The controlled experiment in knowledge-acquisition research

by C. N. Nicholson

This paper is based on a review of the literature about controlled experiments in research on knowledge acquisition. The review was carried out to help the author make decisions about the design of his own experiment comparing two knowledge-acquisition methods. The paper looks critically at six experiments reported in the literature, and proposes a framework within which such empirical work can be viewed. It concludes that some of the apparent difficulties can be resolved, and that controlled experiments can be a useful way of discovering the relationships at work in a knowledge-acquisition project.

Introduction

Case studies and benchmarks have been used widely in research on knowledge-based systems. For example, in a case study, Michalski and Chilausky [1] investigated the effect of the acquisition method in a single domain on the effort needed to acquire the knowledge and on the diagnostic accuracy of the resulting knowledge bases. In a benchmark, Quinlan [2] used several case bases as input to different induction algorithms, and observed the effect of these variables on the diagnostic accuracy of the induced knowledge bases.

But there is a growing awareness [3] that controlled experiments can help advance understanding of how the knowledge source, representation, acquisition method, domain, and engineer affect the effort needed to build a knowledge base and the quality of its performance. Burton and Shadbolt [4, p. 11] argue strongly in favor of controlled experimentation:

> Although one can get useful practical information from case studies, there will always be many factors unique to any particular knowledge-elicitation session. Hence the need for a formal experimental analysis.

Indeed, researchers such as Burton et al. [5], Lundell [6], Stevenson et al. [7], Deffner and Ahrens [8], Adelman [3], and Agarwal and Tanniru [9] have used methods from experimental psychology to explore research questions in knowledge acquisition.

The author's interest in the subject arose from his own need to compare two knowledge-acquisition methods[1] in terms of the effort they demand from a domain expert, and the accuracy of their outcomes. The controlled experiment seemed the ideal way to do the investigation, so a search was made of previous uses of this approach in the field of knowledge acquisition. It is evident that not many

[1] For a comparison of the external features of the two methods (the repertory grid technique and knowledge acquisition from a minimal set of examples), see [10].

researchers have used controlled experiments for this purpose. However, the few that appear in the literature do contain lessons from which the author's own design was able to benefit. These lessons, and their influence on the author's design, are discussed in this paper.

Experiments

This section discusses six experiments reported in the knowledge-acquisition literature.

- *Congruence of representation*

Proposing hypotheses based on Anderson's theory of skill acquisition [11], Lundell argues [6] that while novices store their expertise in declarative memory, or at the conscious level, experts do so in procedural memory, or at the tacit level. Lundell further argues that it ought to be easier to elicit rules from novices than from experts, and that it ought to be easier to obtain typical examples (or what he calls "prototypes") from experts than from novices.

In addition, Lundell conjectures that an artificial neural network (built using prototypes and exemplars obtained from an expert) ought to have greater diagnostic accuracy than a similar knowledge base derived from exemplars and prototypes that have been elicited from novices. Conversely, a set of rules elicited directly from a novice ought to have a higher diagnostic accuracy than a set elicited directly from an expert.

Lundell's "representational congruence" hypothesis asserts that if, for example, a rule-elicitation method is used, it will elicit primarily knowledge stored as rules in the mind of the expert. Lundell's representational and "elicitational congruence" hypotheses involve the following independent variables:

- Elicitation method.
- Expert's level of expertise.
- Knowledge representation in the knowledge base.

These variables are all controllable in an experiment. The dependent variable, which, Lundell argues, is a function of the variables listed above, is diagnostic accuracy of the knowledge base built using the knowledge elicited from a subject. To test his hypotheses, Lundell had to vary the controllable variables in turn, and record the effects on diagnostic accuracy. Taking several observations for each setting of each controllable variable allowed him to increase the reliability of his results. Of course, the subjects themselves are also variable (see, e.g., [3, 12]).

Lundell's experiment is essentially a two-group design, in which each subject fills in four different types of questionnaire. It used a random presentation order in an attempt at eliminating sequence effects.

Two of Lundell's questionnaires were aimed at eliciting rules directly. One he called the "direct rule" questionnaire, and the other the "decomposed rule" questionnaire. These two complemented each other in his subsequent creation of rule bases.

The two other questionnaires were aimed at eliciting examples, from which knowledge could be derived by some kind of machine learning. One of these questionnaires elicited a set of typical examples or cases; this one he called the "prototype elicitation" questionnaire. The fourth questionnaire, which he called the "exemplar questionnaire," consisted of a randomly generated set of undiagnosed hypothetical cases for the subjects to diagnose.

Using these questionnaires for knowledge acquisition appears to impair the external validity of Lundell's experiment. The antecedents and the consequents are given, whereas in practice it seems more usual for these to have to be elicited from the knowledge source by various methods. The considerable amount of knowledge acquisition which clearly went into the preparation of these questionnaires deserves to be acknowledged openly. Moreover, questionnaires are rarely used to acquire knowledge for knowledge-based systems (see, e.g., [13]).

Lundell used the completed questionnaires to build a number of expert systems, but little is said in his dissertation about this process. And without any assurance to the contrary, his readers are left wondering about the scope for introduction of errors at this stage. Still, perhaps this criticism is a bit unfair, because the graphical representation on his questionnaires seems capable of being easily transformed into production rules. In the case of his connectionist networks, it appears obvious that the exemplar and prototype data were simply coded as examples and used to train the networks in the diagnostic task.

Lundell's subjects emerged from his training with a range of levels of expertise in the diagnostic task. Some had become good at it, and others had learned to a lesser extent. Lundell classified his newly trained subjects as either skilled or unskilled. He set his criterion at the median test score, so that half the subjects are "unskilled" and the others "skilled." It appears to be an arbitrary distinction with little basis in theory and little rationale, save that of balancing the sizes of the two groups.

After basing his initial arguments on the theory that experts' skills reside at a tacit level while novices' skills are represented consciously, Lundell appears to make little use of this representational differential that would be expected to exist between his skilled group and his unskilled one.

Perhaps an improvement would have been to use an adaptive questionnaire to gather the same type of data. Under this approach, subjects would interact with a computer program that asks questions based on answers already given. By doing this, he would have introduced

some of the flexibility characteristic of real-world knowledge acquisition, while providing systematic and consistent recording of data.

By creating his own experts in a domain of his own making, Lundell may have sacrificed external validity, but at the same time he gained a ready-made set of test cases against which both the experts themselves and the elicited knowledge bases could be evaluated. He also limited the scope of the task to a size amenable to analysis and experimental control.

- *Thinking aloud*

Stevenson et al. [7] also did an experiment to test a hypothesis implied by Anderson's ACT* (adaptive control of thought) theory [11]. Their hypothesis was that their own method of knowledge acquisition would be more effective than "traditional" methods. They argue that it is wrong to assume that analysis of thinking-aloud protocols accurately unearths the knowledge contained in an expert's automatic productions. What thinking aloud is more likely to do, they argue, is to slow down and even distort the expert's actions. They argue that it is more effective to let the expert perform his task undisturbed except for the scrutiny of a videotape camera and recorder. At some later time, the expert can explain his actions while watching the videotape. These explanations can be used to generate production rules. Stevenson et al. call this method an "evaluation technique."

The experiment of Stevenson et al. tested their hypothesis by varying the acquisition-method treatments to which subjects were exposed. They used a two-group repeated-measures design, although one group (the experts) was very small (two subjects) compared with the other group (eight subjects). All subjects received all treatments, but in the same order (there was no attempt to correct for sequence effects by counterbalancing). But time (more than a day) was allowed between treatments, perhaps to allow the attenuation of any carry-over effects.

Stevenson et al. appear not to have taken the analysis of the data as far as Lundell did. They did not measure the diagnostic accuracy of derived knowledge bases. They did, however, employ a more qualitative approach than Lundell's bald statistical one. They examined the differences between the kinds of constructs that the experts produced and those that the novices produced.

But although Stevenson et al. assert that thinking aloud may be less effective than their evaluation technique, they fail to support this empirically. Or, more precisely, they appear not to have designed their experiment to test this.

- *Computer-assisted knowledge elicitation*

Deffner and Ahrens [8] were not comparing knowledge-acquisition methods; they were simply evaluating the single method embodied in a tool of theirs. This method involves having a domain expert enter rules in a formal language and, as a second stage, refine any ill-defined quantifiers used in the rules. According to [8], deferring the refinement solves the problem of experts "drying up" when they are interrupted and asked to be more precise about quantifiers.

Like Lundell, Deffner and Ahrens used an artificial domain and created experts in it by training their twenty-two subjects. The domain is nutritional prediction in a simulation of a person to be fed from a menu. During training, the subjects are free to display their tendency to explore the domain. This tendency is observed by tracing each subject's interactions with the training software.

Although apparently not so by design, Deffner and Ahrens' experiment is a two-group one. The groups were discovered by *post hoc* cluster analysis of some of the training interaction data. Both groups received the same treatment (elicitation method), but they also had what Deffner and Ahrens assume to be two different levels of expertise. One dependent variable is the accuracy of the generated knowledge base, and this is measured by testing the rules on the simulation. Other dependent variables are the number of rules elicited and the average number of attributes per rule.

Deffner and Ahrens do not say how many of their subjects fall into each group. Nor do they say how they treat the two subjects who do not "fall clearly into one of the two groups." They concede that their tool "may at first sight appear not to be very practical" [2, p. 359], and try to remedy this lack of external validity by suggesting where the use of the tool might fit in a series of knowledge-acquisition stages.

- *Elicitation efficacy*

Whereas Lundell [6] and Stevenson et al. [7] were testing hypotheses, Burton et al. [5] wanted to determine the relative efficacies and efficiencies of different knowledge-elicitation techniques. They wanted to be able to predict which methods would be most appropriate for which circumstances, so that builders of knowledge-based systems would have some empirical basis for their choices.

Burton et al. also stopped short of building knowledge bases, and therefore did not reach as far as measuring diagnostic accuracy. However, they did perform other kinds of evaluation on the elicited knowledge, which they coded as "pseudo-English production rules." In a subsequent experiment, these rules were each rated by the experts on a four-point scale ranging from true to false. Thus, they were able to compare (at least for some of their data) the overall quality of rules resulting from each elicitation technique.

In their experiment, Burton et al. had as independent variables the elicitation method and the expert's personality. They tried to keep the knowledge

representation constant. Their dependent variables were the amount of knowledge elicited per unit time, and the quality of elicited rules.

They also made the distinction between procedural and declarative knowledge. Indeed, they assert that two of their methods (protocol analysis and formal interview) are likely to elicit procedural knowledge, while the others (card sort and laddered grid) are likely to elicit declarative knowledge. But they were forced to conclude that their results did not support this assertion.

Although, like Lundell, they used students as subjects, Burton et al. did not create instant experts. Thus, their claim of expertise is more credible, especially in the light of Anderson's assertion [11] that it takes a long period of practice to create an expert. On the other hand, Burton et al. offer little proof of the subjects' expertise. Burton's subjects were not tested for skill level as Lundell's subjects were.

There is usually some danger of impairing external validity when university students are used as subjects in experiments (see, e.g., [14]). To get some idea of the effect of using students, Burton et al. followed their 1987 experiment with another—this time using "real" experts. The earlier results were vindicated (see [15]).

- *Knowledge engineer as a variable*

Adelman [3] did not build expert systems with the knowledge elicited from his 138 subjects. What he was trying to do was to determine the effect of two variables (knowledge engineer and elicitation method) on the "predictive accuracy" of the knowledge elicited.

He used two methods ("top-down" and "bottom-up") and six knowledge engineers in what he describes as a "2 × 6 factorial" design. However, he appears to have had some difficulty in specifying exactly how the knowledge engineers differed from one another. He finally decided to use the institution from which the knowledge engineer had received his or her training as the dimension on which to group them. With this grouping, he reduced his data to that of a 2 × 3 factorial design. Perhaps it would have been more meaningful to have used either the psychometric profiles of the knowledge engineers to find clusters (as did Deffner and Ahrens [8]), or some aspect of their experience.

One of Adelman's chief concerns was with the quality of a domain expert's expertise. Adelman argues that the expert is a factor in the quality of any elicited knowledge. But, as with his knowledge engineers, he appears not to have decided what attribute of the expert is the variable of concern. Yet there is a theoretical reason for focusing on skill level (see [11]). In addition, both Deffner and Ahrens and Burton et al. have found the expert's personality to be important. Thus, Adelman might have tried to vary these systematically. He did not.

- *Effects of training*

Agarwal and Tanniru [9] used a completely randomized single-factor design to compare unstructured interviewing with "a specific kind of structured interview." They did this to test four hypotheses about the relative efficacy and efficiency of the two methods of knowledge acquisition.

They did well to find as subjects thirty "expert practitioners who were responsible for [a capital budgeting/resource allocation] decision." The subjects were split into three groups of ten, and each group was given one of the three treatments. However, there is some doubt about the consistency with which the treatments were administered in the experiment.

The control group of experts had their knowledge elicited, via unstructured interviews, by what Agarwal and Tanniru call "experienced knowledge engineers." However, only some of these had any experience at eliciting knowledge for expert systems. The others were systems analysts who were experienced at interviewing. Agarwal and Tanniru do not say how many of these interviewers were used, but do admit to having been unable to find enough experienced knowledge engineers.

Novice knowledge engineers, unlike experienced ones, were apparently abundant. Agarwal and Tanniru were therefore able to take care to establish that the novice knowledge engineers all started with comparable lack of experience of the domain and the knowledge-acquisition methods to be used. However, the novice knowledge engineers were given training in only one of the methods (structured interviewing), and left to administer the other method (unstructured interviewing) without the benefit of any training.

As a comparison of two methods, the experiment of Agarwal and Tanniru appears therefore to have been biased toward one method. However, they did succeed in showing that knowledge engineers who receive training in a technique are likely to be more efficient and effective using it than those who try to apply a technique in which they have not been trained.

Conclusions

- *Lessons from the past*

Building a knowledge base can be viewed as the process depicted in **Figure 1**. There are several inputs into knowledge acquisition; various attributes of these inputs interplay to produce some acquired knowledge in a representation formalism. The knowledge and representation are clearly interrelated, with the latter being the form in which the former can exist in a knowledge base. This knowledge base itself exhibits qualities, such as diagnostic or predictive accuracy.

The qualities displayed by the resulting knowledge base are affected by all the variables that provide input to the

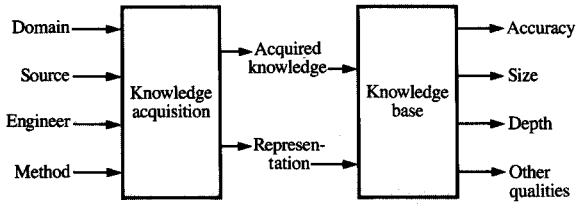

Figure 1

Factors and effects in building a knowledge base.

knowledge base. A major reason that researchers do empirical investigations is to see how these potential independent variables to the left affect the dependent variables to the right.

The understanding gained from experimenting with these variables is likely to bring more predictability to building knowledge bases, and allow knowledge engineers and planners to make choices based on more solid foundations than are available at present. The researchers discussed in this paper have experimented for various purposes: testing hypotheses, evaluating a method or a tool, and looking for correlations. The efforts of these researchers have highlighted various challenges.

For example, because of the need to be consistent and the constraint on time, it can be difficult to control the acquisition method. Some researchers have had to restrict the method [4] or use artificial ones [6]. In addition, there are several attributes of both the knowledge source and the knowledge engineer that deserve attention as variables in their own right (e.g., level of expertise, and personality).

Being sure that you have an expert (determining his or her level of expertise) appears also to be a widespread problem in experiment design. The means used to solve this problem have not been entirely convincing. Some researchers have selected people whose expertise they appear to think unquestionable (e.g., academics who specialize in the topic). They have also selected subjects whose novice status they appear to think indisputable (e.g., first-year university undergraduates). Others have trained and tested their own subjects; but these researchers appear to have difficulty deciding on appropriate criteria for expert and novice.

There are also problems with using accuracy as a measure of expert-system performance or knowledge-base quality. Each researcher defines accuracy in a different way, according to what is convenient for the experiment design. Even with a consistent definition of accuracy, there is still likely to be a problem. If the test cases are taken randomly from a very large set of historical records, the frequency distribution of certain attributes and classes is likely to have certain characteristics. However, if the test cases are exemplars, or even cases that an expert thinks interesting, the frequency distribution of attributes and classes is likely to be quite different. Thus, diagnostic accuracies cannot be meaningfully compared without an accompanying comparison of the source of the test cases.

- *Effects on a new experiment*

The need for consistent application of a knowledge-acquisition method to all subjects can be addressed by modeling the method in a tool, and eliminating the knowledge engineer altogether. However, the use of knowledge-acquisition tools to compare two methods can introduce a confounding variable: the user interface. The author's solution was to build a single tool (SCENIC) embodying the two methods: the repertory grid technique and knowledge acquisition from a minimal set of examples (KAMSE). The tool assumes that an analysis domain has been identified, and takes knowledge acquisition as far as the generation of a knowledge base and the validation of it through batch consultation. The two methods followed the stages shown in **Table 1**.

Where appropriate, routines are shared between the methods. Adhering to the Systems Application Architecture® (SAA™) Common User Access™ (CUA™) guidelines in designing the user interface of SCENIC helped to achieve consistency between the two methods. Menu bars are activated in a standard way, and function keys are used for analogous purposes between methods.

The hypotheses to be tested involved the variables shown in **Table 2** (the stages mentioned in the table are those from Table 1).

One of the criteria imposed on the problem domain to be used is that it had to be an area of knowledge for which experts could be readily found, and in large numbers, so that artificial experts did not have to be created by training. But it is not easy to find domains like this. Several possibilities were explored and rejected: spelling, words, grammar, and conversation. In the end, object identification turned out to be a suitable choice. It could be simple or complex, depending on the limits imposed on the domain. Moreover, it is an area in which there is no shortage of experts. It is an effective surrogate for many other kinds of diagnostic and classification knowledge domains.

The experiment used a repeated-measures within-subject design, in which subjects were randomly assigned to both knowledge-acquisition methods from a Latin square of treatment combinations. After being trained in the use of the relevant part of the tool, subjects then used it to elicit their own knowledge of a restricted object identification

Table 1 Knowledge-acquisition stages for the two methods implemented in SCENIC.

Stage	*Repertory grid technique*	*KAMSE*
1	Listing all elements that exemplify the domain classes	Identifying the classes that cases can belong to
2	Identifying constructs that distinguish elements from each other	Identifying the attributes (e.g., supply voltage) considered in deciding the class of a case, and listing the values (e.g., 3V, 5V, 24V) that each attribute can have
3	Rating all elements on each construct elicited	Describing, without repetition, examples of all classes in terms of attribute descriptors and values
4	Using machine induction to find regularities and distill the grid into a knowledge base	Using machine induction to find regularities and distill the examples into a knowledge base
5	Classifying a set of exemplars and using these to evaluate the knowledge base produced in the previous step	Classifying a set of exemplars and using these to evaluate the knowledge base produced in the previous step

domain. The tool generated knowledge bases by machine induction.

To evaluate each knowledge base, the tool generated 32 exemplars (from random attribute values used in the knowledge base) to be classified by the subject. The classified exemplars were used by a batch consultation process within the tool, which summarized the performance of the knowledge against the expected results. At every stage of the experiment, the tool collected data relevant to the variables being measured. Subjects were allowed a week between treatments to allow carry-over effects to be attenuated.

Not all of the problems are solved in the author's design, but being aware of them helped make the design better than it would otherwise have been.

- *Future directions*

The results of the analysis of data gathered in the experiment will allow conclusions to be drawn about the effects of the two methods. The results may also highlight opportunities for improving the efficiency or efficacy of either method, or of combining the best parts of the two to form a new method.

The controlled experiment still appears to be a promising approach to investigating the relationships at work in knowledge acquisition. As present and future researchers respond to the challenges posed by their predecessors, the quality of research design and the value of the findings are likely to be enhanced.

Acknowledgments

I am indebted to N. A. D. (Con) Connell, Clare Jackson, Jonathan Klein, Mike Vale, Chris Woodford, anonymous reviewers, and my wife Helene for reading drafts at various stages of the development of this paper. They all made useful suggestions which helped shape this final version.

Table 2 Variables involved in the hypotheses.

Hypothesis	*Independent variables*	*Dependent variables*
1	Method	Effort at stage 1
2	Method	Effort at stage 2
3	Method	Effort at stage 3
4	Method	Classification accuracy
5	Method	Total effort (stages 1 to 6)

References

1. R. S. Michalski and R. Chilausky, "Knowledge Acquisition by Encoding Expert Rules Versus Computer Induction from Examples: A Case Study Involving Soybean Pathology," *Int. J. Man-Machine Studies* **12**, 63–87 (1980).
2. J. R. Quinlan, "Induction of Decision Trees," *Machine Learning* **1**, 81–106 (1986).
3. Leonard Adelman, "Measurement Issues in Knowledge Engineering," *IEEE Trans. Systems, Man, & Cybernetics* **19**, No. 3, 483–488 (1989).
4. Mike Burton and Nigel Shadbolt, "Experiments in Knowledge Elicitation," *AISB Quarterly*, Part 65 (Summer Edition), pp. 11–12 (1988).
5. Mike Burton, Nigel Shadbolt, A. P. Hedgecock, and G. Rugg, "A Formal Evaluation of Knowledge Elicitation for Expert Systems: Domain 1," *Research and Development in Expert Systems IV*, S. Moralee, Ed., Cambridge University Press, Cambridge, England, 1987.

6. James Walfred Lundell, "Knowledge Extraction and the Modelling of Expertise in a Diagnostic Task," Ph.D. dissertation, University of Washington, Seattle, 1988.
7. R. J. Stevenson, K. I. Manktelow, and M. J. Howard, "Knowledge Elicitation: Dissociating Conscious Reflections from Automatic Processes," *People & Computers IV*, D. M. Jones and R. Winder, Eds., Cambridge University Press (on behalf of British Computer Society), Cambridge, England, 1988.
8. G. Deffner and R. Ahrens, "On the Use of Formal Language and Ill Defined Quantifiers in Knowledge Acquisition," *Proceedings of the Human Factors Society 33rd Annual Meeting. Perspectives 1989*, Vol. 1, pp. 356–360.
9. Ritu Agarwal and Mohan R. Tanniru, "Knowledge Acquisition Using Structured Interviewing: An Empirical Investigation," *J. Management Info. Syst.* **7**, No. 1, 123–140 (1990).
10. Clive Nicholson, "Learning Without Case Records: A Mapping of the Repertory Grid Technique onto Knowledge Acquisition from Examples," *Expert Syst.* **9**, No. 2, 79–87 (1992).
11. John R. Anderson, "Acquisition of Cognitive Skill," *Psychological Rev.* **89**, No. 4, 369–406 (1982).
12. Alphonse Chapanis, *Research Techniques in Human Engineering*, Johns Hopkins Press, Baltimore, 1959.
13. Margaret Welbank, "An Overview of Knowledge Acquisition Methods," *Interacting with Computers* **2**, No. 1, 83–91 (1990).
14. John Jung, "Current Practices and Problems in the Use of College Students for Psychological Research," *Canadian Psychologist* **10**, No. 3, 280–290 (1969).
15. A. M. Burton, N. R. Shadbolt, G. Rugg, and A. P. Hedgecock, "The Efficacy of Knowledge Elicitation Techniques: A Comparison Across Domains and Levels of Expertise," *Knowledge Acquisition* **2**, 167–178 (1990).

Received September 24, 1990; revised manuscript received August 6, 1992; accepted for publication August 17, 1992

Clive N. Nicholson *IBM United Kingdom Laboratories Ltd., Hursley Park, Winchester SO21 2JN, United Kingdom (NICHO at WINVMJ).* Mr. Nicholson is an information developer with IBM United Kingdom Laboratories Ltd. in Hursley, England. He received a B.Sc. in electrical engineering from the University of the West Indies and an M.B.A. degree from the University of Bath. Mr. Nicholson is currently a doctoral candidate at the University of Southampton, where his research is in knowledge acquisition for knowledge-based systems.