 showing matching attributes.

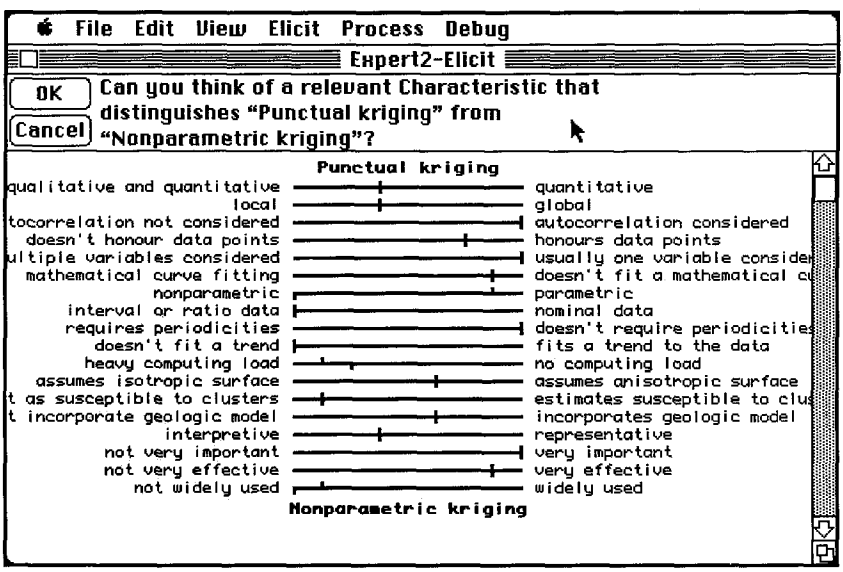

FIG. 8. KSS0 showing matching entities.

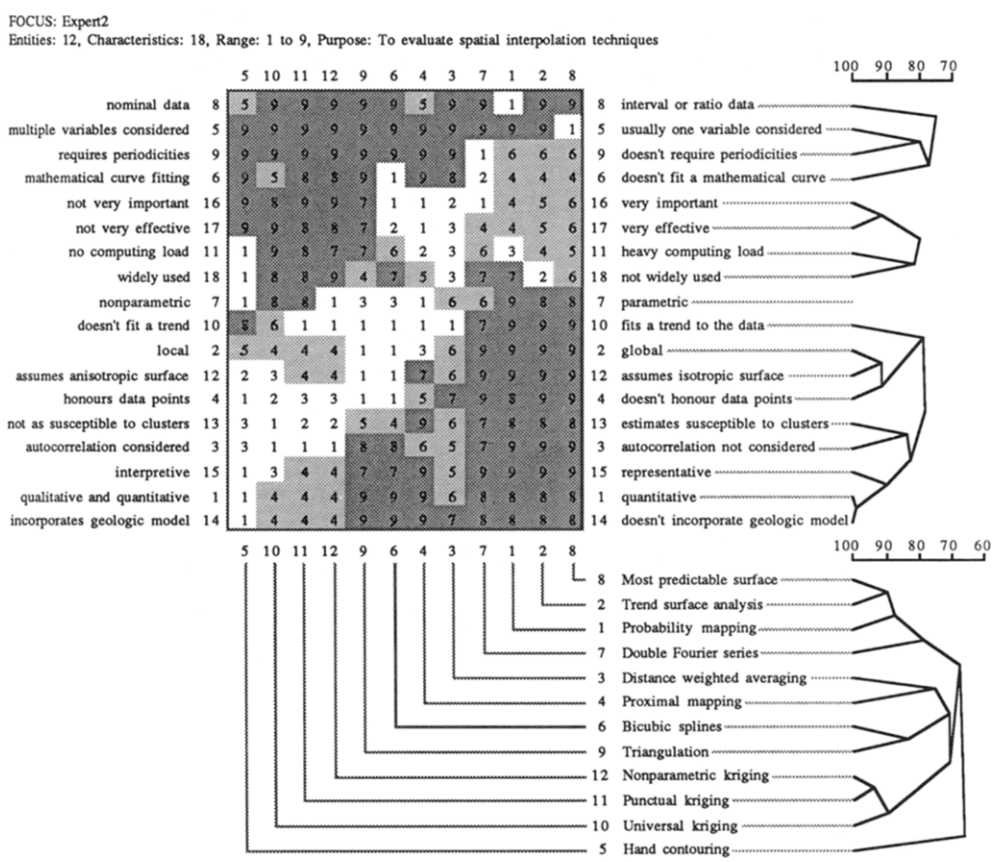

FIG. 9. KSS0 FOCUS display of completed grid.

Methodology

Figure 10 outlines the procedural steps and the data collection points used in the subjective evaluation of KSS0.

The figure shows the first step as procedure 1 (P1) which introduced the subjects to the requirements of KSS0. At this time a specific task was agreed upon and a purpose for the grid elicitation was developed. The first data collection point (G1) consisted of the elicitation of a grid by each subject individually. This provided an initial set of entities and attributes particular to each subject. The second collection point (G2) consisted of the subjects *exchanging* their first grids (entities and attributes) to produce other grids. The next step was a procedure (P2) to have the subjects agree on a common set of entities. This common entity set provided the basis for the next data collection point (G3) which consisted of the elicitation of a grid by each expert using common entities. This elicitation was done, as the other, separately for each expert. P3 represents a wait of over two months before the next data collection point (G4). At this time a repeat of the activity at G3 was completed using a common entity set but the regeneration of attributes. For G5 the subjects

Event	Description
P1	determine task demands and elicitation purpose
G1	grid elicitation using own entity set
G2	grid eliciation using other expert's entities and attributes
P2	experts discuss and produce a common entity set
G3	grid elicitation using common entity set
P3	waiting period: 2 to 3 months
G4	grid elicitation using G3 entity set
G5	grid elicitation using entities and attributes from G3
P4	obtain subjective ratings from entailment output from G3
	Analysis

FIG. 10. Evaluation design event table.

were given the entity and attribute set from G3 and asked to re-rate the entities. The final data collection point (P4) consisted of giving the subjects a selected list of entailments using the ENTAIL algorithm (Gaines & Shaw, 1986) generated from the G3 grid. A set of examples from four separate levels of significance of entailment were selected and randomly presented to subjects who were asked to rate each entailment on a four point scale from correct to incorrect.

THE STUDIES

Two separate studies were completed using specific task demands from two separate and quite different domains. These two studies show the different *types* of expertise used. In the first study, the experts were researchers, and continually re-structured and re-construed their knowledge as a result of feedback from KSS0. This demonstrates *leading edge* knowledge, where experts are brought difficult cases and situations which lead to a reconstruction of their expertise. It is, moreover, an area where there are strongly held opinions and much disagreement. In the second study, the experts were professionals working daily with their knowledge. In this case they have fixed expert knowledge which is common to all experts in this field. Furthermore, one of the experts, a trainer, commonly helps others to structure their knowledge of the field.

The first study involved three researchers in a university geography department each of whom is an expert working in the field of spatial interpolation techniques to produce contour maps. Each was interviewed separately (P1) and was involved independently for an hour or more in the first session on the task of "evaluating spatial mapping techniques" (G1). In the second session, on the following day, each expert exchanged his or her attributes with the others. This involved expert 1 in seeing expert 2's entities and attributes without the ratings, and rating them as he

would have done (G2a). He repeated this for expert 3's grid (G2b). This was done for each expert in turn.

Following this session, all three experts met together to discuss the entities each had used, and to agree both on the meanings of each, and on a common set that all could use. There were twelve of these decided after 30–40 min. In the third session, in the same week, each expert elicited another set of attributes which could include the previous ones elicited, as well as any he or she had seen when doing the exchange procedure. This was G3. Three attributes were added at this stage by the experimenter: *very important–not very important; very effective–not very effective;* and *widely used–not widely used.* Each expert also rated the attributes for importance on a 1–10 scale: 1 most important, to 10 least important. After this session, the experts were asked not to think about the topic more than they usually would, not to discuss the topic with each other, and not to review their attributes before the next session.

In the fourth session, 12 weeks later, each expert in turn elicited another set of attributes using the agreed set of entities from session three producing G4. Immediately this was completed, in about an hour, each expert was shown his entities and attributes from session three, that is, G3 with the ratings removed, and asked to re-rate these on the same scale. This gave G5.

The entailments (Gaines & Shaw, 1986; Shaw & Gaines, 1987) were produced for each expert from G3. These rules were then sampled from each category of significance, $\geq 0{\cdot}15$, 0·14 to 0·10, and 0·09 to 0·05 and randomly presented to the expert (P4) who was asked to rate them on a scale 1—correct, 2—likely to be correct, 3—not likely to be correct, and 4—incorrect.

In the first session, two of the experts used a 5-point scale, and one a 9-point. When agreeing on their common set of entities, they also agreed to use a 9-point scale. Although this seems cumbersome when using numbers it is, in fact, a finer discrimination of a 5-point scale, having all the in-between points available for finer distinctions. All three experts agreed that the system was easy to use, allowed easy correction of errors, and gave them the options that they needed to peruse the data.

In the second study, a trainer and an engineer in the field of gas pipeline valve maintenance provided their expertise. These two individuals were chosen because they were both considered knowledgeable in the area, but also because they brought a different background to the field: one from a field-operator perspective, the other from that of an engineer. The trainer had been involved in training gas pipeline operators in proper maintenance of pipeline valves for 15 years and the engineer owned a company which sold valves. Each was interviewed separately (P1) and a diagram of the area was developed. Common to each diagram was a crucial step of determining the type and make of the valve. This step was necessary because knowledge of the valve type and manufacturer gave important clues to the operator concerning proper maintenance. The task demand was defined as "the determination of valve characteristics". Each subject completed the design requirements except for P4.

THE MEASURES: CONSISTENCY AND CONSTRUING

Two measures were used in this study, *consistency-with* another and *construed-by* another. *Consistency* is measured using *exchange grids.* Do the experts see the topic

in a similar way at the same time? By exchanging entities and attributes they are able to view the topic from the perspective of the other. To do this involves an understanding of what the other's entities and attributes actually *mean* in the domain, or *can I construct a point of view which makes sense of them*? Not only that, but *do I agree with the other's construction of the topic, having understood the perspective of the other*? So *consistency* is defined as the degree of match between one expert's ratings of his/her own entities and attributes and another expert's ratings of the first expert's entities and attributes. The patterning of the entities on a particular attribute is matched against the patterning of the same entities on that same attribute in the other grid.

Consistency over time can also be measured in this way but using the *same* expert's grid on the two occasions with the same entity and attribute labels as before, but new ratings. Only entities and attributes with the same labels in both grids can contribute to this measure of consistency. Operationally, this measure was calculated using the EXCHANGE measure based on the matching algorithm that is described in Shaw (1980) and used in FOCUS, CORE and SOCIOGRIDS. It ranges from 100 for perfect match to 0 for maximum reversal. This measure is picking out the differing use of terminology in the two grids. The following formula was used:

$$\text{G consistency-with G}' \text{ at criterion} = 100 \times \frac{\text{(number of attributes in G matched greater than criterion by } \textit{same} \text{ attributes in G}')}{\text{(number of attributes in G)}}$$

Construing is akin to consistency. The entity labels must be the same in both grids, but separate sets of attributes can be used. In this case, the measure picks out the attribute in the second grid with the best match to the patterning of the entities on the one being considered in the first grid. In this case, the same attribute may be chosen more than once in the second grid if it matches more closely than any other to more than one attribute in the first. Again, the same matching algorithm was used. This measure is picking out the best match between the attributes in the two grids, regardless of terminology. The following formula was used:

$$\text{G construed-by G}' \text{ at criterion} = 100 \times \frac{\text{(number of attributes in G matched greater than criterion by } \textit{any} \text{ attributes in G}')}{\text{(number of attributes in G)}}$$

INTER- AND INTRA-SUBJECTIVE AGREEMENT AND UNDERSTANDING

Operational definitions for the various replication and consistency measures were used to evaluate the knowledge support system. *Intra-subjective understanding,* U_t, was defined as the degree of construing of the earlier grid G3 by the later one G4. Those attributes which were matched on the G3 and G4 grids at a criterion of 80 were used for this calculation, and this was calculated for each expert:

$$U_t = \text{G3 construed-by G4 at 80}$$

Intra-subjective agreement, A_t, was defined as the degree of consistency between one expert's ratings of his/her own entities and attributes in G3 with his/her ratings

of the entities and attributes in the later grid G5. The following formula was used to calculate A_t for each expert:

$$A_t = \text{G3 consistency-with G5 at 80}$$

Inter-subjective understanding, U_i, was defined as the degree of construing of each expert's grid G3 by another expert's G3:

$$U_i = \text{G3E construed-by G3E' at 80}$$

Inter-subjective agreement, A_i, was defined as the degree of consistency between one expert's ratings of his/her own entities and attributes in G1 with another expert's ratings of the first expert's entities and attributes in G2. The following formula was used:

$$A_i = \text{G1E consistency-with G2E' at 80}$$

(N.B. In study 1 there were two G2s: G2a and G2b, so two measures resulted for each expert.)

Finally, *intra-subjective perspective consistency,* C_i, was defined as the degree of consistency of the rules produced by the ENTAIL algorithm with what the expert apparently expected. The ratings of these statements at P4 were then compared to the entailment output from G3. The number of identical ratings and the number of similar ratings (differing by only one value) were selected. The following formula was used to determine C_i:

$$C = \frac{(\text{number of same ratings} + \text{number of similar ratings}) \times 100}{(\text{total number of rules given})}$$

Results

Figures 11 to 18 display results from the two studies. The *intra-subjective understanding* scores, U_t, displayed in Fig. 11 ranged from 63 to 86 in the first study, and from 83 to 100 in the second. This difference in values between the two studies may reflect the fact that the three experts in the first study are academic researchers

Study 1	$U_t \geq 80$	$A_t \geq 80$
E1	62.5	81.2
E2	77.8	94.4
E3	85.7	78.6
Study 2		
E1	100	66.7
E2	83.3	66.7

FIG. 11. Intra-subjective agreement (A_t) and understanding (U_t) results.

Expert Pairs	Ui ≥ 80	Ai ≥ 80
Study 1		
E1, E2	62.5	33.3
E2, E1	61.1	26.7
E1, E3	31.2	8.3
E3, E1	42.9	33.3
E2, E3	44.4	20.0
E3, E2	71.4	33.3
Study 2		
E1, E2	100	12.5
E2, E1	66.7	100

FIG. 12. Inter-subjective agreement (A_i) and understanding (U_i) results.

at a university, and hence tended to be more provisional, exploratory and accepting of different ways of considering the problem. They were at no time fixed in their views of the topic. In the second study, however, the experts were professionals using their expertise in the field, and hence much more decisive and sure of what they were doing and how it should be done. The *intra-subjective agreement* scores, A_t, also displayed in Fig. 11 ranged in the first study from 79 to 94, indicating that the experts were using their terminology in much the same way from one occasion to the other and that KSS0 was able to reflect this consistency. In the second study both were 67, indicating some difference in the use of terminology.

Inter-subjective agreement scores, A_i, are displayed in Fig. 12. That is the extent to which the experts agree, range in the first study from 8 to 33. This shows that the experts disagree in their terminology and in how they view the topic quite extensively.

This can be seen in Fig. 13 where the details of the analysis are shown. E1 has exchanged his techniques and attributes with the other two, E2 and E3. The listing at the top shows the agreement for the pair E1, E2. The highest matched attribute is *interval data–nominal data,* and four attributes are agreed upon over the threshold of 80. In the other listing for E1, E3 the highest matched attribute was also *interval data–nominal data* but was the only one matched above the threshold of 80.

The raw grids for E1 and E2 are shown in Figs 14 and 15, using a 5-point scale, so that the actual disagreements can be seen. For example, the first number on the first line is a 5 in E1's grid and a 1 in E2's. This means that E2 thinks that *probability mapping requires model,* whereas E1 thinks that *probability mapping requires NO model.* On inquiring about this, the explanation given was couched in terms of what one actually means by the term *model,* indicating the difference in terminology which was shown up in the measures of A_i.

In the second study the scores range from 13 to 100. Here it is interesting to note that the trainer could use the terminology of the engineer totally, but the engineer could only use a small part of the trainer's terminology.

Inter-subjective understanding scores, U_i, are also shown in Fig. 12. Here the values are much higher showing that the experts have similar distinctions about the topic even though they differ greatly in how they use the terminology. In the first study these values range from 31 to 71, and in the second from 67 to 100. Figure 16

```
Grid  Ent Att
* E1:  11  12  Expert1
  E2:  11  12  Expert2
  E3:  11  12  Expert3
       11  12  Common Entities & Common Attributes


E1:E2 33.3% over 80.0 (Expert1 attribute-consistency-with Expert2)

  1:   8.3% > 90.9   A2: Interval data - Nominal data
  2:  16.7% > 81.8   A4: Global - Local
  3:  25.0% > 81.8   A5: Intuitive - Mathematical
  4:  33.3% > 81.8   A6: Requires spatial search - Does not require spatial search
  5:  41.7% > 75.0  A10: Difficult to understand - Easily understood
  6:  50.0% > 72.7   A3: Non-polynomial - Polynomial
  7:  58.3% > 72.7   A7: Discontinuous - Continuous
  8:  66.7% > 59.1  A12: Does not consider non-spatial attributes - Considers non-spatial attributes
  9:  75.0% > 56.8  A11: Few points - Many points
 10:  83.3% > 50.0   A8: Does not honour data - Honours data
 11:  91.7% > 47.7   A1: Requires no model - Requires model
 12: 100.0% > 47.7   A9: Linear interpolation - Non-linear interpolation


E1:E3 8.3% over 80.0 (Expert1 attribute-consistency-with Expert3)

  1:   8.3% > 86.4   A2: Interval data - Nominal data
  2:  16.7% > 79.5   A5: Intuitive - Mathematical
  3:  25.0% > 79.5   A7: Discontinuous - Continuous
  4:  33.3% > 77.3   A4: Global - Local
  5:  41.7% > 75.0   A3: Non-polynomial - Polynomial
  6:  50.0% > 68.2  A11: Few points - Many points
  7:  58.3% > 65.9   A1: Requires no model - Requires model
  8:  66.7% > 63.6   A9: Linear interpolation - Non-linear interpolation
  9:  75.0% > 63.6  A10: Difficult to understand - Easily understood
 10:  83.3% > 56.8   A8: Does not honour data - Honours data
 11:  91.7% > 56.8  A12: Does not consider non-spatial attributes - Considers non-spatial attributes
 12: 100.0% > 52.3   A6: Requires spatial search - Does not require spatial search
```

FIG. 13. Analysis of inter-subjective agreement.

FOCUS: Expert1
Entities: 11, Attributes: 12, Range: 1 to 5, Purpose: To evaluate spatial interpolation techniques

		1	2	3	4	5	6	7	8	9	10	11		
Requires no model	1	5	5	4	1	3	1	4	4	5	2	5	1	Requires model
Interval data	2	5	1	1	1	4	4	1	1	1	1	1	2	Nominal data
Non-polynomial	3	5	5	1	1	1	1	5	1	5	1	5	3	Polynomial
Global	4	1	1	3	4	4	4	5	2	1	4	1	4	Local
Intuitive	5	4	4	5	3	2	1	5	4	5	3	5	5	Mathematical
Requires spatial search	6	5	5	2	1	2	3	5	5	5	3	5	6	Does not require spatial search
Discontinuous	7	5	5	4	2	1	3	5	5	5	5	5	7	Continuous
Does not honour data	8	5	5	2	3	1	2	4	4	5	2	1	8	Honours data
Linear interpolation	9	5	5	2	2	3	1	5	5	5	5	5	9	Non-linear interpolation
Difficult to understand	10	2	4	1	4	4	5	1	2	1	4	1	10	Easily understood
Few points	11	1	3	1	5	3	1	3	2	1	4	2	11	Many points
Does not consider non-spatial attributes	12	2	2	2	2	3	3	1	1	2	1	5	12	Considers non-spatial attributes
		1	2	3	4	5	6	7	8	9	10	11		

Probability mapping
Trend surface analysis
Kriging
Distance weighted averaging
Proximal mapping
Hand contouring
Bicubic splines
Double Fourier series
Most predictable surface
Negative exponential surface
Vector trend surface analysis

FIG. 14. Display of expert 1's first grid G1 from first study.

FOCUS: Expert2
Entities: 11, Attributes: 12, Range: 1 to 5, Purpose: To evaluate spatial interpolation techniques

		1	2	3	4	5	6	7	8	9	10	11		
Requires no model	1	1	3	4	2	1	5	1	2	3	3	3	1	Requires model
Interval data	2	5	1	1	1	5	1	1	1	1	1	1	2	Nominal data
Non-polynomial	3	1	5	2	2	1	1	4	3	5	4	5	3	Polynomial
Global	4	4	1	3	3	4	3	5	1	1	2	1	4	Local
Intuitive	5	3	4	3	3	2	1	5	5	4	5	4	5	Mathematical
Requires spatial search	6	4	5	1	2	3	3	3	5	5	5	5	6	Does not require spatial search
Discontinuous	7	3	5	3	3	4	1	2	5	5	5	5	7	Continuous
Does not honour data	8	1	1	3	2	2	5	4	1	1	1	1	8	Honours data
Linear interpolation	9	2	3	5	4	1	1	2	3	3	3	3	9	Non-linear interpolation
Difficult to understand	10	3	3	1	4	5	1	2	2	2	3	2	10	Easily understood
Few points	11	3	4	3	3	5	3	3	4	4	3	4	11	Many points
Does not consider non-spatial attributes	12	4	3	5	5	4	5	1	2	3	5	5	12	Considers non-spatial attributes

1	2	3	4	5	6	7	8	9	10	11
Probability mapping	Trend surface analysis	Kriging	Distance weighted averaging	Proximal mapping	Hand contouring	Bicubic splines	Double Fourier series	Most predictable surface	Negative exponential surface	Vector trend surface analysis

FIG. 15. Display of expert 2's second grid G2 exchanged with expert 1.

shows the matches which were found for the pair E1, E2—the first attribute coming from E1 and the second being the one from E2 which matched the highest *patterning* of the techniques on the first attribute. This was repeated for each of E1's attributes.

Except for the agreement of 100 in the second study, these ratings are remarkably similar. This suggests that KSS0 shows a facility for allowing experts to compare their understanding and to determine a level of consistency. Whether the threshold criteria are high enough to produce the required level of knowledge-base performance or too high to reach the criterion of completeness is a matter for further study.

Intra-subjective perspective consistency sores C_i are displayed in Fig. 17. These show whether the expert finds the rules meaningful, and rates the rules as correct or significant in the same way as the ENTAIL (Gaines & Shaw, 1986; Shaw & Gaines, 1987) algorithm. They range in study 1 from 82 to 92, showing a high level of expected rules appearing in KSS0 output. This indicates a good basis for using the ENTAIL produced rules as input to an expert system shell. A few of these rules from expert E1 are shown in Fig. 18.

CONCLUSIONS AND FUTURE WORK

In knowledge support system evaluation, internal and external problems require attention. Internal factors refer to aspects of the design, its procedures, measures and observations. External factors include aspects which are external to the design but which are present in the experimental environment.

```
Grid  Ent Att
* E1:  12  16  G3
  E2:  12  18  G3
       12  34  Common Entities & Total Attributes

E1<:E2 62.5% over 80.0 (E1G3 attribute-construed-by E2G3)

 1:   6.2% > 88.5    E1A2: Local - Global
                     E2A2: local - global
 2:  12.5% > 87.5    E1A3: Low level data - High level data
                     E2A8:   nominal data - interval or ratio data
 3:  18.8% > 86.5    E1A1: Does not honour data points - Honours data points
                     E2A4:  doesn't honour data points - honours data points
 4:  25.0% > 86.5    E1A7: Short distance autocorrelation - Long distance autocorrelation
                     E2A2:                                 local - global
 5:  31.2% > 86.5    E1A9: New geographical technique - Old geographical technique
                    E2A18:            not widely used - widely used
 6:  37.5% > 85.4   E1A16: Not widely used - Widely used
                    E2A18: not widely used - widely used
 7:  43.8% > 83.3    E1A5: Discontinuous - Continuous
                     E2A2:          local - global
 8:  50.0% > 82.3    E1A4: Mathematically complex - Mathematically simple
                    E2A11:   heavy computing load - no computing load
 9:  56.2% > 81.2   E1A10:   Hard to adapt to multivariate - Easy to adapt to multivariate
                     E2A5: usually one variable considered - multiple variables considered
10:  62.5% > 81.2   E1A12:   Does not require spatial search - Requires spatial search
                    E2A13: estimates susceptible to clusters - not as susceptible to clusters
11:  68.8% > 79.2    E1A6: Does not require a priori model - Requires a priori model
                     E2A2:                                  local - global
12:  75.0% > 79.2   E1A11:      Few points - Many points
                    E2A18: not widely used - widely used
13:  81.2% > 76.0   E1A13:          Does not use polynomial - Uses polynomial
                     E2A6: doesn't fit a mathematical curve - mathematical curve fitting
14:  87.5% > 76.0   E1A15: Not very effective - Very effective
                    E2A17: not very effective - very effective
15:  93.8% > 72.9   E1A14: Not very important - Very important
                    E2A18:    not widely used - widely used
16: 100.0% > 71.9    E1A8:        Models the stationarity - Assumes stationarity
                     E2A3: autocorrelation not considered - autocorrelation considered
```

FIG. 16. Analysis of inter-subjective understanding.

The variability of the subjects presents a constant internal problem in evaluation studies of knowledge support systems. The variance of the outcome measures due to the user is explicably tied to and requires isolation from the variance due to the knowledge support system. Procedures and measures which accomplish this are required for an accurate assessment of the knowledge support system.

Evaluation studies require careful selection of subjects and selection of task demands. Subjects are likely to present different levels of knowledge and experience when faced with the same task. The effects of this difference on outcome measures are not easily determined.

Study 1	Ci
E1	88.2
E2	91.9
E3	82.2

FIG. 17. Intra-subjective perspective consistency results (C_i) from first study.

Entailments (Cutoff 0.15)

Local *implies* Short distance autocorrelation
Long distance autocorrelation *implies* Global
Global *implies* Long distance autocorrelation
Requires a priori model *implies* Long distance autocorrelation
Short distance autocorrelation *implies* Local
Mathematically simple *implies* Old geographical technique
Global *implies* Does not honour data points
Requires a priori model *implies* Global
Short distance autocorrelation *implies* Does not require a priori model
Very important *implies* Widely used
Honours data points *implies* Local
Local *implies* Does not require a priori model
Honours data points *implies* Short distance autocorrelation
Long distance autocorrelation *implies* Does not honour data points
Not widely used *implies* Not very important
Does not honour data points *implies* Global
Continuous *implies* Does not honour data points
Old geographical technique *implies* Mathematically simple
Does not honour data points *implies* Continuous
Requires a priori model *implies* Does not honour data points
Does not require a priori model *implies* Local
Does not require a priori model *implies* Short distance autocorrelation
Does not require a priori model *implies* Many points
Discontinuous *implies* Honours data points
Honours data points *implies* Does not require a priori model
Does not honour data points *implies* Long distance autocorrelation
Global *implies* Continuous
Local *implies* Many points
New geographical technique *implies* Mathematically complex
Not very important *implies* Not widely used
Widely used *implies* Very important

FIG. 18. Some rules produced by the ENTAIL algorithm from expert 1.

The extent to which training and/or prior exposure to the knowledge support system affects measured outcomes is also a factor in evaluation. To what extent did prior training alter the subject's approach to the knowledge support system? This issue also suggests another: the effect due to growing familiarity with the knowledge support system as the experiment unfolds. Subjects in the second study tended to "regress to the extreme" as they produced their grids. They produced fewer intermediate ratings at G5 than at G1 or G2. This may indicate "refinement" of the ratings or a phenomenon which occurs from the continued use of the same information for the same task demand.

Novelty effects and situational effects also require identification, elucidation and, ultimately, control. What influence did the novelty of using KSS0 have on the subjects and subsequent measures? Did the amount of guidance given while the subjects worked on the system affect the measurements? Did the subjects have the right amount of time to complete the elicitations?

To what extent were the measured outputs of intra-subjective agreement and understanding affected due to the development of new understanding by the subject from one time to another? Lower scores may indicate a "new" or "better" understanding of the material. This development may also be attributed to the demands of the experiment.

Measures have been put forward which tease out differences in terminology between experts, and show the content of this difference. To a knowledge engineer, it seems that if experts talk about their topic of expertise in the same terms then they mean the same thing, and if they talk in different terms they mean different things. However, it can be seen from the results of this study that the knowledge engineer, or even another expert, may be mistaken. The experts in the first study were clearly accustomed to disagreeing on terminology, a fact which may not always be made explicit. The consistency measure used for inter-subjective agreement showed up the divergence in use of terminology between the experts (8–33), but when used for one expert alone for intra-subjective agreement gave much higher values (79–94). This measure, then, shows a good range of discrimination in the two cases within experts and between experts. This is not so apparent in the second study where the experts were professionals using their skills and knowledge everyday in an industrial situation. A similar rationale can be applied in the construed-by measures of inter- and intra-subjective understanding.

In the first study there was clearly inconsistency among the experts, and this was reflected by the results from the knowledge support system. This was therefore a good population for study, as it highlighted a controversy and/or disagreement which would be likely to occur between experts and users of an expert system, even if it did not occur among the experts themselves.

There are many more ways of processing the data, especially in terms of the rules derived from the various grids, and this is being done. The preliminary results of the intra-subjective perspective consistency values show that the rules produced by ENTAIL are those largely expected by the particular expert concerned.

One of the other directions in which we are continuing this study, is to look into the objective validation of knowledge support system. That is, to analyse the text of some research papers on this topic, using TEXAN. A comment made several times by more than one of the experts in the first study was that although this topic is taught in an undergraduate geography course, there is no textbook which covers the material in the depth that they were forced to think about it, even after the first session.

Financial assistance for this work has been made available by the Natural Sciences and Engineering Research Council of Canada. The KITTEN and KSS0 systems are made available by the Centre for Person Computer Studies. We would like to thank our colleague Brian Gaines for discussion of this work, and for access to KSS0. We would also like to thank Robin Edmond, Karen Dutton-Marion, Alvin Simms, Frank Sparks and Nigel Waters for sharing their expertise with us.

References

ANDERSON, A. (1975). Policy experiments: selected analytical issues. *Soc. Methods and Research,* **4,** 1.

BERNSTEIN, I. N. (1976). *Validity Issues in Evaluation Research.* Beverly Hills, CA: Sage.

BURTON, A. M., SHADBOLT, N. R., HEDGECOCK, A. P. & RUGG, G. (1987). A formal evaluation of knowledge elicitation techniques for expert systems. *Proceedings of the first European Workshop on Knowledge Acquisition for Knowledge-Based Systems.* Reading, UK.

CAMPBELL, D. T. & STANLEY, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

CARLSON, W. M. (1979). The new horizon in business information analysis. *Data Base,* **10**(4), 3–9.

CHOW, S. L. (1987). *Experimental Psychology: Rationale, Procedures and Issues.* Calgary: Detselig Enterprises.

CLEAVES, D. A. (1987). Cognitive biases and corrective techniques: proposals for improving elicitation procedures for knowledge-based systems. *International Journal of Man–Machine Studies,* **27,** 155–166.

ESHELMAN, L., EHRET, D., MCDERMOTT, J. & TAN, M. (1987). MOLE: a tenacious knowledge acquisition tool. *International Journal of Man–Machine Studies,* **26,** 41–54.

GAINES, B. R. (1987). Rapid prototyping for expert systems. In: M. OLIFF, Ed., *Proceedings of International Conference on Expert Systems and the Leading Edge in Productions, Planning and Control,* pp. 213–214. University of South Carolina.

GAINES, B. R. & SHAW, M. L. G. (1986). Induction of inference rules for expert systems. *Fuzzy Sets and Systems,* **18**(3), 315–328.

JOHNSON, P. E. (1986). Cognitive models of expertise. *Symposium on Expert Systems and Auditor Judgment,* University of Southern California.

JOHNSON, P. E., ZAULKERNAN, I. & GARBER, S. (1987). Specification of expertise. *International Journal of Man–Machine Studies,* **26,** 161–181.

KAHN, G. NOWLAN, S. & MCDERMOTT, J. (1985). MORE: an intelligent knowledge acquisition tool. *Proceedings of the Ninth Joint Conference on Artificial Intelligence,* Los Angeles, (CA) (August).

KELLY, G. A. (1955). *The Psychology of Personal Constructs.* New York: Norton.

MARCUS, S. (1987). Taking backtracking with a grain of SALT. *International Journal of Man–Machine Studies,* **26,** 383–398.

MICHALSKI, R. S. & CHILAUSKY, R. L., (1981). Knowledge acquisition by encoding expert rules versus computer induction from examples: a case study involving soybean pathology. In: E. H. MAMDANI & B. R. GAINES, Eds, *Fuzzy Reasoning and its Applications,* London: Academic Press.

NAYLOR, T. H., BALINTFY, J. L., BURDICK, D. S. & CHU, K. (1966). *Computer Simulation Techniques.* New York: Wiley.

O'KEEFE, R. M., BALCI, O. & SMITH, E. P. (1987). Validating expert system performance. *IEEE Expert,* **2(4),** 81–90.

SHAW, M. L. G. (1980). *On Becoming a Personal Scientist.* London: Academic Press.

SHAW, M. L. G. (1982). PLANET: some experience in creating an integrated system for repertory grid applications on a microcomputer. *International Journal of Man–Machine Studies,* **17,** 345–360.

SHAW, M. L. G. & GAINES, B. R. (1983). A computer aid to knowledge engineering. *Proceedings of British Computer Society Conference on Expert Systems,* Cambridge, pp. 263–271 (December).

SHAW, M. L. G. & GAINES, B. R. (1986). Interactive elicitation of knowledge from experts. *Future Computing Systems,* **1**(2), 151–190.

SHAW, M. L. G. & GAINES, B. R. (1987). KITTEN: Knowledge initiation and transfer tool for experts and novices. *International Journal of Man–Machine Studies,* **27,** 251–280.

SIEGEL, P. (1986). *Expert Systems: A Non-Programmer's Guide to Development and Application.* TAB, Blue Ridge Summit, PA.

SOWA, J. F. (1984). *Conceptual Structures: Information Processing in Mind and Machine.* Reading, MA: Addison-Wesley.

TRAVERS, R. M. V. (1969). *An Introduction to Educational Research,* 3rd edn. MacMillan: London.

VANDIERENDONCK, A. (1975). Inferential simulation: hypothesis-testing by computer simulation. *Nederlands Tijdschrift voor de Psychologie, (Amesterdam),* **30,** 677–700.

WATERMAN, D. A. (1986). *A Guide to Expert Systems.* Don Mills, Ontario: Addison-Wesley.