PROCEEDINGS OF THE INTERNATIONAL CONFERENCE
ON FIFTH GENERATION COMPUTER SYSTEMS 1988,
edited by ICOT. © ICOT, 1988

VALIDATION IN A KNOWLEDGE ACQUISITION SYSTEM WITH MULTIPLE EXPERTS

Mildred L. G. Shaw

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

ABSTRACT

Knowledge Support System Zero (KSS0) is a knowledge support system providing an integrated set of tools for knowledge acquisition for Fifth Generation computer systems. A model for knowledge support system evaluation and validation is discussed. KSS0 has been evaluated against the model, and experiments are reported in knowledge acquisition of spatial interpolation techniques for contour maps. Results are described on validation experiments to show the extent to which this system can replace standard interviewing techniques:

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

A framework for identifying **consensus**, **correspondence**, **conflict** and **contrast** with multiple experts is described.

1 INTRODUCTION

Shaw and Gaines (1983, 1986) emphasized 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 and Gaines 1983). The experiments were based on a spectrum of those with knowledge of business record keeping using an automated personal construct elicitation methodology (Shaw and Gaines 1986).

Various forms of analysis were applied to the resulting grids. The constructs or attributes elicited were compared with the BIAIT constructs by analysing with ENTAIL 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 methodology 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-generated the management hierarchy of the firm from the construct systems of staff involved in managing garment faults.

2 AN OUTLINE OF KNOWLEDGE SUPPORT SYSTEM ZERO

KSS0 combines a number of different sources, including text, expert interviews and expert decision-making, and a number of different techniques, including text analysis, entity and attribute elicitation, clustering and inductive rule generation. It is written in Pascal and runs on the Apple Macintosh family of computers to provide a highly interactive and graphic knowledge acquisition environment. At the heart of KSS0 is an object-oriented knowledge base in which knowledge is formally represented as a multiple-inheritance digraph of classes, objects and properties. Such a structure generalizes the entity and attribute grids used in several early knowledge acquisition systems (Shaw and Gaines 1983, Boose 1985) and has proved both general and powerful in a variety of applications (Boose and Bradshaw 1987, Gaines 1987, Diederich et al 1987).

The main tools in KSS0 are shown in Figure 1:

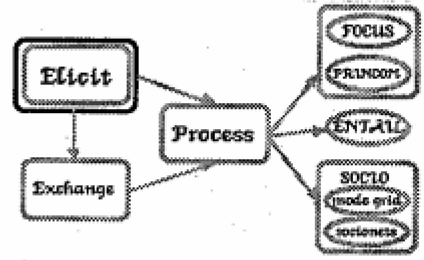
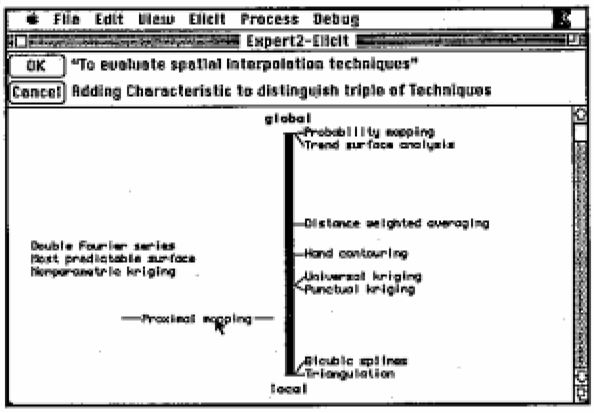


Figure 1. Some tools in Knowledge Support System Zero.

- **Elicit** accepts specifications of entities within a domain and provides an interactive graphical elicitation environment within which the experts can distinguish entities to derive their attributes. The resultant class is continuously analyzed to provide feedback prompting the expert to enter further entities and attributes.
- **Exchange** allows interactive exchange and comparison of entities and attributes from multiple experts in a domain.
- **FOCUS** hierarchically clusters entities and attributes within a domain prompting the experts to add higher-level entities structuring the domain.
- **PRINCOM** spatially clusters entities and attributes within a domain prompting the experts to add higher-level entities structuring the domain.
- **ENTAIL** induces logical entailments enabling the attributes of an element, or the evaluations of a decision-making situation in a domain, to be derived from other attributes.
- **SOCIO** compares the structures for the same domain generated by different experts, or the same expert at different times or from varying perspectives.

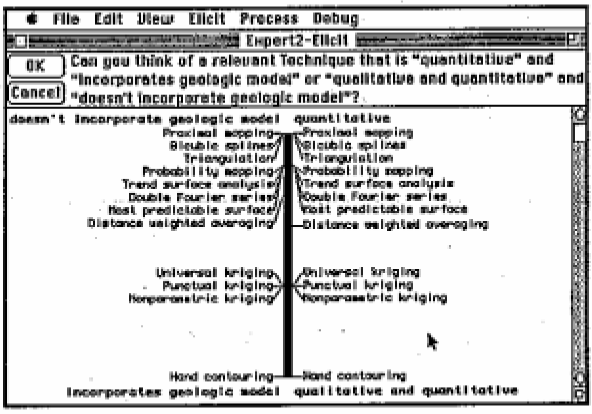
Part of the interaction with one of the geographers is used to give a flavour of Elicit. Figure 2 shows the common entities agreed on by all the geographers being placed on the attribute **local — global**. Some techniques have already been rated, some are waiting in a list on the left, and **proximal mapping** is being dragged on to the attribute bar. Both the attribute labels and the techniques can be edited any time, and also the placing of the techniques. Figure 3 shows two matched attributes, with both sets of ratings shown, so that they can be compared. Again, any one is moveable if the expert wants to adjust anything.



"To evaluate spatial interpolation techniques"
Adding Characteristic to distinguish triple of Techniques

global
Probability mapping
Trend surface analysis
Distance weighted averaging
Double Fourier series
Most predictable surface
Nonparametric kriging
Hand contouring
Universal kriging
Punctual kriging
Proximal mapping
Bicubic splines
Triangulation
local

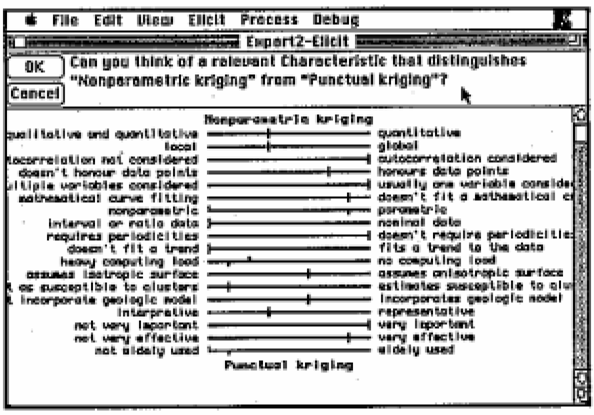
Figure 2. KSS0 showing click and drag elicitation.



Can you think of a relevant Technique that is "quantitative" and "incorporates geologic model" or "qualitative and quantitative" and "doesn't incorporate geologic model"?

doesn't incorporate geologic model | quantitative
Proximal mapping | Proximal mapping
Bicubic splines | Bicubic splines
Triangulation | Triangulation
Probability mapping | Probability mapping
Trend surface analysis | Trend surface analysis
Double Fourier series | Double Fourier series
Most predictable surface | Most predictable surface
Distance weighted averaging | Distance weighted averaging
Universal kriging | Universal kriging
Punctual kriging | Punctual kriging
Nonparametric kriging | Nonparametric kriging
Hand contouring | Hand contouring
incorporates geologic model | qualitative and quantitative

Figure 3. KSS0 showing matching attributes.



Can you think of a relevant Characteristic that distinguishes "Nonparametric kriging" from "Punctual kriging"?

Nonparametric kriging

qualitative and quantitative	quantitative
local	global
tocorrelation not considered	autocorrelation considered
doesn't honour data points	honours data points
ultiple variables considered	usually one variable conside
mathematical curve fitting	doesn't fit a mathematical c
nonparametric	parametric
interval or ratio data	nominal data
requires periodicities	doesn't require periodicitie
doesn't fit a trend	fits a trend to the data
heavy computing load	no computing load
assumes isotropic surface	assumes anisotropic surface
. as susceptible to clusters	estimates susceptible to clu
. incorporate geologic model	incorporates geologic model
interpretive	representative
not very important	very important
not very effective	very effective
not widely used	widely used

Punctual kriging

Figure 4. KSS0 showing matching techniques.

Techniques may also be matched, as shown in the screen in Figure 4. 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. Figure 5 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.

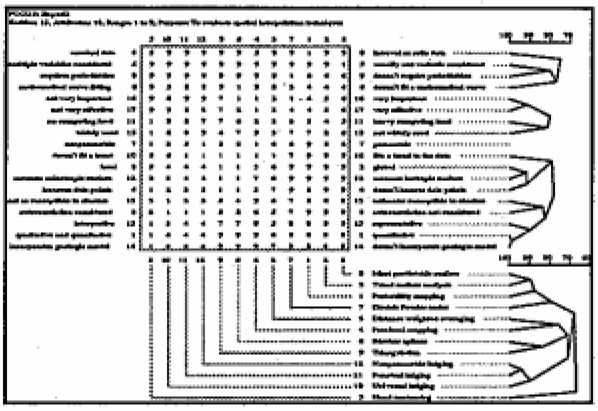


Figure 5. KSS0 FOCUS display of completed grid.

Exchange grids are used for the measurement of understanding and agreement between either two experts or on two occasions (Shaw 1980). To do this two people, possibly experts with differing points of view, each elicit a grid in an area of common knowledge or experience. Each may choose his own entities independently of the other, and elicit and rate his attributes quite separately. Each then can **Exchange** his grid, that is use the other's entities and attributes but not the rating values. This is then completed by the other expert. By comparing pairs of these grids it is possible to map the extent of overlap of the agreement and understanding between the two experts. The **Exchange** procedure allows multiple experts to be aware of and explore the meaning of each other's terminology in the domain. The experts can then understand and extend their own thinking and problem-solving capabilities with ideas from other experts. They are able to see the relationship of their points of view to those of others; explore differing terminology for the same attributes; become aware of differing attributes having the same terminology; extend their own conceptual systems with those of others; provide others with attributes they have found valuable; and explore a problem-solving domain using the full group resources.

The **Process** facility allows single and multiple grids to be analysed. The **FOCUS** algorithm is a distance-based hierarchical cluster analysis technique that sorts the entities and attributes into a linear order such that those closest together in the space are also closest together in the order. It has the advantage in presentation that the sorting is used only to re-present the original grid re-organized by the neighbourness of attributes and elements. It is left to the user to construe his own personal meaning into the result and confirm this directly in terms of the original data. The tree diagrams show the entity and attribute clusters and are imposed on the re-sorted original grid data (Shaw 1980). The program provides a hierarchical clustering of an expert's conceptual system that preserves the data elicited from him so that the sources of the analysis are evident and can be discussed. The **PRINCOM** algorithm is another distance-based cluster analysis using standard principal component analysis and giving the same results as Slater's INGRID (Slater 1976, 1977). This analysis has been used widely because it gives a visually meaningful map of some of the relations between entities and attributes. The program gives a non-hierarchical cluster analysis based on principal components that can be used to gauge the major dimensions along which an expert is making distinctions.

The **ENTAIL** algorithm is a logical analysis of the conceptual system taking the expert's distinctions to be fuzzy predicates (Zadeh 1972, Gaines 1976). The program derives asymmetric implications between the attribute poles so that one can infer how a new entity might be placed on one attribute given how it is placed on others (Gaines and Shaw 1981). It can also provide rules for input to an expert system shell (Gaines and Shaw 1986). **ENTAIL** gives a dependency analysis of the attributes in a grid by deriving significant logical entailments consistent with the data. From a clustering point of view this analysis is significant because it can show up one-way relations that may be missed by distance-based clustering techniques.

However, in the present context it is more significant that the entailments may be treated as inference rules to be used in an expert system. Figure 6 shows some results from the **ENTAIL** analysis of the grid of Figure 5. The entailments are shown with two values in the range from 0 to 1: first, the truth value of the hypothesis, and second, the information content (uncertainty reduction generated) of asserting the hypothesis. For example, L1 -> R14 has a truth value of 0.88, and an information content of 0.25. The information content measures the significance of the hypothesis and is used to ensure that trivial entailments consistent with the data are pruned (Gaines and Shaw 1986).

```
ENTAIL     Truth Information (Cutoff 0.15) Entailment
R14->L 1   1.00  0.30 incorporates geologic model ->
                      qualitative and quantitative
R 1->L14   1.00  0.26 quantitative -> doesn't
                      incorporate geologic model
L 1->R14   0.88  0.25 qualitative and quantitative ->
                      incorporates geologic model
L15->R14   0.75  0.25 interpretive -> incorporates
                      geologic model
L15->L 1   0.75  0.24 interpretive -> qualitative and
                      quantitative
R 4->R12   0.75  0.22 honours data points -> assumes
                      anisotropic surface
R16->R17   0.88  0.22 very important -> very effective
L17->L16   0.88  0.22 not very effective -> not very important
L 1->L15   0.88  0.21 qualitative and quantitative ->
                      interpretive
L 4->L12   0.88  0.21 doesn't honour data points ->
                      assumes isotropic surface
L12->L 4   0.75  0.21 assumes isotropic surface ->
                      doesn't honour data points
R14->L15   0.88  0.21 incorporates geologic model ->
                      interpretive
R12->R 4   0.88  0.20 assumes anisotropic surface ->
                      honours data points
L 4->R 2   0.75  0.19 doesn't honour data points -> global
R 3->R14   0.62  0.18 autocorrelation considered ->
                      incorporates geologic model
R 3->L15   0.62  0.18 autocorrelation considered -> interpretive
L12->R13   0.75  0.18 assumes isotropic surface ->
                      estimates susceptible to clusters
L12->R15   0.88  0.18 assumes isotropic surface -> representative
L13->L15   0.62  0.18 not as susceptible to clusters ->
                      interpretive
L15->R12   0.88  0.18 interpretive -> assumes anisotropic
                      surface
```

Figure 6. Some rules produced by the **ENTAIL** algorithm from Expert 2.

The **SOCIO** program provides facilities for comparing and contrasting multiple sources of expertise. It is an extension of SOCIOGRIDS (Shaw 1980) for deriving socionets and mode attributes from groups of individuals construing the same class of elements. Its objective is to take different classes representing the same domain and compare them for their structure, showing the similarities and differences. It may be regarded as the implementation of a simple form of analogical reasoning. The grids may have the same entities and attributes but possibly differing values. **SOCIO** analyzes the matches between the entities and the attributes in the grids according to the values, and shows those entities and attributes that are similar and those which are different. A typical application is to see whether experts agree on the definitions of classes by asking them to separately fill in the values for a domain definition **Exchanged** between them. If they agree then it is **consensus**, and if they disagree it is **conflict**.

If the grids have different attributes but the same entities **SOCIO** analyzes the matches between the attributes in the grids, and for each attribute in the original shows the closest matching attribute in the other class. A typical application is to see whether experts are using different terminologies for the same attributes by asking them to define attributes separately and fill in the values for a domain defined through an agreed set of elements. Here, **correspondence** between entities or attributes can be identified. If there is no similarity in either terminology or the use of the attributes, this is **contrast**. Figure 7 shows the four possibilities which can occur in this intersection of the experts' conceptual structures.

		Terminology	
		Same	Different
Attributes	Same	***Consensus*** Agreement in the terms and vocabulary used, and the conceptual structures in the system are used in the same way.	***Correspondence*** Different terms and vocabulary are used, but the conceptual structures are the same.
	Different	***Conflict*** Two experts use the same terms and vocabulary but in a conceptually different way.	***Contrast*** Experts use different terms and vocabulary in different ways, having contrasting conceptual structures.

Figure 7. Consensus, correspondence, conflict and contrast among experts.

When a number of grids representing the same domain are available, **SOCIO** also provides two other forms of analysis. The first is that it tries to derive "modal" entities or attributes that reflect a consensus among the experts by extracting those which occur as highly matched entities or attributes across the majority of classes. A typical application is to reach consensus on critical concepts that are associated with a rich vocabulary at differing levels of abstraction. The other is that it derives a set of socionets showing the degree to which each expert is able to make the same distinctions as another expert, even if they use different terminology. A typical application is to validate the knowledge acquisition process by determining whether the structure derived conforms with known relations between the experts (Shaw and Gaines 1989).

3 EVALUATION OF KNOWLEDGE-BASED TOOLS

Evaluation studies of knowledge-based systems (specifically expert systems) have been reported and methods for validation have been outlined. A typical example is Siegel (1986) who points out the need to evaluate the output of expert systems and he outlines in good 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 and 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 of the performance of the system against the problem it was designed to solve. Knowledge acquisition was considered difficult to design and problematic at best. 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 1986) 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 for 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 (1983) who compared the inductively derived rules from AQ11 with direct representations of the expert's knowledge. Other automated systems such as AQUINAS (Boose and Bradshaw 1987), MORE (Kahn et al 1985), SALT (Marcus 1986) and MOLE (Eshelman et al 1987), which have been designed as 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 et al (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 the evaluation of the systems per se. What appears lacking is a useful framework to guide acquisition tool evaluation. The following section outlines a model for the evaluation and validation of knowledge acquisition systems.

4 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 explication of components and requirements: determine the external criteria which act as comparison points for the system under evaluation; recognize levels of operation and measurement in the external criteria; determine and describe the input and output of the system under evaluation; require clarification, justification and theoretical support of the methods of processing inputs to outputs (e.g. the algorithms); carefully select populations for study.

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 and Gaines, 1987). These are also incorporated into the evaluation framework.

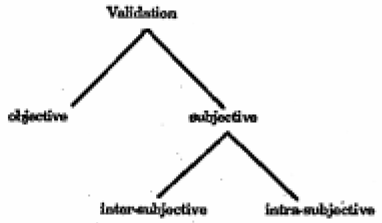


Figure 8. A validation model.

Figure 8 shows two distinct aspects of validation: those concerned with objective 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 looks at objective aspects of a system.

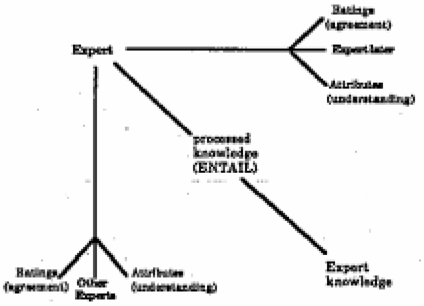


Figure 9. The framework for the study.

The subjective aspects relate to the way in which the knowledge acquisition system acquires the knowledge that the expert has. Since it is 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.

Figure 9 shows the model developed for this study. It was necessary to see if there was mutual agreement across experts and 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 **ENTAIL** analysis performed by the system on each expert's knowledge was seen as correct or not.

5 METHODOLOGY

Figure 10 outlines the procedural steps and the data collection points used in the subjective evaluation of KSS0 (Shaw and Woodward 1988). It shows the first step as Procedure 1 (P1) which introduced the experts 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 expert individually. This provided an initial set of entities and attributes particular to each expert. The second collection point (G2) consisted of the experts exchanging their first grids (entities and attributes) to produce other grids. The next step was a procedure (P2) to have the experts 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 was the other, separately for each expert. P3 represents a wait of over twelve weeks 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 experts 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 experts a selected list of entailments using the **ENTAIL** algorithm (Gaines and 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 the experts who were asked to rate each entailment on a four point scale from correct to incorrect.

Event	Description
P1	determine task demands and elicitation purpose
G1	grid elicitation using own entity set
G2	grid elicitation 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: 12 weeks
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

Figure 10. Evaluation design event table.

Two separate studies were completed using specific task demands from two separate and quite different domains, but only the geography one will be reported here. It 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 minutes. 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.* 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, twelve 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 and Shaw 1986, Shaw and Gaines 1987) were produced for each expert from G3. These rules were then sampled from each category of significance, ≥ .15, .14 to .10, and .09 to .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.

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.

6 THE MEASURES

Two measures were used in this study, **consistency-with** another and **construed-by** another. **Consistency** is measured using **Exchange.** 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 **SOCIO** 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:

G consistency-with G' at criterion = 100 x

$$\frac{\text{(number of attributes in G matched greater than criterion by } \underline{\text{same attributes}} \text{ in G')}}{\text{(number of attributes in G)}}$$

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:

G construed-by G' at criterion = 100 x

$$\frac{\text{(number of attributes in G matched greater than criterion by } \underline{\text{any attributes}} \text{ in G')}}{\text{(number of attributes in G)}}$$

Operational definitions for the various construing 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. 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:

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

7 RESULTS

Figures 11 to 16 display the results. The **intra-subjective understanding** scores, U_t, displayed in Figure 11 ranged from 63 to 86; and the **intra-subjective agreement** scores, A_t, from 79 to 94, indicating that each expert was using her/his terminology in much the same way from one occasion to the other and that KSS0 was able to reflect this consistency.

Expert	$U_t \geq 80$	$A_t \geq 80$
E1	62.5	81.2
E2	77.8	94.4
E3	85.7	78.6

Expert Pairs	$U_i \geq 80$	$A_i \geq 80$
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

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

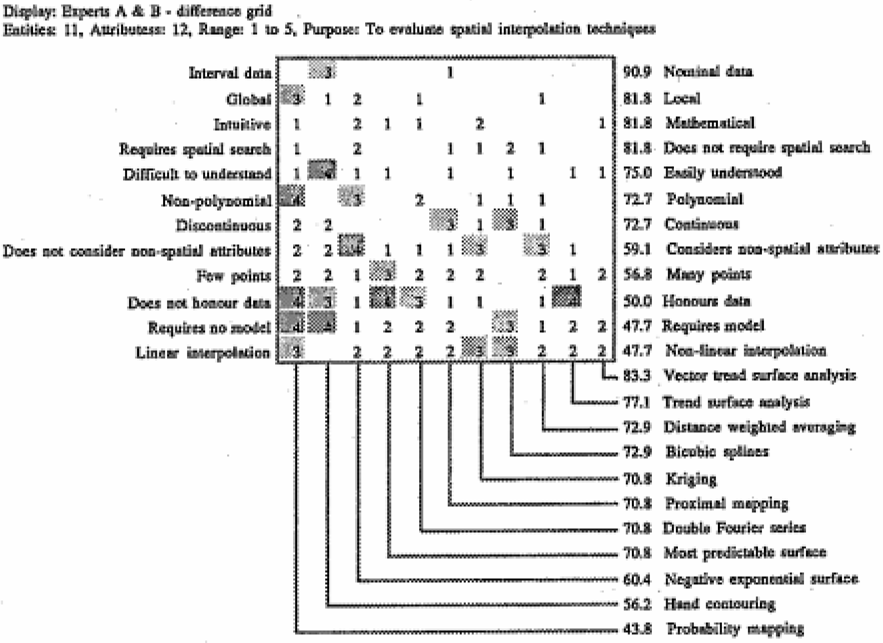
Fig. 12.Inter-subjective agreement (A_i) and understanding (U_i) results.

```
E1:E2 33.3% over 80.0 (E1 attribute-consistency-with E2)
 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 (E1 attribute-consistency-with E3)
 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
```

Figure 13. Analysis of inter-subjective agreement.

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



Display: Experts A & B - difference grid

Entities: 11, Attributess: 12, Range: 1 to 5, Purpose: To evaluate spatial interpolation techniques

Interval data		3				1						90.9	Nominal data
Global	3	1	2		1				1			81.8	Local
Intuitive	1		2	1	1		2				1	81.8	Mathematical
Requires spatial search	1		2			1	1	2	1			81.8	Does not require spatial search
Difficult to understand	1	4	1	1		1		1		1	1	75.0	Easily understood
Non-polynomial	4		3		2		1	1	1			72.7	Polynomial
Discontinuous	2	2				3	1	3	1			72.7	Continuous
Does not consider non-spatial attributes	2	2	4	1	1	1	3		3	1		59.1	Considers non-spatial attributes
Few points	2	2	1	3	2	2	2		2	1	2	56.8	Many points
Does not honour data	4	3	1	4	3	1	1		1	4		50.0	Honours data
Requires no model	4	4	1	2	2	2		3	1	2	2	47.7	Requires model
Linear interpolation	3		2	2	2	2	3	3	2	2	2	47.7	Non-linear interpolation

83.3 Vector trend surface analysis
77.1 Trend surface analysis
72.9 Distance weighted averaging
72.9 Bicubic splines
70.8 Kriging
70.8 Proximal mapping
70.8 Double Fourier series
70.8 Most predictable surface
60.4 Negative exponential surface
56.2 Hand contouring
43.8 Probability mapping

Figure 14. The difference grid for experts E1 and E2.

This can be seen in Figure 13 where the details of the **SOCIO** analysis are shown. E1 has exchanged his techniques and attributes with the other two experts, E2 and E3. The listing at the top of Figure 13 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. Inspection of the raw grids for E1 and E2 show the reason for these results. The first number on the first line is a 1 in E1's grid and a 5 in E2's. This means that E1 thinks that *probability mapping requires model*, whereas E2 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. This is an example of **conflict** where E1 and E2 are using the same terminology in conceptually different ways. There is also conflict shown in the concept of *linear interpolation—nonlinear interpolation*.

Figure 13 can be redrawn as a difference grid where rating values (in this case 1 to 5) for E2's ratings of E1's entities on his attributes are subtracted from E1's similar rating values respectively. Figure 14 shows this with the entities and attributes about which they agree the most in the top right corner, shown by no difference or a difference of only 1; and those with most disagreement towards the bottom left, shown by the maximum difference of 4 or a large difference of 3. Hence from this difference grid, the substance of the agreements and the disagreements can easily be identified and discussed by the two experts.

```
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
```

Figure 15. Analysis of inter-subjective understanding.

Inter-subjective understanding scores, U_i, are also shown in Figure 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. These values range from 31 to 71, and Figure 15 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. The cumulative percentage is given of those with matches greater than the value shown, and the attribute from E2 which best matches each from E1 is shown beneath it. This was repeated for each of E1's attributes.

It can be seen that:

- The first or highest match which accounts for 6.2% of the attributes has a level of 88.5 out of a possible 100 if they were identical. That is, both experts are using the attribute *local— global* in the same way. This is an example of **consensus**.
- The first and second matches together account for 12.5% of all attributes, and they are matched over the level of 87.5. However, this is not consensus since when E1 uses the attribute *low-level-data—high-level data* , E2 is using the attribute *nominal data—interval or ratio data* . This is a difference in terminology which indicates that their levels of abstraction are different in their construing of this topic. This shows the two experts in **correspondence**.
- The third match again shows **consensus**, with both experts using the attribute *does not honour data points—honours data points* in the same way.
- The fourth match shows a difference in terminology, or a **correspondence** between *short-distance autocorrelation—long-distance autocorrelation* and *local—global*. Notice that *local—global* used by E2 was also used in the first match indicating that E1 has two attributes *short-distance autocorrelation—long-distance autocorrelation* and *local—global* which are used similarly to each other but with different terminology, whereas E2 has only one. In fact, looking on to the seventh and eleventh matches, it can be seen that E1 has two more attributes *discontinuous—continuous* and *does not require a priori model—requires a priori model* which correspond to E2's single attribute *local—global*. This shows a differences in richness of concepts not necessarily making new distinctions in the class so far defined by the entities.
- The eighth match, still over the level of 82, again shows **correspondence**. It shows the attribute *heavy computing load—no computing load* is being used by E2 to correspond to *mathematically complex—mathematically simple* used by E1. This is a difference in terminology corresponding to a correlation in the real-world.

Each match can be examined in this way for each pair of experts to determine levels of **consensus** and **correspondence**. **Contrast** occurs when the experts use different terms and vocabulary in different ways, showing that they have contrasting conceptual structures. These results suggest that KSS0 provides 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.

Study 1	CI
E1	88.2
E2	91.9
E3	82.2

Figure 16. Intra-subjective perspective consistency results (C_i).

Intra-subjective perspective consistency scores C_i are displayed in Figure 16. 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 and Shaw 1986, Shaw and Gaines 1987) algorithm. They range from 82 to 92, showing a high level of expected rules appearing in the 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 E2 were shown in Figure 6.

8 CONCLUSIONS

KSS0 is a knowledge support system providing an integrated set of tools for knowledge acquisition. **Elicit** provides facilities for eliciting the important dimensions of an expert's thinking on a topic; **Exchange** extends this to share entities and attributes between experts and elicit differences in perspective and terminology as well as disagreements on the topic. **SOCIO** processes results from several experts to reveal the similarities and differences in the concept systems of different experts, or the same experts at different times, construing a domain defined through common entities or attributes. It can be used to focus discussion between experts on those differences between them which require resolution, enabling them to classify them in terms of differing terminologies, levels of abstraction, disagreements, and so on. It provides a framework for identifying consensus, correspondence, conflict and contrast in a knowledge acquisition system with multiple experts.

Some preliminary studies have been reported on validating this knowledge support system, and 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, user of the system, or even another expert, may be mistaken. The experts 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, but when used for one expert alone for intra-subjective agreement gave much higher values. This measure, then, shows a good range of discrimination within experts and between experts. A similar rationale can be applied in the construed-by measures of inter- and intra-subjective understanding.

ACKNOWLEDGEMENTS

Financial assistance for this work has been made available by the Natural Sciences and Engineering Research Council of Canada. The KSS0 system is made available by the Centre for Person Computer Studies. I would like to thank my colleagues Brian Gaines and Brian Woodward for discussion of this work, and for help in conducting the studies.

REFERENCES

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

Boose, J.H. (1985). A knowledge acquisition program for expert systems based on personal construct psychology. **International Journal of Man-Machine Studies 20**, 21-43.

Boose, J.H. & Bradshaw, J.M. (1987). Expertise transfer and complex problems: using AQUINAS as a knowledge acquisition workbench for knowledge-based systems. **International Journal of Man-Machine Studies (26)**, 3-28.

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, U.K.

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

Cleaves, D.A. (1987) Cognitive biases and corrective techniques: proposals for improving elicitation procedures for knowledge-based systems. **International Journal of Man-Machine Studies 26**, 155-166.

Diederich, J., Ruhmann, I. & May, M. (1987). KRITON: a knowledge-acquisition tool for expert systems.**International Journal of Man-Machine Studies 26**, 29-40.

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. (1976). Foundations of fuzzy reasoning. **International Journal of Man-Machine Studies, 8**(6), 623-668 (November).

Gaines, B.R. (1987). Rapid prototyping for expert systems. Oliff, M. (Ed). **Proceedings of International Conference on Expert Systems and the Leading Edge in Productions, Planning and Control.** pp. 213-241. University of South Carolina.

Gaines, B.R. & Shaw, M.L.G. (1981). New directions in the analysis and interactive elicitation of personal construct systems. Shaw, M.L.G., Ed. **Recent Advances in Personal Construct Technology.** pp. 147-182. London: Academic Press.

Gaines, B.R. & Shaw, M.L.G. (1986). Induction of inference rules for expert systems. **Fuzzy Sets and Systems, 18**, 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.

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

Michalski, R.S. (1983). A theory and methodlogy of inductive learning. In Michalski, J. Carbonell, T & Mitchell, E. (eds) **Machine Learning.** Palo Alto: Trago Publishing

O'Keefe, R.M., Balci, O., & Smith, E.P. (1987). Validating expert system performance. **IEEE Expert 2**(4) 81-90 (Winter).

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**, 263-271 (December). Cambridge.

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 Tools for Experts and Novices. **International Journal of Man-Machine Studies, 27**, 251-280.

Shaw, M.L.G. & Gaines, B.R. (1989). A methodology for recognizing consensus, correspondence, conflict and contrast in a knowledge acquisition system. **International Journal of Man-Machine Studies**, (in press).

Shaw, M.L.G. & Woodward, J.B. (1988). Validation in a knowledge support system: consistency and.construing with multiple experts. **International Journal of Man-Machine Studies**, (in press).

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

Slater, P., Ed. (1976). **Dimensions of Intrapersonal Space: Volume 1.** London: John Wiley.

Slater, P., Ed. (1977). **Dimensions of Intrapersonal Space: Volume 2.** London: John Wiley.

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

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

Zadeh, L.A. (1972). Fuzzy languages and their relation to human intelligence. **Proceedings International Conference on Man and Computer.** Basel: Karger.