VALIDATION OF EXPERT SYSTEMS:

PERSONAL CHOICE EXPERT --

A FLEXIBLE EMPLOYEE BENEFIT SYSTEM

Michael C. Sturman
George T. Milkovich

CENTER FOR ADVANCED HUMAN RESOURCE STUDIES

NEW YORK STATE SCHOOL OF INDUSTRIAL AND LABOR RELATIONS

CORNELL UNIVERSITY

WORKING PAPER #92-33

This paper has not undergone formal review or approval of the faculty of the ILR School. It is intended to make results of Center Research, conferences, and projects available to others interested in human resource management in preliminary form to encourage discussion and suggestions.

ABSTRACT

A method for validating expert systems, based on psychological validation literature and Turing's "imitation game," is applied to a flexible benefits expert system. Expert system validation entails determining if a difference exists between expert and novice decisions (construct validity), if the system uses the same inputs and processes to make its decisions as experts (content validity), and if the system produces the same results as experts (criterion-related validity). If these criteria are satisfied, then the system is indistinguishable from experts for its domain and satisfies Turing's "imitation game."

The methods developed in this paper are applied to a human resource expert system, *Personal Choice Expert* (PCE), designed to help employees choose a benefits package in a flexible benefits system. Expert and novice recommendations are compared to those generated by PCE. PCE's recommendations do not significantly differ from those given by experts. High inter-expert agreement exists for some benefit recommendations (e.g. Dental Care and Long-Term Disability) but not for others (e.g. Short-Term Disability and Life Insurance). Insights offered by this method are illustrated and examined.

INTRODUCTION[1]

Increasingly, expert systems offer the promise of becoming managerial decision aids. Much of the research on managerial expert systems focuses on their feasibility [2] [7], case studies describing their applications [14] [24], evaluations of alternative methods of knowledge acquisition and representation [12] [13] [15] [25] [27], and other issues related to user interfaces [11].

One critical, yet relatively under-analyzed issue involves the validation of expert systems. Validity refers to the degree to which a system makes decisions that correspond to decisions made by experts. The literature on expert system validation emphasizes the need for validation [3] [17], offers a framework for understanding validation [18], and reports attempts to validate expert systems [5] [21]. A substantial body of knowledge on validation exists beyond the domain of expert systems, particularly in the validation of psychological tests used to predict future behaviors or assess behavioral attributes [8] [16] [22].

Our study draws upon the psychological validation research as well as the earlier work on expert system validation. Building on this earlier work, we offer an approach for system validation for decisions without predefined correct answers and where expertise is not well mapped and often implicit. We apply this approach to a human resource expert system developed to aid employees in making decisions in a flexible benefit environment. Although this study uses a human resource expert system, it should also be relevant to other systems where decision makers face similar conditions.

THE *PERSONAL CHOICE EXPERT* SYSTEM

Personal Choice Expert (PCE) is an expert system designed to help employees choose among benefits alternatives from a flexible benefits plan. The system was designed and field tested with employees at a major computer manufacturer [9].

INSERT FIGURE 1 ABOUT HERE

The need for the system sprung from three sources of pressure. First, under the flexible benefits program, employees faced a highly complex decision task. As figure 1 depicts, for a relatively simple flexible benefits plan, employees confront a large number of possible benefit combinations. (e.g. 4 health plans * 4 dental plans * 4 short term disability plans * 4 long term disability plans * 4 life insurance plans * 4 accidental death and dismemberment plans * 4 spouse life insurance plans exceeds 16,000.) In more complex plans, such as for the one used by the company in this study, employees face decisions about even greater possible combinations of benefits. Additionally, employees receive information and advice from numerous sources. These include enrollment requirements from the company, information booklets containing details on the coverage and levels of the various benefits and their options, and advice from benefit counselors, their coworkers, and others. Combined, the amount of data and advice coming from numerous sources probably makes employees relatively uncertain and uneasy about their decisions. Hence, the expert systems was intended to offer employees some expert advice and guidance for making their benefit decisions.

The second reason for developing the expert system was costs. The manufacturer reported that it employed several benefits experts whose duties included counseling employees in making their benefit decisions. The company felt, though, that the counseling task was taking up too much of their benefit experts' time. For example, one benefits expert told of long queues of employees outside her office during benefits enrollment time, all having questions about making benefits choices. She even had "to sneak out the back door of her office" when she wanted to eat lunch. Clearly, some less costly aid would help alleviate the demand on the benefit professionals' time, and it would allow them more time to devote their expertise to higher priority tasks, such as attempting to contain escalating benefit costs rather than counseling employees.

Finally, the expert system serves to empower employees by acting as their decision aid. By offering employees this aid, the employer attempts to signal that it is really the employee's responsibility to make informed choices: to select the combination of benefits that best fits each person's circumstances. The benefits options and the decision aids are the employer's responsibility but the actual decisions rest with the employees.

To build the system, the design team became familiarized with the terminology and task by reviewing the company's benefits literature and through discussion with benefits managers. This entailed reading the company's benefits manuals, practicing benefits selection using the company's plan, and utilizing the company's spread-sheet like program as a benefits selection aid. (See [23] for a more detailed description of the design process and detailed specifications of PCE.)

A typical session of an employee using *PCE* involves four phases: the direction, question, benefits calculation, and the output phases. The direction phase, including the introduction screen, informs the user about what the program does and how to use it. During this phase, the program includes multiple disclaimers so that employees know that the program provides recommendations, not a definite best answer. Once the user reads all the directions, the program moves to the question phase.

INSERT FIGURE 2 ABOUT HERE

In the question phase, the user proceeds through the screens sequentially, answering all the appropriate questions on each screen. For some screens, the program can produce multiple versions; however, users will only see one version of each window. For example, if users indicate that they are not married, the program will not ask for spouse's income, or level of spouse's alternate life insurance.

When the employee enters all the necessary information, the program advances to the benefits calculation phase. Here, the program computes the benefits recommendations and the corresponding costs. Once PCE determines its recommendations, the system provides the benefits recommendations on a screen that mimics the actual benefits enrollment form that the company requires the employee to complete.

PCE is intended to aid employees in making their benefits decisions. Thus, it includes many features to help make the system user friendly. The program is not intended to limit the freedom that employees have regarding their benefits decisions. Thus, employees are ensured that the system only offers recommendations and guidance to them. It is intended to act as their personal decision aid. Further, the system can serve as a simulation device so employees, if they desire, can ignore the system recommendations completely, or they can repeatedly enter different scenarios to "see" the expert recommendations under changing conditions.

VALIDATION

Validity is the degree to which inferences and decisions are justified by evidence [22]. For expert systems, this entails analyzing the decision-making capabilities of a system [3] [18].

Much of the psychological literature on validation focuses on ascertaining the validity of devices. These devices generally predict an individual's behavior in a given context (e.g. in a job or school). Expert systems differ from this context in that they mimic the decisions someone deemed to be an "expert" makes. Thus, although similarities exist between psychological and expert system contexts, the specific methods must be adapted so that expert systems can be evaluated [3] [5] [17] [18].

The validation process consists of three methods; they include construct validation, content validation, and criterion-related validation [1] [6] [10] [17] [18]. The approaches employed in expert system validation to date fall within these three areas.

Construct validation attempts to verify that a measurement device actually measures what it purports to measure [6] [16]. A construct can be said to be valid only if it can be shown that the operation produces results that agree with those achieved with alternative operationalizations of the same construct, and conversely, that it produces different results than do operations thought to tap different constructs [20]. For expert systems, this means that the system must perform like an expert, and expert decisions must differ from novice decisions. Although researchers have discussed the possibility that no clear answer may exist for certain situations [3], a method of analyzing expert reliability has not been reported in the literature. Using this approach, however, construct validity for an expert system could be inferred through high expert reliability and significant differences between expert and novice recommendations.

Content validation addresses the extent to which a particular measure represents the content universe of the property being measured [10]. For example, in industrial psychology this form of validation verifies that a device samples all the critical behaviors and knowledge that encompasses the process in question [1] [16] [22]. For expert systems, this means the system's logic mimics the process experts use to make their decisions. Note that content validation is not concerned with *what* the expert system decides; it only attempts to ascertain if the process and logic that the system uses to reach its decisions are similar to those used by experts.

One test of content validity involves comparing the models derived during knowledge acquisition to the methods used by the experts. If experts who were not involved in the system's construction agree with the models specified in the system, then the content of the system's decision process would appear valid. However, due to the resources and time required to take this approach, this tactic has generally not been used to evaluate expert systems.

A second test of content validity, used as part of the validation process for the MYCIN expert system [3] [21], involves determining which of the questions that the system asks are appropriate for solving the problems facing the system. Additionally, experts can be asked if the system neglected to ask any pertinent questions. Through this method, it is possible to infer the nature of the inputs to the expert system. This should give some insight into the information the system uses to make decisions.

In essence, content validation applied to expert systems entails examining whether the system collects all appropriate data (i.e. it uses the same variables that experts do), excludes inappropriate data (i.e. does not use variables which experts do no use), and that it uses the appropriate data like an expert to reach its decisions. The degree to which these three criteria are satisfied indicates the degree to which the system's content can be said to be valid.

Criterion-related validation measures the statistical relationship that exists between a given index and a criterion score [1] [22]. An example is the ability of an employment test to predict future performance or turnover. The actual processes of the system are not investigated; rather, the system is treated like a "black box" and only the inputs (employment test) and outcomes (performance) are analyzed.

In this form of validation, a system would be evaluated by comparing system decisions to the correct answers. A valid system would produce the same answers as the experts. In many decision situations, particularly those involving employees, a single correct answer simply does not exist. Rather, several possible answers may be "correct." When there is no "gold standard" to serve as a perfect answer, the experts' decisions usually serve as the standard. In our study, the correct choices given perfect future knowledge is simply not available; hence, expert recommendations, based on information available at the time of the decision, served as the correct decisions. A second problem with criterion-related validation to expert systems is that experts do not always agree with each other. System validity will thus be limited by the variability in expert decisions. Construct validation may show that expert-quality decisions do not exist. If relatively low variability in expert decision occurs, there may still exist varying levels of expertise. Thus, it may be necessary to group expert recommendations and to determine how system recommendations compare to the experts' collective judgments.

This portion of criterion-related validation overlaps with construct validation. This is because reliability of expert responses is necessary for there to be high agreement when comparing the system to the grouped experts' responses. Although this information can be obtained without considering the construct of expertise, some degree of reliability of decisions is essential if an expert system is going to be a useful aid for employees.

In essence, criterion-related validation measures the relationship between the decisions developed by the system and decisions developed by human experts [17]. Within that task, both the reliability of the expert decisions and the validity of the system must be determined. A high degree of criterion-related validity should thus be demonstrated through high agreement between the expert system and the experts.

In summary, the concepts of construct, content, and criterion-related validation, while commonly used to evaluate psychological tests, can also be applied to validating expert systems. However, these methods must be adopted to be applied to expert systems. For construct validation, this means analyzing the construct of expertise in general. Specifically, this entails examining both the differences between experts and novices, and the degree of agreement among experts. Content validation of expert systems involves analyzing the inputs to the system and evaluating the variables and logic the system uses to develop its answers. Criterion-related validation calls for comparing the decisions made by the system to the decisions made by the experts. Combined, these three forms of validation can serve as a means to evaluate validity of an expert system.

THE TURING TEST FRAMEWORK

The basic premise of the Turing Test is that a "thinking machine" has been successfully developed if an individual cannot determine if a conversation or task is being performed by a machine or a person [26]. When this differentiation is impossible, artificial intelligence has been obtained. The Turing Test has been applied to the evaluation of expert systems [18]. Using Turing's "imitation game", if it is impossible to tell from a set of decisions which were made by experts and which were made by an expert system, then the system is valid. The Turing test and validation can be integrated by determining the following: one, if expertise exists in the given domain (construct validation); two, if the system uses the correct data and decision processes to reach each of its decisions (content validation); and three, if the system produces answers like an expert (criterion-related validation). A system's validity can be inferred from the degree to which it exhibits these characteristics.

EVALUATION OF PERSONAL CHOICE EXPERT

Method and Design

To evaluate the validity of *PCE*, the authors developed a questionnaire consisting of 10 actual employee scenarios. The scenarios were based on actual employee information and included information on family status, levels of alternate benefit coverage, spending information, and other descriptors that may be relevant to the decision process (e.g. spouse's level of income, other sources of income, hours employee exercises per week). An example is shown in figure 3. Experts and novices filled out the questionnaire by making benefits decisions for each scenario. Subjects identified the variables they used in their decisions and described any other information they would have liked when making their recommendations.

INSERT FIGURE 3 ABOUT HERE

Thirteen people were identified by the company officials as benefits experts, eight of whom completed the questionnaire. This yields a response rate of 62%. The experts mean tenure at the company was 10 years. Average tenure as a benefits counselor was 7 year. The sample of novices consisted of fifteen college students who had no previous knowledge of the firm's benefits program.

Measures of agreement in this study are calculated using the Kappa statistic. Kappa, as developed by Cohen [4], is a measure of agreement across raters when the data is non-ordinal:

$$K = (\pi_o - \pi_e) / (1 - \pi_e)$$

π_o = Frequency of agreement

π_e = Expected frequency of agreement due to chance.

Kappa equals 1 when there is perfect agreement, and it equals 0 when the agreement is equal to that expected by chance.

The validity of the system is determined by evaluating the extent to which the system satisfies the conditions for validity for each of the three types of validity discussed above through the framework of the Turing Test.

Analysis of Construct Validity

To infer construct validity, inter-expert agreement and expert to novice agreement must be analyzed. The degree of agreement among experts' benefits recommendations were analyzed. A significance test was performed to determine if expert agreement is above 50% (kappa > 0.50). A similar test was performed to determine if novices' recommendations were significantly different from those made by experts. Thus, a test was performed to see if intra-novice agreement was significantly below intra-expert agreement. These analyses were performed on each benefit category.

The results of these tests are in figure 4. Experts agreed on over 50% of their recommendations (p <= 0.0001) for employees' Dental and Long-Term Disability insurance coverages. Expert agreement was significantly below 50% for Short-Term Disability, Employee Life, and Accidental Death and Dismemberment insurance (p <= 0.001). For the benefit types of Health Care (mean agreement = 53%) and Spouse Life Insurance (mean agreement = 45%), neither benefit types tested significantly above or below 50% agreement. Analysis of intra-novice agreement revealed that novices had agreement below 50% (p <= 0.01) for the benefit categories of Health Care, Short-Term Disability, Life Insurance, and Accidental Death and Dismemberment Insurance. Intra-novice agreement was significantly above 50% (p <= 0.01) for Dental Care and Long-Term Disability decisions. As with experts, Spouse Life Insurance (mean agreement = 50%) did not test significantly above or below the 50% mark.

INSERT FIGURE 4 ABOUT HERE

Tests on the difference between novice and expert recommendations also yielded mixed results. Expert recommendations differed significantly from novice recommendations for Health Care decisions (p <= 0.05) and Long-Term Disability decisions (p <= 0.01). For all other benefit types, no significant difference was discovered.

The following inferences can be drawn from these results. First, high inter-expert agreement exists only for Dental and Long-Term Disability Insurance recommendations. There is low inter-expert reliability when they make recommendations for Short-Term Disability Insurance, Life Insurance, and Accidental Death and Dismemberment Insurance. It seems, therefore, that the construct of expertise is not uniform over all benefits types. Perhaps some of the firm's benefits experts are more knowledgeable in certain types of benefits than others. Or perhaps, there may be no agreement among experts at all for certain benefits types. Perhaps if a larger number of experts existed then agreement may have increased by virtue of large N size. But under conditions where only a few experts exist, our results show that no uniform reasoning was exhibited by the company's experts for decisions on Short-Term Disability Insurance, Life Insurance, and Accidental Death and Dismemberment Insurance.

Analysis of Content Validity

The content validity of PCE was analyzed by comparing the inputs used by experts to those used by the expert system. The expert system uses 23 variables as input for recommendations on the seven benefit types under consideration. Of these 23 variables, 11 were cited as necessary by all the experts. Six additional variables were used by seven of the eight experts. The rest of the variables were used by at least half of the experts in the study.

It is still necessary, though, to determine if experts use variables which are not used by the system. A total of 14 other variables, i.e. variables not used by the system, were either included in the survey to see if experts used them or written in the space provided by the experts for this purpose. Ten of these variables, all of which were written in by one of two experts, were only cited by a single person. Four other variables, though, were cited by more than one person.

These four variables included gender, if the employee has a spouse at the company, if the employee is a smoker, and the number of hours per week that the subject exercises. Gender was only cited as relevant by three of the eight experts. Thus, it would appear that the system is not flawed for not using gender. However, seven experts thought that knowing if the employee agreement exists only for Dental and Long-Term Disability Insurance recommendations. There is low inter-expert reliability when they make recommendations for Short-Term Disability Insurance, Life Insurance, and Accidental Death and Dismemberment Insurance. It seems, therefore, that the construct of expertise is not uniform over all benefits types. Perhaps some of the firm's benefits experts are more knowledgeable in certain types of benefits than others. Or perhaps, there may be no agreement among experts at all for certain benefits types. Perhaps if a larger number of experts existed then agreement may have increased by virtue of large N size. But under conditions where only a few experts exist, our results show that no uniform reasoning was exhibited by the company's experts for decisions on Short-Term Disability Insurance, Life Insurance, and Accidental Death and Dismemberment Insurance.

Analysis of Content Validity

The content validity of PCE was analyzed by comparing the inputs used by experts to those used by the expert system. The expert system uses 23 variables as input for recommendations on the seven benefit types under consideration. Of these 23 variables, 11 were cited as necessary by all the experts. Six additional variables were used by seven of the eight experts. The rest of the variables were used by at least half of the experts in the study.

It is still necessary, though, to determine if experts use variables which are not used by the system. A total of 14 other variables, i.e. variables not used by the system, were either included in the survey to see if experts used them or written in the space provided by the experts for this purpose. Ten of these variables, all of which were written in by one of two experts, were only cited by a single person. Four other variables, though, were cited by more than one person.

These four variables included gender, if the employee has a spouse at the company, if the employee is a smoker, and the number of hours per week that the subject exercises. Gender was only cited as relevant by three of the eight experts. Thus, it would appear that the system is not flawed for not using gender. However, seven experts thought that knowing if the employee had a spouse who worked at the company was important, and six felt that knowing if the individual was a smoker and the number of hours of exercise was important.

INSERT FIGURE 5 ABOUT HERE

Although it is impossible to determine what the effects of these variables were with the data from this study, it is possible to make some estimates. First, data on spouse employment at the company was probably used by experts because recommendations for the family would be based on benefits choices by both individuals. The expert system solicits similar information through questions about alternative spousal coverage.

The other two variables which experts use but the system does not are if the employee smokes and the number of hours of exercise that the employee participates in per week. It is likely that these variables were used as a signal of health for determining health coverage, disability insurance, and/or life insurance. PCE does not use this information for two reasons. First, PCE uses projected medical expenses rather than proxies as the signal for health when determining its health care choice recommendations. Second, the system bases choices of the other insurances on the financial requirements of the employee.

Given the sample size of experts, it is difficult to determine if omitting these two variables signals deficient content. However, we judge that PCE, overall, collects very similar information to that used by the experts. Thus, although some content differences do exists, arguably there is a high degree of content validity in the expert system.

Analysis of Criterion-Related Validity

The next phase of validation is to test whether the system actually produces answers that differ from experts. Decisions generated by the expert system for all seven benefit categories were compared to those made by experts.

Within each benefit category, the system's recommendations were compared to the decisions made by each of the experts. The level of agreement was calculated by comparing the system's recommendation for a given employee and a given benefit category to the recommendations from each of the experts. Expert recommendations were not averaged because the data was nominal, and modal data was not used because it would not have differentiated between low and high intra-expert agreement.

System-expert agreement was then compared to the intra-expert agreement. The results show there is no significant difference (at alpha = 0.10) between the system-expert comparisons and the intra-expert comparisons, leading us to conclude that the system generates results indistinguishable from the experts. These results are reported in figure 6. Although doubts have been raised regarding the utility of "expert" decisions, the expert system, PCE, makes recommendations like an expert for all seven benefit types.

INSERT FIGURE 6 ABOUT HERE

CONCLUSIONS

Validation is a critical step in the knowledge engineering process. Much of an expert system's utility hinges upon the validity of the recommendations that it provides to the decision maker. Thus, methods of expert system validation must be developed and tested. The concepts of construct, content, and criterion-related validation can be applied to the task of validating expert systems, but not without some modification. These types of validity can be modified for expert systems through the framework of the Turing Test. This entails determining how much the system replicates the decision processes and results of experts in addition to analyzing the differences in responses between experts and novices. Or more generally, when validating an expert system, it is necessary to continuously question if the system is acting like a human expert.

Our analysis of this method using the human resource expert system, *Personal Choice Expert*, revealed that the company's experts did not uniformly agree on recommendations across all benefit types. High intra-expert agreement exists for some benefit decisions, specifically Dental Care Insurance and Long-Term Disability Insurance; however, there is low agreement on Short-Term Disability Insurance, Life Insurance, and Accidental Death and Dismemberment Insurance decisions. The intra-expert agreement for Health Care Insurance and Spouse Life Insurance choices is less clear. Because expert reliability serves to limit validity, it is impossible for PCE to be a perfectly valid device for all benefit types.

On the other hand, PCE decisions do not significantly differ from expert recommendations. Thus it would appear that the system produces responses like experts for all benefit categories. Additionally, for four of the seven benefit types, the system-expert agreement is greater than the average level of intra-expert agreement, implying that the expert system may be able to provide better advice than a single human expert could for some benefit types.

It is obvious that the system is not perfectly valid; however, it would have been naive to have expected perfect expert agreement, complete randomness of novice decisions, and perfect agreement between experts and the expert system. It seems clear that there are significant issues for the company's experts to resolve with regard to their decisions about Short-Term Disability Insurance, Life Insurance, and Accidental Death and Dismemberment Insurance. PCE does perform like the sample of experts for all benefit types; however, the process of construct validation shows that company "experts" may be expert in name only because the "experts" and "novices" perform similarly for certain benefit types. Thus, while we can conclude that PCE performs like the benefits counselors and that it performs like an expert for some benefit types, PCE's expertise for some types of benefits decision making is open to question because of the limits imposed by the reliability of the company's experts. This suggests that for certain categories of management decisions, more research on what constitutes an expert is required.

Thus, while it is essential to evaluate the decision making ability of an expert system before its implementation, it is also necessary to examine how expertise is defined. Further, it is necessary to study how expert systems like PCE will be used, what effects they may have on employee behaviors and attitudes, and the effects of the system on the decision task. Steps have been taken to formulate methods for ascertaining system validity, but further research needs to be performed to understand the issues of expert definitions and system implementation.

[1] We wish to thank the Center for Advanced Human Resource Studies for its financial support of this project. We also thank John Boudreau, John Hannon, and Marty Wells for their suggestions and criticisms.

REFERENCES

[1] Arvey, Richard D.; Faley, Robert H. *Fairness in Selecting Employees*. New York: Addison-Wesley, 1988.

[2] Briggs, Steven; Doney, Lloyd D. Eight HR Expert Systems Now. *Computers in Personnel,* 1989, Fall, 10-14.

[3] Buchanan, Bruce G.; Shortliffe, Edward H. *Rule-based expert systems: The MYCIN experiments of the Stanford heuristic programming project.* Reading, MA: Addison-Wesley, 1984.

[4] Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, 20, 37-46.

[5] Cohen, Paul R.; Howe, Adele E. Toward AI research methodology: Three case studies in evaluation. *IEEE Transaction on Systems, Man, and Cybernetics*, 1989, 19(3), 634-646.

[6] Gatewood, Robert D.; Feild, Hubert S. *Human Resource Selection.* Philadelphia, PA: Dryden Press, 1990.

[7] Goul, Michael; Tonge, Fred. Project IPMA: Applying decision support system design principles to building expert-based systems. *Decision Sciences*, 1987, 18, 448-467.

[8] Guion, Robert M. Personnel Assessment, Selection, and Placement. In Marvin D. Dunnette & Leaetta M. Hough (Eds.) *Handbook of Industrial and Organizational Psychology (Volume 2).* Palo Alto, CA: Consulting Psychologists Press, 1991.

[9] Hannon, John M.; Milkovich, George T.; Sturman, Michael C. The effect of a flexible benefits expert system on employee decisions and satisfaction. Presented at the Academy of Management, Las Vegas, NV, August 1992.

[10] Kerlinger, Fred N. *Foundations of Behavioral Research.* New York: Holt, Rinehart & Winston, 1973.

[11] Lamberti, Donna M.; Wallace, William A. Intelligent interface design: An empirical assessment of knowledge presentation in expert systems. *MIS Quarterly*, 1990, September: 279-308.

[12] Lenk, Peter J.; Floyd, Barry D. Dynamically updating relevance judgments in probabilistic information systems via users' feedback. *Management Science*, 1988, 34 (12), 1450-1459.

[13] Liang, Ting-Peng. A composite approach to inducing knowledge for expert system design. *Management Science*, 1992, 38(1), 1 - 17.

[14] McMillan, Claude. An Expert Scheduler for Part-Timers. *Computers in Personnel*, 1989, Fall, 22-26.

[15] Mendel, Max B.; Sheridan, Thomas B. Filtering information from human experts. *IEEE Transactions on Systems, Man, and Cybernetics*, 1989, 36(1), 6-16.

[16] Nunnally, Jum C. *Psychometric Theory.* New York: McGraw-Hill, 1978.

[17] O'Leary, Daniel E. Validation of expert systems— with application to auditing and account expert systems. *Decision Sciences*, 1987, 18, 468-486.

[18] O'Leary, Daniel E. Methods of validating expert systems. *Interfaces*, 1988, 18(6), 72-79.

[19] Rich, Elain; Knight, Kevin. *Artificial Intelligence.* New York: McGraw-Hill, 1991.

[20] Sackett, Paul R; Larson, James R. Jr. *Research strategies and tactics.* In Marvin D. Dunnette & Leaetta M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology (Volume One).* Palo Alto, CA: Consulting Psychologists Press, 1990.

[21] Shortliffe, Edward H. *MYCIN: A rule-based computer program for advising physicians regarding antimicrobial therapy selection.* PhD dissertation, Stanford University, CA, 1974.

[22] Society for Industrial and Organizational Psychology. *Principles for the Validation and Use of Personnel Selection Procedures (Third Edition).* College Park, MD: Society for Industrial and Organizational Psychology, 1987.

[23] Sturman, Michael C.; Milkovich, George T.; Hannon, John M. Development of a Human Resources Expert System: *Personal Choice Expert.* Working Paper, Center for Advanced Human Resource Studies, School of Industrial and Labor Relations, Cornell University; 1992.

[24] Sviokla, John J. An Examination of the Impact of Expert System on the Firm: The Case of XCON. *MIS Quarterly*, 1990, June, 127-140.

[25] Tou, Julius T. Knowledge engineering revisited. *International Journal of Computer & Information Sciences,* 1985, 14 (3), 123-133.

[26] Turing, A. M. Computer machinery and intelligence. In Edward A. Feigenbaum & Julian Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill, 1963.

[27] Wright, George; Ayton, Peter. Eliciting Modelling Expert Knowledge. *Decision Support Systems*, 1987, 3(1), 13-26.

Figure 1

The Flexible Benefits Decision

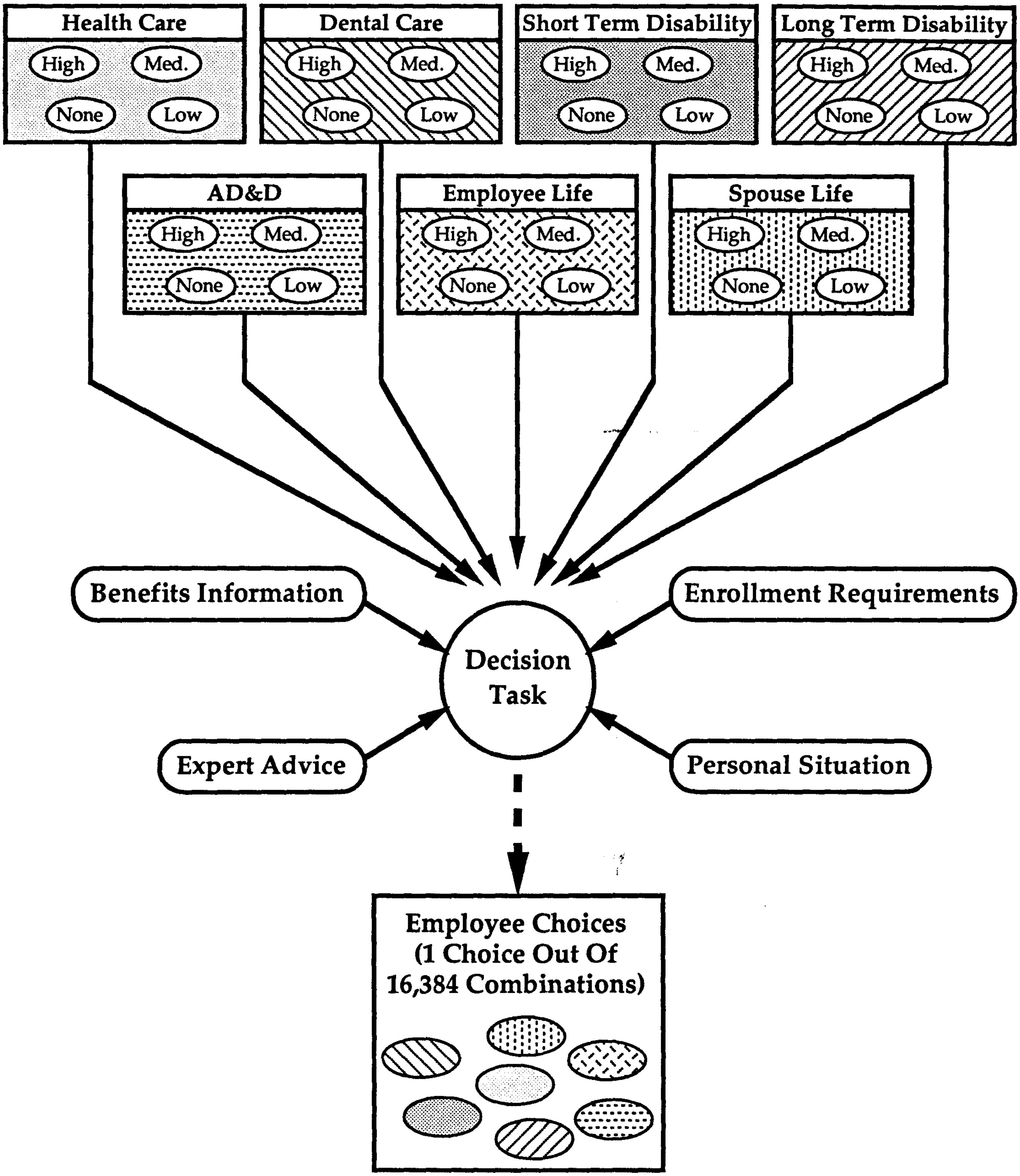

Figure 2

Using *Personal Choice Expert*

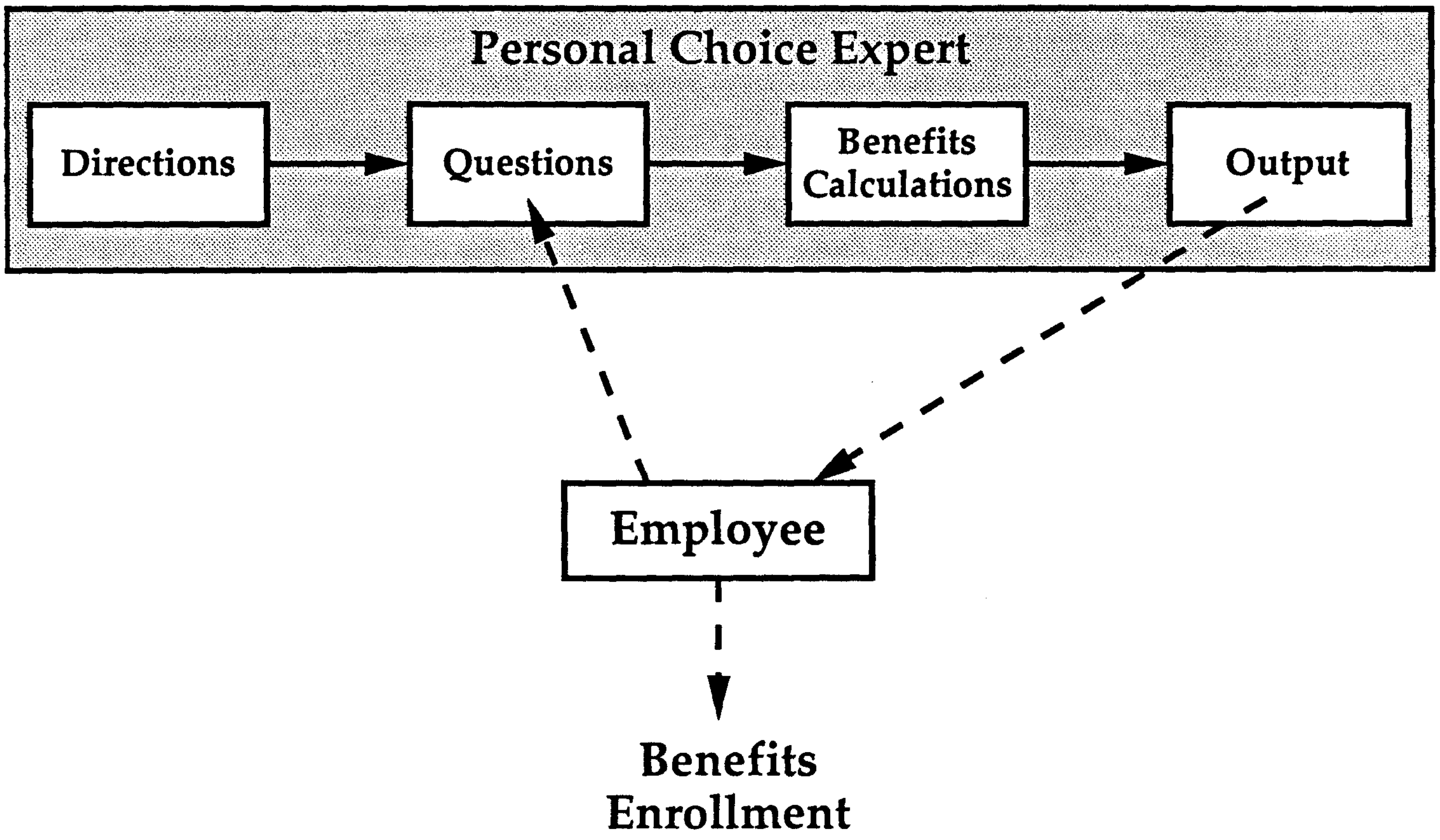

Figure 3

Sample Employee Scenario

Personal Information	Age: 32 Sex: Female Marital Status: Married Number of Legal Dependents: 1 Number of Special Dependents: 0
Income Information	Base Pay from NCR: $25,000 Total Pay from NCR: $27,500 Additional Income (outside of NCR): $3000 Savings: $4200 Spouse's Income: $25,000 Flexible Benefits Credits: 3520 Amount Saves per Year: $3000 Amount Spends per Year on Non-essentials: $6000
Alternate Insurance Coverage Information	Covered by a non-NCR Short Term Disability Plan: No Covered by Other Life Insurance Plan: Yes Value of Other Life Insurance Plan: $100,000 Covered by Other Dental Plan: No Spouse Covered by His or Her Own Life Insurance Plan: Yes Value of Spouse's Other Life Insurance Plan: $180,000 Equivalent of Other Health Care Coverage: None
Expenses Information	Medical Expenses (before insurance) per Year: $2000 Dental Expenses (before insurance) per Year: $320 Maximum Number of Vacation Days Willing to SELL: 0 Minimum Number of Vacation Days Wanting to BUY: 2 Amount Spent per Year on Dependent Care: $400
Other Information	Is Spouse Employed by NCR: No Is the Individual a Smoker: No Number of Hours of Exercise per Week: 5

Figure 4

Results of Construct Validation

Benefit Type	Intra-Expert Agreement (Kappa)	Intra-Novice Agreement (Kappa)	Novice-Expert Agreement (Kappa)
Health Care Insurance	0.53	0.39 [3]	0.45 [5]
Dental Care Insurance	0.65 [1]	0.63 [4]	0.65
Short Term Disability	0.34 [2]	0.39 [3]	0.37
Long Term Disability	0.85 [1]	0.58 [4]	0.69 [6]
Life Insurance	0.19 [2]	0.26 [3]	0.21
Accidental Death and Dismemberment Insurance	0.17 [2]	0.20 [3]	0.18
Spouse Life Insurance	0.45	0.50	0.44

[1] Level of intra-expert agreement is significantly above 0.50 ($p < 0.0001$)

[2] Level of intra-expert agreement is significantly below 0.50 ($p < 0.001$)

[3] Level of intra-novice agreement is significantly below 0.50 ($p < 0.01$)

[4] Level of intra-novice agreement is significantly above 0.50 ($p <0.01$)

[5] Intra-Expert agreement is significantly greater than Novice-Expert agreement ($p < 0.05$)

[6] Intra-Expert agreement is significantly greater than Novice-Expert agreement ($p < 0.01$)