
Michael C. Sturman
Cornell University, mcs5@cornell.edu

George T. Milkovich
Cornell University, gtm1@cornell.edu

Follow this and additional works at: https://scholarship.sha.cornell.edu/articles

Recommended Citation

This Article or Chapter is brought to you for free and open access by the School of Hotel Administration Collection at The Scholarly Commons. It has been accepted for inclusion in Articles and Chapters by an authorized administrator of The Scholarly Commons. For more information, please contact hotellibrary@cornell.edu.

If you have a disability and are having trouble accessing information on this website or need materials in an alternate format, contact web-accessibility@cornell.edu for assistance.

Abstract

A method for validating expert systems, based on validation approaches from psychology and Turing's "imitation game," is demonstrated using a flexible employee benefits expert system. Psychometric validation has three aspects: the extent to which the system and expert decisions agree (criterion-related validity), the inputs and processes used by experts compared to the system (content validity), and differences between expert and novice decisions (construct validity). If these criteria are satisfied, then the system is indistinguishable from experts for its domain and satisfies the Turing Test.

Personal Choice Expert (PCE) was designed to help employees of a Fortune 500 firm choose benefits in their flexible benefits system. Its recommendations do not significantly differ from those given by independent experts. Hence, if the system-independent expert agreement (criterion-related validity) were the only standard, PCE could be considered valid. However, construct analysis suggests that re-engineering may be required. High intra-expert agreement exists only for some benefit recommendations (e.g., dental care and long-term disability) and not for others (e.g., short-term disability, accidental death and dismemberment, and life insurance). Insights offered by these methods are illustrated and examined.

Keywords
Decision Support Systems

Disciplines
Benefits and Compensation

Comments

Required Publisher Statement

This article or chapter is available at The Scholarly Commons: https://scholarship.sha.cornell.edu/articles/351

Michael C. Sturman and George T. Milkovich
Cornell University

A method for validating expert systems, based on validation approaches from psychology and Turing’s “imitation game,” is demonstrated using a flexible employee benefits expert system. Psychometric validation has three aspects: the extent to which the system and expert decisions agree (criterion-related validity), the inputs and processes used by experts compared to the system (content validity), and differences between expert and novice decisions (construct validity). If these criteria are satisfied, then the system is indistinguishable from experts for its domain and satisfies the Turing Test.

Personal Choice Expert (PCE) was designed to help employees of a Fortune 500 firm choose benefits in their flexible benefits system. Its recommendations do not significantly differ from those given by independent experts. Hence, if the system-independent expert agreement (criterion-related validity) were the only standard, PCE could be considered valid. However, construct analysis suggests that re-engineering may be required. High intra-expert agreement exists only for some benefit recommendations (e.g., dental care and long-term disability) and not for others (e.g., short-term disability, accidental death and dismemberment, and life insurance). Insights offered by these methods are illustrated and examined.

Introduction

Increasingly, expert systems offer the promise of becoming managerial decision aids. Much of the research on managerial expert systems focuses on their feasibility [2] [6], case studies describing their applications [12] [22], evaluations of alternative methods of knowledge acquisition and representation [10] [11] [13] [23] [25], and other issues related to user interfaces [9].

One critical, yet relatively under-analyzed issue involves validation of expert systems. Expert system validity is 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] [15], offers a
framework for understanding validation [16], and reports attempts to validate expert systems [4] [18]. However, there is no consensus about the methods appropriate for validation, and virtually overlooked is the substantial body of knowledge on validation in psychology. Reporting the validities of psychological tests used to predict future behaviors or assess behavioral attributes is standard procedure for judging tests [7] [14] [19].

Our study integrates the psychometric validation research with the existing work on expert system validation. Although some work has described how psychometric validation applies to expert system validation [15] [16], our study demonstrates a methodology on a system we designed. Additionally, the method applies to domains where decisions do not have predefined correct answers and where expertise is not well mapped. We demonstrate this method with a human resource expert system developed to aid employees in making decisions in a flexible employee benefit environment. Although this study uses a human resource expert system, the method should also be relevant to other systems where decision makers face similar conditions.

The Turing Test Framework

Validity is the degree to which inferences and decisions are justified by evidence [20]. Much of the psychometric literature on validation focuses on ascertaining the validity of devices generally used to predict an individual’s behavior in a given context (e.g., in a job or school). Expert systems operate in different contexts in that they mimic the decisions made by someone deemed to be an “expert.” Although similarities exist between psychological and expert system contexts, adaptations are required so that expert systems can be evaluated [3] [4] [15] [16].

The Turing Test, which sets a standard for the success of artificial intelligence, has been applied to the evaluation of expert systems [16] [17]. The test proposes that a “thinking machine” has been successfully developed if an individual cannot determine whether a given conversation or task is being performed by a machine or by a person [24]. Using Turing’s “imitation game,” if it is impossible to tell from a set of decisions those that were made by experts and those that were made by an expert system, then the system is valid.

The Turing Test applies to expert system validation through the implementation of three psychometric validation methods: criterion-related validation, content validation, and construct validation [1] [5] [8] [15] [16]. The system must produce answers like an expert. This is analyzed through criterion-related validation. The system must also collect information like a human expert. This is tested through content validation. For a computer program to be indistinguishable from an expert, it must be
possible to distinguish between experts and novices. This can be verified through construct validation. A system's validity can be inferred by the degree it exhibits these characteristics.

**Validation**

Criterion-related validation measures the statistical relationship that exists between a given index and a criterion score [1] [19]. The system processes are not investigated; rather, the system is treated like a “black box” and only the inputs (e.g., employment test) and outcomes (e.g., performance) are analyzed. In essence, criterion-related validation as applied to an expert system measures the relationship between decisions developed by the system and decisions developed by human experts [15]. A valid system would produce the same answers as the experts. However, in many decision situations, particularly those involving employees, a single correct answer simply does not exist. Rather, several possible answers may be equally correct. When there is no gold standard, or perfect answer, the experts’ decisions usually serve as the standard. In our study, the correct choice given perfect future knowledge is simply not available. Expert recommendations, based on information available at the time of the decision, served as the correct decisions. System validity will thus be limited by the variability in expert decisions (i.e., expert reliability) [14].

Content validation addresses the extent to which a particular measure represents the content universe of the property being measured [8]. In industrial psychology, this form of validation verifies that a device samples all the critical behaviors and knowledge that encompass the process in question [1] [14] [19]. For expert systems, it reflects the degree to which the system’s logic mimics the processes experts use to make their decisions. Content validation is not concerned with the decisions reached by the expert system; it is concerned with how the expert system reaches its decisions. Content validation focuses on determining which questions used by the system are critical for solving the problems facing the system. This analysis, similar to part of the process used to validate the MYCIN expert system [3] [18], involves comparing the inputs used by the expert system to those required by an independent group of experts. Essentially, content validation applied to an expert system entails examining whether the system collects all appropriate data (uses the same variables that independent experts do), excludes inappropriate data (does not use variables which experts do not use), and uses these variables in a manner similar to independent experts.

Construct validation attempts to verify that a device actually measures what it purports to measure [5] [14]. A test can be said to be valid only if it can be shown that it 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 [8]. A construct valid expert system must perform like experts and differ from novices. Construct validity can thus be inferred through criterion-related validity, high expert reliability, and significant differences between expert and novice recommendations.

**Evaluation of Personal Choice Expert**

*Method and Design*

Personal Choice Expert (PCE) is an expert system designed to help employees choose among benefits alternatives from a flexible benefits plan [20]. The system design was based on an actual flexible benefits program of a Fortune 500 firm and field tested on employees [21]. It generates specific recommendations for each of the various benefit forms and coverage options offered to employees. PCE is based on three company benefit managers’ expertise supplemented by information provided by its benefit consultants (e.g., booklets and comparative data).

To evaluate PCEs validity, we developed a questionnaire with 10 actual employee scenarios. (Due to limits on expert time, 10 were all that were feasible to collect.) The scenarios were based on actual employee information: family status, levels of alternate benefit coverage, spending information, and other potentially relevant descriptors (e.g., spouse’s income, other sources of income, and hours of exercise per week). The data provided included both fields that were and were not required by the expert system. Independent experts (experts not involved in the design of the system) and novices filled out the questionnaire by making benefits decisions for each scenario. Subjects identified variables they used in their decisions and described any other information they would have liked when making their recommendations.

Sixteen people were identified by the company as benefits experts. Three experts were involved in the system design. The remaining 13 were used in the validation study. Of the 13 experts surveyed for the study, 8 completed the questionnaire (62 percent response rate). The experts’ mean tenure at the company was 10 years; average tenure as a benefits counselor was 7 years. Novices were 15 college students who had no previous knowledge of the benefits program, benefits decision making, or work experience in benefits administration.

Generally, employees solicit advice from one of these benefit counselors who services their geographic units. We considered employing a Delphi procedure to generate consensus among the benefits counselors. However, this would not accurately mimic the approach used by employees or by
the experts. Employees seek out one benefits counselor. Benefits counselors do not consult with their peers to answer questions. Thus, to best understand the validity of the system, the independent recommendations of each of the experts needed to be obtained.

To measure agreement, the number of agreements between all the experts divided by the total number of potential agreements was calculated. This calculation yields a proportion, equaling 1 when agreement is perfect, and 0 when there is no agreement. These proportions can then be compared to cutoff values or other proportions using standard t-tests.

Analysis of Criterion-Related Validity

The question here is whether the system’s and independent experts’ answers are equivalent. To answer this, we compared the recommendations of the system and those of the independent experts for each benefit category. The proportion of agreement was calculated for each benefit category by comparing the system’s recommendations for the 10 employee scenarios to the recommendations of the 8 independent experts for the same employees.

T-tests were performed to determine if recommendations of the expert system differed significantly from those of independent experts. For each benefit category, the mean intra-expert agreement, mean system-expert agreement, total number of comparisons in each benefit category, and the p-values for differentiating between the proportions of intra-expert agreement and system-expert agreement are shown in Table 1.

Note that the level of expert system-expert agreement was never significantly lower than the mean level of intra-expert agreement (the agreement among the independent experts). That is, each of the independent experts agreed with the expert system’s recommendations at least as much as they agreed with the other experts’ recommendations. In the case of life insurance and accidental death and dismemberment insurance, the expert system-expert agreement was lower than the mean intra-expert agreement; yet for these situations the p-values (based on one-tailed f-tests) were .45 and .10, respectively. Thus, the expert system exhibits criterion-related validity for all benefit categories; the recommendations provided by the independent experts did not significantly differ from those of the expert system.

In three other cases-health care insurance, short-term disability, and spouse life insurance-the expert system-expert agreement was slightly higher but otherwise undifferentiated from the intra-expert agreement, with p-values (based on one-tailed t-tests) of .44, .06, and .39, respectively. For the other two benefit categories-dental care and long-term disability-the expert system-expert agreement
was significantly higher than the mean intra-expert agreement, with p-values (based on one-tailed f-tests) of less than .05. The system performs better, that is, it captures more common rater variance, than the average pairing of independent human experts for these two benefit types.

### Table 1: Results of criterion-related validation.

<table>
<thead>
<tr>
<th>Benefit Type</th>
<th>Proportion of Intra-Expert Agreement (Number of Comparisons)</th>
<th>Proportion of System-Expert Agreement (Number of Comparisons)</th>
<th>Difference Between System and Experts (p values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health care insurance</td>
<td>.53 (N=280)</td>
<td>.54 (N=80)</td>
<td>.44</td>
</tr>
<tr>
<td>Dental care insurance</td>
<td>.66 (N=280)</td>
<td>.78 (N=80)</td>
<td>.02</td>
</tr>
<tr>
<td>Short-term disability</td>
<td>.34 (N=280)</td>
<td>.44 (N=80)</td>
<td>.06</td>
</tr>
<tr>
<td>Long-term disability</td>
<td>.86 (N=273)*</td>
<td>.92 (N=79)*</td>
<td>.04</td>
</tr>
<tr>
<td>Life insurance</td>
<td>.26 (N=273)*</td>
<td>.25 (N=79)*</td>
<td>.45</td>
</tr>
<tr>
<td>Accidental death and dismemberment insurance</td>
<td>.27 (N=273)*</td>
<td>.20 (N=79)*</td>
<td>.10</td>
</tr>
<tr>
<td>Spouse life insurance</td>
<td>.45 (N=273)*</td>
<td>.47 (N=79)*</td>
<td>.39</td>
</tr>
</tbody>
</table>

*Number of comparisons is less due to one missing value on one of the experts’ questionnaires.

### Analysis of Content Validity

Content validity compares the inputs used by the independent experts with those used by the expert system. Determining content validity involves three steps: qualitatively analyzing the variables used/not used; determining if the level of intra-expert agreement regarding which variables the system should use is above the chosen cutoff (here we demonstrate a 50 percent cutoff to signify a majority); and determining if the experts agree with the content selected by the expert system.

The expert system uses 23 variables to generate its recommendations. Of these 23 variables, 11 were cited as necessary by all 8 independent experts, 6 other variables were cited by 7, and at least half of the experts rated the other 6 variables as necessary. Experts reported using 14 variables not used by PCE. Ten of these variables were cited by only one expert. Thus, it is not unreasonable to assume that this did not signify deficient content. However, four other variables were cited by more than one
person: gender (cited by three experts), if the employee has a spouse at the company (seven cited), if
the employee is a smoker (six cited) and the number of hours per week that the subject exercises (six
 cited). The frequencies of variables used by the system and experts, and the frequencies of variables
used by more than one expert but not by the system are reported in Table 2.

The precise effects of these variables are impossible to determine from our data, though it is
possible to make some estimates. Although gender was used by more than one expert, given that less
than half of the experts used the variable, it was determined that it was not deficient content on the
part of the expert system to not use the variable. Experts probably found the variables of spouse
employment at the company useful because recommendations would be based on benefits choices by
both individuals. The expert system solicits similar information by specifically asking questions regarding
alternative health insurance, dental insurance, life insurance, spouse life insurance, and short term
disability insurance, in addition to asking questions about spouse income. Thus, to include the question
in the expert system logic seems redundant. This view was also expressed by the three experts who
helped design the system during the knowledge acquisition phase.

The other two variables which the independent experts use but the system does not are
whether the employee smokes and the number of hours the employee exercises per week. These
variables are likely signals of health for determining medical 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. This projection, as recommended by the experts when the system was designed, is
based on employee forecasts of future medical expenses. Second, the system bases choices of the other
insurances on the financial requirements of the employee.

Determining whether the omission of these variables reflects deficient content validity is
difficult given the relatively small number of independent experts. However, PCE overall collects very
similar information. Thus, although some content differences do exist, there is arguably a high degree of
content validity.

This analysis is qualitative, which makes it difficult to replicate for validating other systems.
Nevertheless, it provides insights into the variables used and omitted by the expert system. Intra-expert
agreement regarding variables that should or should not be used involves comparing each expert to the
group as a whole (not including themselves). The proportion of agreement was calculated, equaling the
number of times that the subject agreed with each of the other experts (or all the experts for the expert
system) to use or not use each variable, divided by the total number of potential agreements. For
example, if the system used all the variables listed in Table 2, then the frequency of agreement would equal 181 (the sum of the right hand column in Table 2), divided by 216 (8 x 27 variables), or 84 percent. If the expert system used all the same variables except for age, this value would equal 80 percent (173/216). If there is perfect agreement, this proportion will equal 1. If there is no agreement, then the value will be 0. The proportions for each expert, the expert system, and the mean expert agreement are reported in Table 3. The level of intra-expert agreement is significantly greater than 50 percent (one-tailed t-test; p<.01).

The proportion of expert system agreement is not significantly different from the mean expert agreement (two-tailed f-test; p=. 1006). In fact, the mean agreement for the expert system is actually marginally significantly greater than the mean intra-expert agreement (one-tailed f-test; p=.0503).

As described above, the proportion of agreement is not a sufficient statistic for estimating content validity (unless it equals 1.00). It is possible that just by neglecting one question, a decision maker or expert system would not have the appropriate information to make a “correct” decision. For this reason, it is important that the quantitative and qualitative methodologies are employed for evaluating a system’s content validity.

In the PCE case, our specifications for the three part test of content validity are satisfied. However, we recognize that other analysts may set higher cutoffs, thus requiring greater agreement between the experts used in the system design and those used later for this validation demonstration.

Analysis of Construct Validity

Construct validity involves ascertaining the degree of both intra-expert agreement (i.e., agreement among the independent experts) and expert to novice agreement. This requires two separate tests. First, intra-expert agreement was analyzed. Second, intra-expert agreement was compared to novice-expert agreement. These analyses were performed for each benefit category.

To test for intra-expert agreement, a significance test was performed for each benefit category to determine if the frequency of expert agreement was above 50 percent. Levels of agreement were calculated for each expert as the agreement between the individual expert and the other experts. The measure was obtained for each of the eight experts and each benefit category. Frequencies were compared using f-tests of proportions.
It should be noted that the 50 percent value for the significance test is arbitrary. We chose it to signify a majority of agreement. If the level of agreement exceeded 50 percent (using a one-tailed t-test), we concluded that intra-expert agreement existed. Alternatively, we could assume intra-expert agreement only if the frequency of agreement equaled 1.00 (i.e., perfect agreement). The best method depends on the task domain and the requirements of the system designers.

We performed similar tests to determine whether novices' recommendations significantly differed from those of the experts. The novice-expert agreement was calculated as the agreement

<table>
<thead>
<tr>
<th>Variables Used by Experts and Expert System</th>
<th>Number of Experts Who Use Variable (Out of 8 Experts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Age</td>
<td>8</td>
</tr>
<tr>
<td>2. Martial status</td>
<td>8</td>
</tr>
<tr>
<td>3. Number of legal dependents</td>
<td>8</td>
</tr>
<tr>
<td>4. Number of special dependents</td>
<td>7</td>
</tr>
<tr>
<td>5. Base pay</td>
<td>7</td>
</tr>
<tr>
<td>6. Total pay</td>
<td>4</td>
</tr>
<tr>
<td>7. Additional income</td>
<td>5</td>
</tr>
<tr>
<td>8. Savings</td>
<td>4</td>
</tr>
<tr>
<td>9. Spouse's income</td>
<td>8</td>
</tr>
<tr>
<td>10. Flexible benefits credits</td>
<td>7</td>
</tr>
<tr>
<td>11. Amount spent per year on non-essentials</td>
<td>4</td>
</tr>
<tr>
<td>12. Covered by another short term disability plan</td>
<td>8</td>
</tr>
<tr>
<td>13. Covered by another life insurance plan</td>
<td>8</td>
</tr>
<tr>
<td>14. Value of other life insurance plan</td>
<td>8</td>
</tr>
<tr>
<td>15. Covered by another dental plan</td>
<td>7</td>
</tr>
<tr>
<td>16. Spouse covered by his or her own life insurance plan</td>
<td>8</td>
</tr>
<tr>
<td>17. Value of spouse's other life insurance plan</td>
<td>8</td>
</tr>
<tr>
<td>18. Equivalent of other health care coverage</td>
<td>8</td>
</tr>
<tr>
<td>19. Medical expenses per year</td>
<td>8</td>
</tr>
<tr>
<td>20. Dental expenses per year</td>
<td>7</td>
</tr>
<tr>
<td>21. Maximum number of vacation days willing to sell</td>
<td>6</td>
</tr>
<tr>
<td>22. Minimum number of vacation days willing to buy</td>
<td>6</td>
</tr>
<tr>
<td>23. Amount spent per year on dependent care</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variables Used by Experts but not by Expert System</th>
<th>Number of Experts Who Use Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is spouse employed by the company</td>
<td>7</td>
</tr>
<tr>
<td>Is the individual a smoker</td>
<td>6</td>
</tr>
<tr>
<td>Number of hours exercise per week</td>
<td>6</td>
</tr>
<tr>
<td>Gender</td>
<td>3</td>
</tr>
</tbody>
</table>

Variables Used by Expert System but not by Experts

None

Table 2: Results of content validation.
between each of the novices and the group of experts and compared to intra-expert agreement for each benefit category. One-tailed significance tests were used to determine if the level of novice-expert agreement was significantly below intra-expert agreement.

The results of these tests are shown in Table 4. Experts agreed on over 50 percent of their recommendations \( (p \leq .0001) \) for employees' dental (66 percent) and long-term disability (86 percent) insurance coverages. Expert agreement was significantly below 50 percent for short-term disability (34 percent), employee life (26 percent), and accidental death and dismemberment (27 percent) insurances \( (p \leq .0001) \). For health care (53 percent) and spouse life insurance (45 percent), neither benefit types tested significantly above or below 50 percent. Tests comparing novice and expert recommendations also yielded mixed results. Expert and novice recommendations differed for health care \( (p \leq .01) \), dental care \( (p \leq .05) \), long-term disability decisions \( (p \leq .0001) \), life insurance \( (p \leq .01) \), and accidental death and dismemberment insurance \( (p \leq .001) \). For short-term disability and spouse life insurance, no significant differences were discovered.

In summary, high intra-expert agreement exists only for dental and long-term disability insurance recommendations. There is low intra-expert reliability for short-term disability insurance, life insurance, and accidental death and dismemberment insurance. The construct of expertise does not seem uniform over all benefits types. In other words, being an expert in long-term disability insurance

### Table 3: Frequency values for estimating content validity.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Frequency of Agreement (All Variables)</th>
<th>Frequency of Agreement (All Variables with N&gt;1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1</td>
<td>.56</td>
<td>.70</td>
</tr>
<tr>
<td>Expert 2</td>
<td>.80</td>
<td>.78</td>
</tr>
<tr>
<td>Expert 3</td>
<td>.77</td>
<td>.74</td>
</tr>
<tr>
<td>Expert 4</td>
<td>.78</td>
<td>.82</td>
</tr>
<tr>
<td>Expert 5</td>
<td>.82</td>
<td>.80</td>
</tr>
<tr>
<td>Expert 6</td>
<td>.70</td>
<td>.65</td>
</tr>
<tr>
<td>Expert 7</td>
<td>.79</td>
<td>.77</td>
</tr>
<tr>
<td>Expert 8</td>
<td>.80</td>
<td>.78</td>
</tr>
<tr>
<td>Expert System</td>
<td>.80</td>
<td>.78</td>
</tr>
<tr>
<td>Mean</td>
<td>.75</td>
<td>.76</td>
</tr>
</tbody>
</table>

Notes: Frequency of agreement equals the number of times each expert (or the expert system) agrees with the other experts on whether to include each of the variables, divided by the total number of possible agreements.

The first column reports the frequency of intra-expert agreement for all variables mentioned by the experts. This includes the 23 used by the expert system and the 14 variables included by various experts. The total number of comparisons equals 259 \( (7 \times 37) \) for each of the experts and 296 \( (8 \times 37) \) for the expert system.

The second column reports the frequency of intra-expert agreement for all variables except those reported by only a single person. Included in these calculations are the 23 variables used by the expert system plus the four variables included by two or more of the various experts. The total number of comparisons is 189 for the experts and 216 for the expert system.
does not mean that one is necessarily an expert in health care decisions. Possibly, it is easier to become an expert for certain benefit types than it is for others, or there may be no agreement among experts for certain benefit types. If a larger number of experts existed, then agreement may have increased by virtue of large N, however, under conditions where only a few experts exist (which is almost self-evident from the definition of expert), our results show that experts did not exhibit uniform reasoning for decisions on short-term disability insurance, life insurance, and accidental death and dismemberment insurance.

Table 4: Results of construct validation.

<table>
<thead>
<tr>
<th>Benefit Type</th>
<th>Proportion of Intra-Expert Agreement (Number of Comparisons)</th>
<th>Proportion of Novice-Expert Agreement (Number of Comparisons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health care insurance</td>
<td>.53 (N=280)</td>
<td>.44 (N=1200)</td>
</tr>
<tr>
<td>Dental care insurance</td>
<td>.66 (N=280)</td>
<td>.59 (N=1200)</td>
</tr>
<tr>
<td>Short term disability</td>
<td>.34 (N=280)</td>
<td>.37 (N=1200)</td>
</tr>
<tr>
<td>Long term disability</td>
<td>.86 (N=273)</td>
<td>.68 (N=1185)</td>
</tr>
<tr>
<td>Life insurance</td>
<td>.26 (N=273)</td>
<td>.18 (N=1185)</td>
</tr>
<tr>
<td>Accidental death and dismemberment insurance</td>
<td>.27 (N=273)</td>
<td>.17 (N=1185)</td>
</tr>
<tr>
<td>Spouse life insurance</td>
<td>.45 (N=273)</td>
<td>.43 (N=1185)</td>
</tr>
</tbody>
</table>

1 Level of Intra-Expert agreement is significantly above .50 (p<.0001)
2 Level of Intra-Expert agreement is significantly below .50 (p<.0001)
3 Intra-Expert agreement is significantly greater than Novice-Expert agreement (p<.05)
4 Intra-Expert agreement is significantly greater than Novice-Expert agreement (p<.01)

Conclusions

Validation is a crucial step in the knowledge engineering process. Much of an expert system’s utility hinges upon the validity of the recommendations it provides to the decision maker. Thus, methods of expert system validation need to be developed. Psychometric concepts of criterion-related, content, and construct validation can be applied to validating expert systems, but require some modification. A summary of the three validation tests for each benefit category is presented in Table 5. The table summarizes the three parts of the validation process as applied to PCE by saying whether or not evidence existed which satisfied each of the validation sections.
The criterion-related analysis suggests that expert system decisions do not significantly differ from expert recommendations. It would appear that the system produces responses like our sample of independent experts for all benefit categories. Additionally, for three of the seven benefit types, the system-expert agreement exceeds the average intra-expert agreement, implying that the expert system may be able to out-perform a single human expert for some benefit types.

The content validity analysis suggests the expert system uses inputs that are similar to those used by experts. Although some discrepancies exist between the experts and the expert system, there is reason to believe that the expert system demonstrates content validity.

Yet the evidence discovered in the construct analysis reveals that the company experts (three used during the design versus eight used during validation) do not uniformly agree across all benefit decisions. While high intra-expert agreement does exist for some benefit decisions, specifically dental care insurance and long-term disability insurance, 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. Comparing expert to novice decisions, significant differences occur for health care, dental care, long-term disability, life, and accidental death and dismemberment insurance.

The expert system is therefore valid for dental care and long-term disability decisions. That is, the expert system performed like a human expert (and different from a human novice) for decision making in these benefit categories. However, there is mixed evidence for the other benefit categories. For health care insurance, even though the intra-expert agreement was not significantly above 50 percent, given that experts significantly differ from novices, the expert system has the potential to contribute in this area because the expert system offers more consistency than what novices obtain by themselves. Thus, although the expert reliability was not statistically significantly above the 50 percent cutoff, it may be a useful decision aid.
For the benefits categories of life insurance and accidental death and dismemberment insurance, the construct validity evidence is mixed. The level of intra-expert agreement was significantly below the 50 percent cutoff, but expert agreement was significantly higher than novice agreement. Although this signifies some potential benefit of the system for novice users, the system does not seem to be valid for these categories.

For the final two benefit types—short-term disability and spouse life insurance—intra-expert agreement is low and no significant difference exists between expert and novice decisions. Hence, for these benefit categories, there is no evidence of expertise for an expert system to mimic.

A somewhat surprising finding of this study is the variance of agreement levels over the seven benefit categories. Benefits experts had agreement ranging from a high of 86 percent agreement (long-
term disability) to a low of 26 percent agreement (life insurance). While this is not directly relevant to the demonstration of the validation methodology, it does have serious implications for the design of a flexible benefits expert systems, as well as for the field of flexible benefits in general. Perhaps decision making for some benefits categories is simple, while for others it is either highly complex or no true “correct” answer exists. Future research on flexible benefits should address the issues of defining benefits decision-making expertise, determining the complexity of flexible benefits decision making, and deriving methods for evaluating decision-making quality.

Our study has a number of limitations. As mentioned throughout the paper, many of the criteria are based on the judgments of the designers. For example, we chose to look for values significantly above 50 percent agreement, but a different cutoff could be used. Clearly, research on the sensitivity of these cutoffs to the conclusions and actions taken is needed. The sample size of experts is limited inevitably with expert systems. If experts were not in short supply, then it is unlikely that an expert system would be attractive. Regardless of the universality of this statement, it is at least true for the domain in our study.

Due to the data collected, the content was analyzed for the system as a whole instead of by benefit type. Although we were able to draw conclusions regarding the content validity, we believe content validation may be best served by analyzing it by type of decision as well.

Calling for more research is the norm. However, in the case of examining the process for validating expert systems, such a call is clearly warranted. This study only demonstrates a single application of a validation technique. Future work could entail either testing this methodology in other domains or refining it. Validation is an integral step in knowledge engineering. The triple validation approach demonstrated in this paper should help in expert system design and implementation. [Received: June 1, 1993. Accepted: January 3, 1995.]

References


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


Sturman, M. C., Milkovich, G. T., & Hannon, J. M. *Expert system effects on employee decisions and attitudes.* Under review.

