2nd AAAI Knowledge Acquisition for Knowledge-Based Systems Workshop,
Banff, Canada, Oct., 1987

Validation of a Knowledge Support System

Mildred L. G. Shaw & J. Brian Woodward
Department of Computer Science,
University of Calgary
Calgary, Alberta,
Canada T2N 1N4

ABSTRACT

At the previous workshop on Knowledge Acquisition for Knowledge-Based Systems in 1986, criteria for a Knowledge Support System (KSS) were discussed, and a preliminary version of KITTEN (Knowledge Initiation and Transfer Tools for Experts and Novices) was described and demonstrated on Apollo workstations. A three-stage model for KSS evaluation and validation is discussed. 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 this system can replace standard interviewing techniques. Some of the question asked were:

1) To what extent does an expert find the generated rules meaningful?
2) To what extent do other experts agree with the expert's ratings of her/his elements on her/his constructs?
3) Can the same constructs be obtained from the expert after a significant time lapse?
4) Can the same ratings of the elements on the constructs be obtained from the expert after a significant time lapse?

INTRODUCTION

In the last Workshop on Knowledge Acquisition for Knowledge-Based Systems, in it was clear that several researchers were considering the great 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 and Gaines (1983, 1986a) 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. Again, 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.

Various forms of analysis were applied to the resulting grids. The constructs 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 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

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.

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 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 (1986a, 1986b) 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.

FRAMEWORK FOR EVALUATION

In discipline research the 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. 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 usefulness of operative effectiveness but also to help guide future development and improvement of the system under evaluation.

A fifth requirement of evaluation research is the proper 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 cannot pose a threat to internal and external validity (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 reliability 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 model for Knowledge Support Systems (KSSs). Applying the above criteria to KSSs, evaluation stages and specific evaluation tasks are outlined for the evaluation activity. General design considerations have been elucidated for KSSs (Shaw and Gaines, 1986b). These are also incorporated into the evaluation framework.

The evaluation model in Figure 1 presents three distinct levels of activity. These distinctions represent the various levels of operation of KSSs. Primary Level evaluation is concerned with

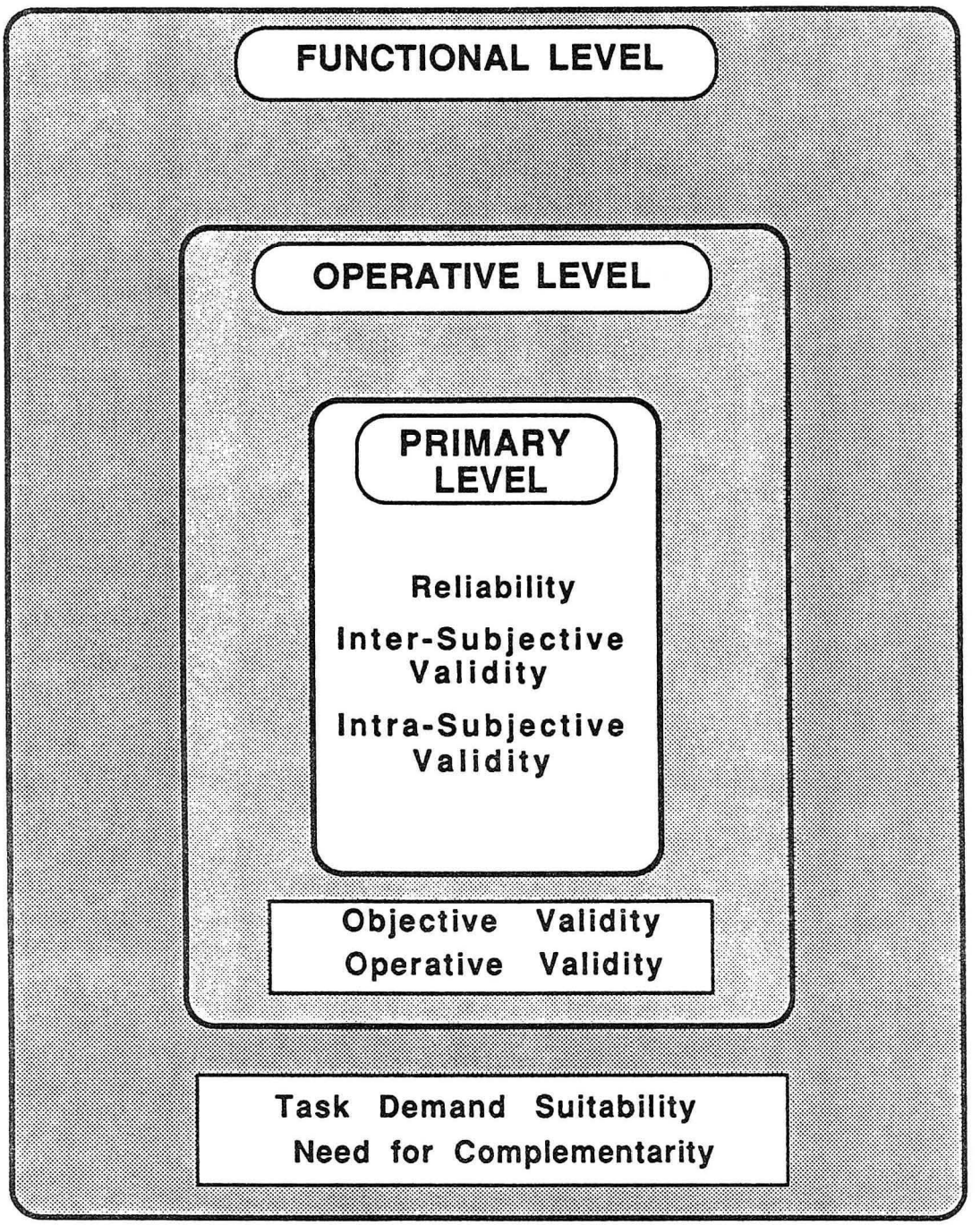


Figure 1. Evaluation Model for Knowledge Support Systems

very basic attributes of KSS operation: the systems reliability over time and the degree to which system output demonstrates basic measures of validity.

Primary Level

A measure of diachronic reliability (Chow, 1987) appears most suitable for evaluating KSSs. This measure reflects the degree of similarity, over time, of the input required for the system and of the calculated output. A KSS reflects a high degree of reliability in the sense that given the same, specific task demands, similar information would be required as input to the system on separate occasions. Also, the output of the KSS, the results of the algorithms, would reflect a high degree of similarity over time. Strictly speaking, this approach to reliability is strongly influenced by the reliability of the user of the KSS. However, the KSS can be viewed as an aid to the user which helps the user maintain consistency over time given the same task demands. Also, KSSs are not used in isolation but are used by individuals so any measure of a KSS must be done with those would would use the system. If reliability measures are low, the system cannot be considered consistent or helpful. For a KSS, this could indicate a wide variance in input and output which would affect any subsequent procedures or uses.

Although **reliability** stands as a necessary requirement of any KSS, other basic requirements demand attention. The KSS must also demonstrate various forms of **validity**. In this sense, validity refers to the extent to which the KSS produces what it is supposed to produce. At the Primary level of evaluation, validity is defined by two measures. The first is **intra-subjective validity**. Based on the concepts of content and face validity (Travers, 1969; Campbell & Stanley, 1963), this type of validity is defined as the extent to which the KSS output is a representative sample of that output. Valid KSSs produce output which has a high degree of similarity with what is apparently expected. The absence of this form of validity 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 KSS output.

Another, complementary measure of validity at this level of evaluation is that of **inter-subjective validity**. Based on the concept of concurrent validity (Chow, 1987; Travers, 1969) this type of validity 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 KSS, then the system can be said to afford a degree of inter-subjective validity. The extent to which two or more people produce the the same output represents the degree of validity.

Even though experts may seldom agree, if a task is properly defined, specifically outlined and agreed upon as important, then a certain degree of similarity can be expected in the outcome of the KSS. 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 validity demonstrated by various experts, while using the system. In the same way as for reliability, the KSS can be seen as an aid to users.

Operative Level

Once the KSS has demonstrated acceptable levels of reliability and intra- and inter- subjective validity, it can be evaluated against more objective criteria. **Objective validation** refers to the degree to which the output of a KSS agrees with well established facts in the domain. This can be accomplished by measuring the degree of convergence of KSS output with established facts from reputable experts and or from reputable printed material. Also, a measure of the degree of divergence of KSS output from known non-truths can be ascertained: the degree to which the KSS output is dissimilar from known falacies. Valid KSS output demonstrates a high degree of

similarity (convergence) with established domain facts and a high degree of dissimilarity (divergence) from falacies.

Another form of validation at this level concerns the usefulness of the output of the KSS. **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 KSS is able to reduce the amount of unnecessary or superfluous output represents a measure of operative validity.

Functional Level

Once the KSS has demonstrated acceptable degrees of objective and operative validity, its functional role attributes require assessment. In other words, the KSS 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 KSS requires testing using a variety of task demands apart from those used in the last level of evaluation. Measures of the degree of KSS reliability and validity on similar task demands and on dissimilar task demands begins to indicate the limits of the system's usefulness, effectiveness and areas of operation.

At this level of evaluation it is necessary to elucidate any and all support processes or complementary procedures which are required prior to or after using the KSS. Effective use of the KSS may require support to meet its general requirements. Shaw & Gaines (1986b) have outlined these general requirements for a KSS.

METHODOLOGY

Figure 2 outlines the procedural steps and the data collection points used in the Primary Level evaluation of KSS0 which is the grid elicitation and analysis part of KITTEN (Shaw & Gaines, 1986b). The version used runs on the Apple Macintosh. 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 elements and constructs particular to each subject. The second collection point (G2) consisted of the subjects exchanging their first grids (elements and constructs) to produce other grids. The next step was a procedure (P2) to have the subjects agree on a common set of elements. This common element set provided the basis for the next data collection point (G3) which consisted of the elicitation of a grid by each expert using common elements. 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 element set but the regeneration of constructs. For G5 the subjects were given the element and construct set from G3 and asked to re-rate the elements. 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 the subjects 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 element set
G2	grid eliciation using other expert's elements and constructs
P2	experts discuss and produce a common element set
G3	grid elicitation using common element set
P3	waiting period: 2 to 3 months
G4	grid elicitation using G3 element set
G5	grid elicitation using elements and constructs from G3
P4	obtain subjective ratings from entailment output from G3
	Analysis

Figure 2. Evaluation Design Table

The Studies

Two separate studies were completed using specific task demands from two separate and quite different domains. 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 constructs with the others. This involved Expert 1 in seeing Expert 2's elements and constructs 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 elements 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 constructs 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 constructs 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 constructs for importance on a 1 to 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 constructs before the next session.

In the fourth session, twelve weeks later, each expert in turn elicited another set of constructs using the agreed set of elements from session three producing G4. Immediately this was completed, in about an hour, each expert was shown his elements and constructs 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 were produced for each expert from G3. These rules were then sampled from each category of significance, $\geq .15$, .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.

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 elements, 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

Operational definitions for the various reliability and validity measures were used to evaluate the KSS. **Construct reliability, R_c,** was defined as the degree of similarity in constructs generated in G3 and G4. Operationally, this measure clustered the sets of constructs from both grids using the FOCUS algorithm (Shaw, 1980) and measuring how many clusters above a cut-off of 75% similarity were mixed. Those constructs which were matched on the G3 and G4 grids at 75 % or better were compared to the total number of possible matches between G3 and G4. Figure 3 shows how the calculation was made for expert E2 from study 1. The following formula was used to calculate R_C for each subject:

$$R_C = \frac{\text{(number of clusters from both G3 and G4)}}{\text{(total number of clusters)}} \times 100$$

Rating reliability, R_r, was defined as the degree of similarity in element ratings in G3 and G5. To calculate this measure, the MINUS algorithm (Shaw, 1980) was used to subtract two grids which have the same elements and constructs. The result was a percentage degree of difference between the two grids. The following formula was used to calculate R_r for each subject:

$$R_r = (100 - \text{difference percentage in G3 and G5})$$

Inter-subjective validity for constructs, V_c, was defined as the similarity in constructs in G3 for each of the experts. The grids for each expert at G3 were combined, and the same measure used as described in construct reliability. In order to be matched, the clusters from study 1 were counted as mixed if they contained constructs from two or three experts.

$$V_C = \frac{\text{(number of clusters from all G3s)}}{\text{(total number of clusters)}} \times 100$$

Inter-subjective validity for ratings, V_r, was investigated by looking at the **exchange grids**. Do the experts see the topic in a similar way at the same time? By exchanging elements and constructs they are able to view the topic from the perspective of the other. To do this involves an understanding of what the other's elements and constructs 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?** This was defined as the degree of similarity between one subject's ratings of his/her own elements and constructs and another subjects's ratings of the first subjects's elements and constructs. Operationally, this measure was calculated using the MINUS algorithm to subtract one grid from another resulting in a difference score. The following formula was used:

$$V_r = (100 - \text{difference percentage in G1 and G2})$$

[N.B. In study 1 there were two G2s: G2a and G2b, so two measures resulted.]

Finally, **intra-subjective validity, V_i,** was defined as the degree of similarity 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 V_i:

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

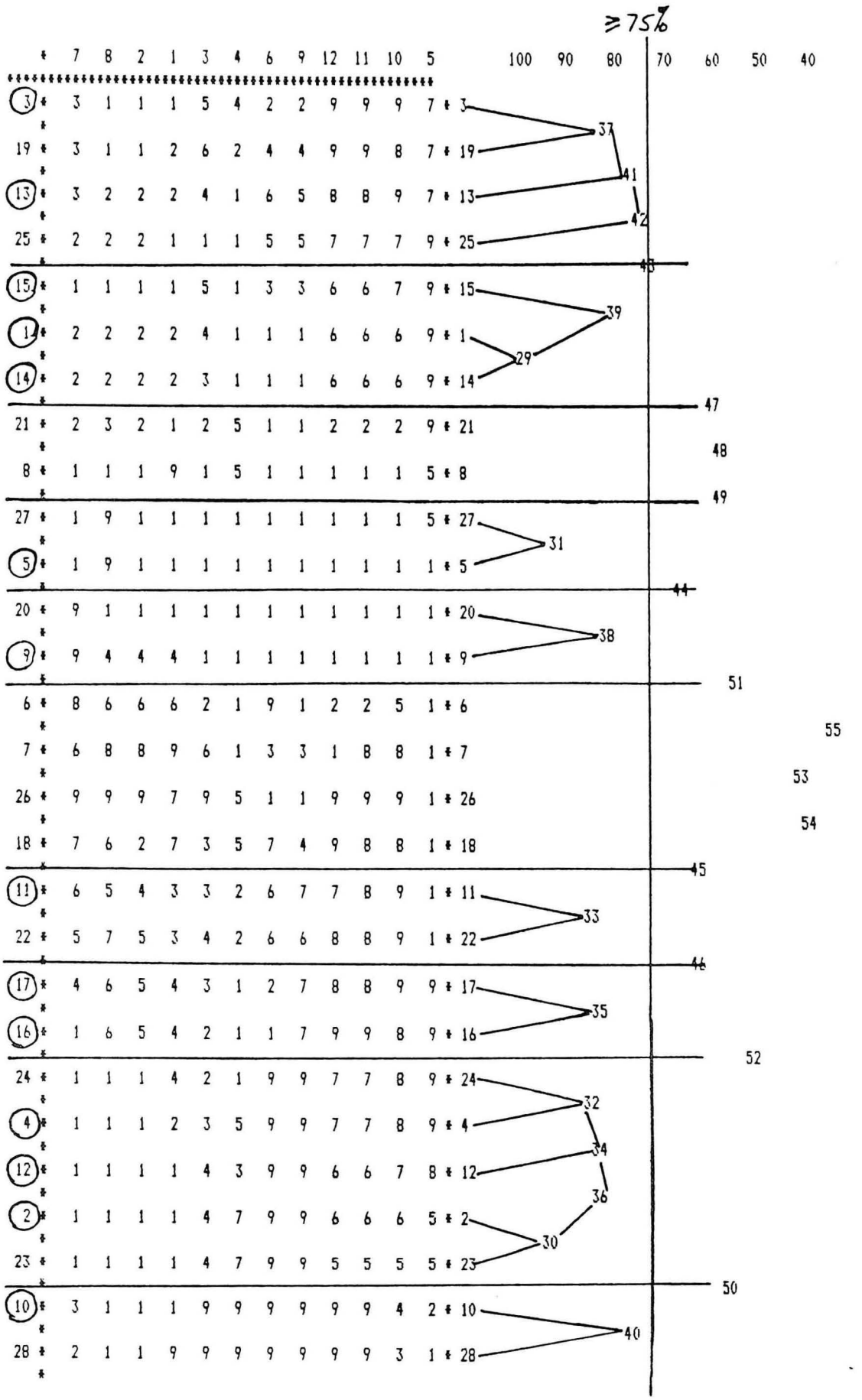


Figure 3. Construct reliability = (number of clusters from both G3 and G4) x 100
(total number of clusters)

= 6 / 8 x 100 = 75%

RESULTS

Figures 4 to 6 display the results from the two studies. The **construct reliability** scores displayed in Figure 4 ranged from 50% to 83% in the first study, but were 100% in the second. This may reflect the fact that the three experts in the first study are academic researchers at a university, and hence tended to be very 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 professional 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 **rating reliability** scores also displayed in Figure 4 ranged in both studies from 84% to 89%, indicating that all experts were using their ideas in much the same way from one occasion to the other and that KSS0 was able to reflect this consistency.

Inter-subjective validation scores, displayed in Figure 5, that is to what extent the experts agree, range in the first study from 68% to 76%. It is interesting to note that the lowest scores of 68% both come from the other experts trying to use the constructs of the senior researcher. This could indicate that he has a greater variety of distinct constructs to use in such a situation that they do. In the second study the scores range from 67% to 100%. Here it is interesting to note that the trainer could use the constructs of the engineer totally, but the engineer could only use two thirds of the trainers constructs.

Except for the 100% agreement in the second study, these ratings are remarkably similar. This suggests that KSS0 shows a facility at allowing experts to compare their understanding and to determine a level of agreement. Whether the percentages 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 validation scores are displayed in Figure 6. These show whether the expert finds the rules meaningful, and rates the rules as correct or significant in the same way as the ENTAIL algorithm. They range in study 1 from 82% to 92%, showing a high level of expected rules appearing in the KSS output.

CONCLUSIONS AND FUTURE WORK

In KSS evaluation, internal and external contamination factors 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.

The variability of the subjects presents a consistent internal problem in evaluation studies of KSSs. The variance of the outcome measures due to the user is explicably tied to and requires isolation from the variance due to the KSS. Procedures and measures which accomplish this are required for an accurate assessment of the KSS. A measure which determines the extent that a KSS helps the user maintain consistency of input would be useful.

Developing operational definitions and formulae of reliability and validity measures require refinement and testing. The application of standard measures may be misleading and/or artificial when used with KSSs. Appropriate definitions of measurement need to be developed for KSS evaluation to be meaningful and useful.

Study 1	Rc	Rr
E1	50%	86.2%
E2	75%	89.4%
E3	83.3%	85.3%
Study 2		
E1	100%	84.4%
E2	100%	89.2%

Figure 4. Reliability Results

Expert Pairs	%
Study 1 E1's agreement with E2	76.1
E2's agreement with E1	68.2
E1's agreement with E3	72.2
E3's agreement with E1	68.8
E2's agreement with E3	72.6
E3's agreement with E2	71.1
Study 2 E1's agreement with E2	67.4
E2's agreement with E1	100

Figure 5. Inter-Subjective Validity for Ratings

Study 1	Vi
E1	88.2%
E2	91.9%
E3	82.2%

Figure 6. Intra-Subjective Validity for Rules

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.

The extent to which training and/or prior exposure to the KSS affects measured outcomes is also a factor in evaluation. To what extent did prior training alter the subject's approach to the KSS? This issue also suggests another: the effect due to growing familiarity with the KSS 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 reliability affected due to the development of new understanding by the subject from one time to another? Lower scores for reliability may indicate a "new" or "better" understanding of the material. This development may also be attributed to the demands of the experiment.

There are many more ways in which the data is being processed, especially in terms of the rules derived from the various grids. In this paper we have looked at the similarities ane differences between experts and within experts over a 2 to 3 month time interval. Clearly, the content of these similarities and differences can be identified and related to external criteria.

One of the directions in which we are continuing this study, is to look into the objective validation of KSS. 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.

ACKNOWLEDGEMENTS

Financial assistance for this work has been made available by the National Sciences and Engineering Research Council of Canada. The KITTEN system is made available by the Centre for Person Computer Studies; the initial Apollo implementation is being carried out at the Knowledge Sciences Institute, University of Calgary. We would like to thank our colleague Brian Gaines for discussion of this work, and Robin Edmond, Karen 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 (August).**

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

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

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.

Eshelman, L.,Ehret, D. McDermott,J & Tan, M (1987). MOLE: a tenacious knowledge acquisition tool. International Journal of Man-Machine Studies,Vol.26 (1),41-54

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. (1986). Induction of inference rules for expert systems. **Fuzzy Sets and Systems 18**(3) 315-328 (April).

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

Johnson, P.E. (1986b).Specification of expertise:knowledge acquisition for expert systems. **Proceedings of Knowledge Acquisition for Knowledge-Based Systems Workshop**. Banff.

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

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

Marcus, S. (1986) Taking backtracking with a grain of SALT. **Proceedings of Knowledge Acquisition for Knowledge-Based Systems Workshop**. Banff.

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

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**(3), 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. (1986a). Interactive elicitation of knowledge from experts. **Future Computing Systems, 1**(2) 151-190.

Shaw, M.L.G. & Gaines, B.R. (1986b). Techniques for Knowledge Acquisition and Transfer. **Proceedings of the Workshop on Knowledge Acquisition for Knowledge-Based Systems,** November, Banff, 39-0 — 39-13.

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, Massachusetts: Addison-Wesley.

Travers, R.M.V. (1969) **An Introduction to Educational Research (Third Edition).** MacMillan: London.

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