VALIDATION OF AN EXPERT SYSTEM FOR NEW CONSTRUCTION AND RETROFIT OF BRIDGE RAILS

Prakit PREMTHAMKORN
Dean
Graduate School
Mahanakorn University of Technology
51 Cheum Sampan Road, Nong Chok
Bangkok 10530, Thailand
Fax: +66-2-988-4040
E-mail: prakit@mut.ac.th

Pannapa HERABAT
Visiting Professor
School of Civil Engineering
Asian Institute of Technology
P.O. Box 4, Klong Luang
Pathumthani 12120, Thailand
Fax: +66-2-516-2126
E-mail: pannapa@ait.ac.th

Abstract: Bridge Rail Expert System (BREXS) is a knowledge-based system developed to aid novice bridge engineers in coping with selection and installation of bridge railing for the state of Texas. The domain of bridge railing has distinct characteristics that there are multiple domain experts and there is not a single most accurate solution to a problem. Therefore there is no individual or source of expertise that can be used to judge performance of the expert system. To this end, a special validation procedure has been developed and embedded to the development cycle of BREXS. This paper presents formal validation methodology of BREXS, and discussed the results. It was concluded that BREXS exhibits acceptably high level of expertise that is comparable to experienced bridge engineers. Nonetheless, the process revealed some weaknesses that called for minor revision of the system.

1. INTRODUCTION

Verification and validation are acknowledged as an integral part of the development process of an expert system. Currently, there is no validation methodology that is accepted as a standard. Expert system development teams need to design their own testing procedure according to the nature of the domain. Important characteristics of the bridge rail design domain are: (1) there are multiple domain experts and, (2) there is not a single most accurate solution to a problem. In other words, there is no individual or source of expertise that can be used to judge performance of the expert system.

A special validation procedure has been developed for BREXS. Details of the procedure as well as application to BREXS are discussed by Premthamkorn (1993) and Mitri (1991). In this paper, the validation methodology is outlined and the significance of the results are discussed.

2. BRIDGE RAIL EXPERT SYSTEM

There are over twenty-five bridge rail designs that have been crashed tested and certified by the Texas Department of Transportation (TxDOT) to use with highway bridges. Nominally, any standard rail can be used in any location on any type of highway bridge. However, this generic interchange is not always practical and appropriate. Effective design of new or retrofit bridge rails depends on many interrelated factors. These include adherence to standard specifications, structural adequacy, safety of drivers and pedestrians, and economical considerations. Chosen bridge railing design should match

the specific site needs authority (Guide Specifications for Bridge Railings, 1989). Requirements for bridge railing performance differ significantly from site to site.

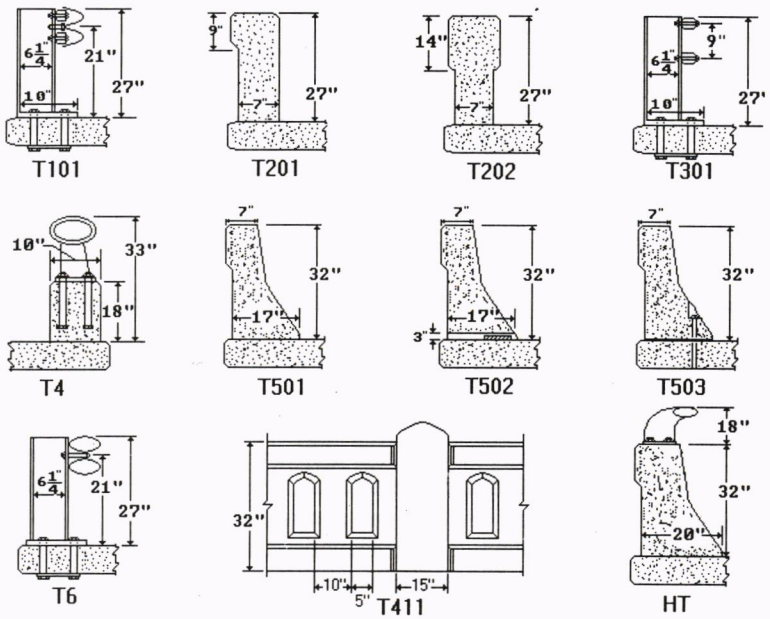

FIGURE 1. Examples of Cross-Sections of Standard Bridge Rails

Problems in bridge rail design and retrofit can not be solved by algorithmic approaches. Bridge engineers have to rely mainly on their experience and expertise to solve this class of problems. Lack of consensus on design guidelines and scarcity of expertise necessitates a decision-supporting tool for novice engineers. The domain of bridge rail design and retrofit is also appropriate for implementation of an expert system in that the scope is relatively narrow and yet complex enough to justify the cost and effort of development.

Bridge Rail Expert System (BREXS) is a knowledge-based system developed to aid novice engineers in dealing with selection and installation of highway bridge railing. The system is designed to support both new construction and retrofit projects. Motivation for development stems from a need to integrate domain expertise and knowledge from complementary disciplines to provide a unified guidance for bridge railing practice. Development goal is to customize bridge railing knowledge-bases and databases, and incorporate them with other existing analytical computer codes to solve complex bridge railing problems. BREXS consists of two subsystems that deal with new construction and retrofit problems. The system has been delivered to TxDOT for operational use since 1993.

The objective of development of BREXS is to incorporate knowledge and expertise compiled from experts that are hierarchically and geographically dispersed into a single system that is accessible to statewide residencies. Databases and existing algorithmic programs are integrated into the system to enhance accuracy of the solution. *BREXS* consists of two subsystems that deal with new construction and retrofit problems. Primary users of the system are expected to be moderately experienced to inexperienced bridge engineers. Documented knowledge as well as expertise from bridge railing experts around

the state of Texas are incorporated into its problem solving capabilities. Knowledge bases and databases are specifically designed to make recommendations for Texas rails.

In the domain of bridge rail design, the state highway engineers have the highest level of expertise. District engineers and resident engineers have lower levels of expertise, respectively. Expertise of bridge engineers is not identical, although their knowledge tends to overlap. In other words, the domain not only has more than one expert, but expertise from each level of engineers is necessary in deriving a solution to a problem. Discrepancies in knowledge and differences of opinions among these experts is the norm. Furthermore, since bridge rail recommendations are subjective, there is no single most accurate solution for a given problem. Accuracy of the solution is, therefore, difficult to measure.

While resident engineers play a major role in bridge rail design, they are not the only group of personnel that possesses the knowledge and experience necessary for selecting new and retrofit bridge rails. General design guidelines and policies originate and are controlled by officials that have the highest authority (Guide Specifications for Bridge Railings,1989). State highway engineers provide guidelines to district and resident engineers in various formats that range from formal guidelines and specifications to less formal memoranda and short courses in specific topics. In addition, state level engineers are also supervised by the Federal Highway Administration. Figure. 2 depicts the hierarchy of authority in highway engineering.

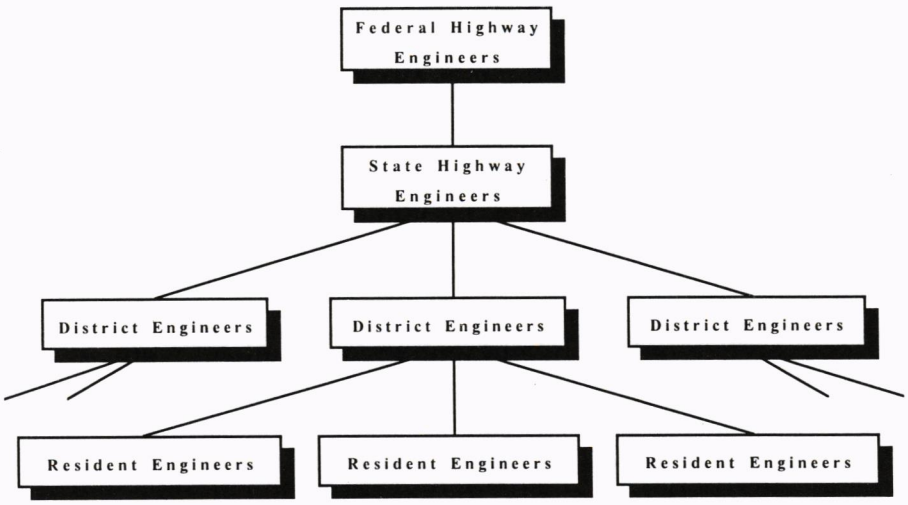

FIGURE 2. Hierarchy of Authority in Highway Design

3. VALIDATION METHOD

A formal validation method that can be applied to an expert system with multiple experts such as BREXS has been developed (Mitri 1991). Objectives of the development include:

(1) Embedding validation into the development life cycle

(2) Finding a technique for performance measures of an expert system.

The developed method is qualitative and consists of the following four approaches:
(1) Face Validation: Purpose of the face validation is to ensure that the system correctly identifies the problems being addressed and covers the intended scope of the domain. The goal is to ensure that the problem-solving approach is sufficiently well structured and formulated. Knowledge must also be correctly represented.

Face validation for BREXS was performed by a combination of the development team, lower level experts (district and resident engineers), and high level experts (state level engineers). The development team examined the system based on information extracted from the knowledge acquisition process. Experts at various hierarchical levels also assessed the usefulness of the system.

(2) Field Testing: This approach to validation involves the use of the prototype system in the field under real-world conditions for an extended period of time. Users are asked to report problems or provide feedback on system performance based on various aspects. Application of this approach is limited to a non-critical domain where error can be tolerated during the development phase. It is preferable that testers have at least a moderate level of expertise so that errors can be properly detected and feedback can be provided.

(3) Turing Test: In this test, performance of the expert system is evaluated without disclosing its identity (Turing, 1963). Purpose of the test is to exclude possible bias of the testers toward or against the expert system.

A Turing test was performed on BREXS by a high level expert (state engineer). Case data and three sets of solutions were provided to the expert. Among these three sets of solutions, one set was generated by BREXS and the other two are solutions recommended by lower level experts (district engineers). Based on an assumption that a machine may not be able to solve the problem as well as the human experts, the expert performing the Turing test was asked to identify the solution generated by BREXS from the three sets of solutions.

(4) Sensitivity Analysis: The purpose of this test is to examine effects on the solution when system parameters or input variables are varied. Information obtained from this test is useful for fine tuning the system performance by adjustment of certain system parameters. It is also useful for identifying inferior performance caused by over-sensitivity or under-sensitivity of the solution to certain parameters.

These four approaches are qualitative techniques that rely primarily on subjective comparisons. While lacking the mathematical precision of qualitative techniques, these approaches are more flexible and less demanding (O'Leary, 1990). Qualitative techniques are especially useful for validating prototype systems where time and cost are important constraints.

Figure 3 shows a schedule of validation approaches that are aligned chronologically with respect to the three major stages of development. Face validation starts during development of the prototype in order to detect major errors in presentation of knowledge. The development team cooperates with higher rank experts during this test.

Documentation from knowledge acquisition is useful in checking for discrepancies between knowledge that is acquired and its representation in the knowledge bases. Knowledge engineers must be able to explain reasoning behind solutions made by the system so that errors can be tracked. Feedback from experts during the test also serves to expand the prototype. For BREXS, this phase of testing was primarily performed with help from state level engineers and a few selected district engineers.

Field testing can be started as soon as the initial prototype is completed. It is performed until after the system is delivered. This testing involves experts at lower ranks and includes potential users of the system. The test must be performed at sites where bridge rail design actually takes place. Testers are asked to give their own recommendations before running the system. Relevant case data, recommendations from testers, and recommendations from the expert system are recorded. Tests include both standard and site-specific cases. Standard cases are those which are prepared by the development team. Site-specific cases are prepared by the testers from actual designs performed on a certain site. Ideally, standard and site-specific cases should statistically represent every aspect of the rail selection that might be encountered in the domain and test sites, respectively.

A Turing test should be performed after the system has been expanded from its initial prototype. Standard test cases from field test are useful in this step. Finally, sensitivity analysis can be performed independently from the other tests.

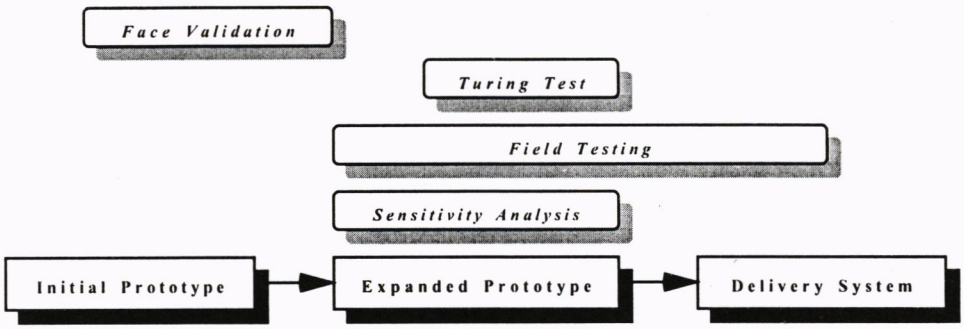

FIGURE 3. Schedule of Validation Approaches during Development Life Cycle

A delivery system normally evolves after several iterations of validation and revision of the expanded prototype (Figure 4). Unless there are major changes in logic of the problem solving approach, face validation is normally performed once during development of the initial prototype. Sensitivity analysis, field testing, and the Turing test need to be performed in order to measure performance of the system after each iteration.

4. PERFORMANCE MEASURE

There are eight performance measures that should be examined during the process of system validation. Each performance measure is briefly described in what follows:

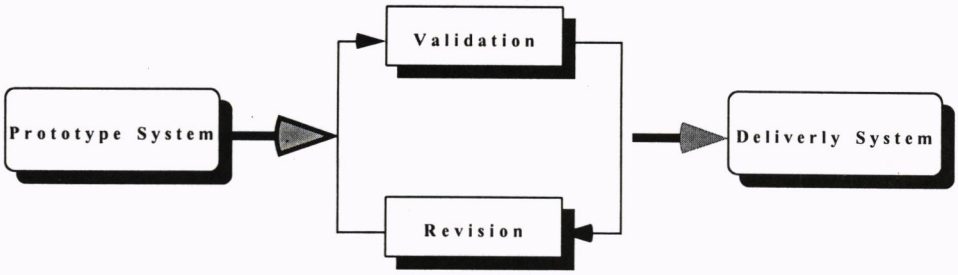

FIGURE 4 Life Cycle of System Validation and Revision

(1) Accuracy of the Solution: This measure indicates the degree of agreement between the solution generated by the system and human experts.

(2) Sensitivity: This measure reveals the degree of change in the solution caused by a variation of a certain parameter.

(3) Turing Test: This measure assesses performance of the system in comparison with referential human experts.

(4) Robustness: This measure reveals how well the system performs under a wide variety of possible conditions. This is to ensure that the system not only performs well with common problem cases but also with unusual cases.

(5) Realism: This measure compares the approach to problem solution taken by the system to the approach taken by human experts.

(6) Appeal: This measure reveals appeal of the system to users. This includes ease of use of the user-interface and ease of use of the system as a whole.

(7) Breadth: This measure concerns the capability of the system to cope with a variety of problem contexts in which it is expected to perform.

(8) Reliability: This measure concerns the degree of correctness of the solution. The system must yield a solution that is statistically trustworthy to within a certain degree.

Not all of these measurements are appropriate and applicable to all expert systems. For expert systems with multiple experts, accuracy of the solution and the Turing test are recommended for inclusion in the test evaluation.

5. VALIDATION OF BREXS

The four validation approaches described in chapter 3. have been applied to BREXS. A minimum level of system performance was targeted. If validation showed that the system did not reach or exceed the set threshold, the system had to go through one more iteration of revision, expansion, and then re-validation. If results were satisfactory, further system revision and expansion would not be performed. The acceptable performance levels were set as follows:

(1) The system must present an adequate level of agreement with both state level experts and lower level experts (district and resident engineers). The degree of agreement must be comparable to that found among human experts.

(2) In the Turing test, the system must yield 80% uncertainty in the judgment of the expert. That is, the expert must not be able to identify BREXS in more than 20% of the total test cases.

(3) Sensitivity analysis must yield proper sensitivity of the system. The system must exhibit an appropriate level of sensitivity of the solution to changes in input parameters.

For face validation, the state expert was asked to review the prototype and embedded knowledge. Issues covered in this stage of testing are: scope of the problem, knowledge representation, and accuracy of the solution. Validity of the graphical user interface was also included in the evaluation. The reasoning process was examined using the transcript generation capability of NEXPERT *OBJECT*. Knowledge bases were tested by the development team along with independent experts. For continuity, knowledge acquisition for prototype expansion was performed concurrently with face validation during each meeting with the domain experts.

Field testing took place after the initial prototype had been expanded. Primary objective of the test was to ensure that the expanded system performs adequately within its scope of the domain under actual field conditions. Test sites chosen were Bryan, Houston, San Antonio district offices, and the main TxDOT office in Austin. Engineers at these four test sites were asked to run 15-20 test cases. Ten of these cases were actually supervised by other district offices. The rest were actual cases supervised by that office. Choices of recommendation of each standard rail were strongly recommended, recommended, slightly recommended, or not recommended.

A Turing test was conducted after the field test had been completed. A state level expert, who was acknowledged to have a high level of domain expertise, was interviewed for this purpose. For each test case, three sets of solutions were given to the state expert. Among the three sets of solutions, two were recommended by district engineers and one was a solution suggested by BREXS. The state expert was asked to attempt to identify from three sets of different solutions the one that had been generated by BREXS.

Finally, a sensitivity analysis was performed by members of development team. About 30 parameters were systematically varied while changes in the output were monitored.

6. RESULTS

Results of the validation of BREXS are summarized according to the approach as follows:

6.1 Results of Face Validation

Face validation was performed during the prototype stage of the development. BREXS correctly identified the addressed problems. Domain experts were in substantial agreement that the knowledge was correctly represented.

6.2 Results of Field Testing

Since there is no single correct answer to bridge railing problems, accuracy of the solution generated by BREXS can only be subjectively assessed with reference to its agreement with the solutions of an expert. A measure called a "score of agreement" is evaluated based on the following factors:

1) Agreement in the top recommended rail type.
2) Agreement in the number of alternative rail types.
3) Agreement in ranking the rail types.
4) Agreement in not recommending rail types.
5) Agreement in reasoning.

The score lies between 0 and 100. The more the two answers agree, the higher the assigned score. Therefore a score of 100 means that there is a perfect match between two answers.

BREXS scored an average of 59.4 and 39.9 on agreement of overall recommendation in comparison with recommendations made by local experts and the state expert, respectively. BREXS shows a slightly higher level of agreement with respect to the topmost recommendation. Table 1 summarizes the score of agreement between BREXS and human experts at the four test sites.

TABLE 1. Score of Agreement between BREXS and Human Experts

Test Site (1)	Number of Test Cases (2)	Average Score of Overall Recommendation (3)	Average Score of Top Recommended Rail (4)
Bryan	12	68.8	78.3
Houston	16	57.9	48
San Antonio	19	54.6	64.4
Local Experts (Total)	47	59.4	62.4
State Expert	15	39.9	66.1

Agreement among human experts was also assessed. Scores of the test revealed a relatively low level of agreement in overall recommendations among them. Agreement on the top recommendation is slightly higher but still relatively low in an absolute sense. Average agreement of the local experts among themselves and with the state experts is consistently lower than that of BREXS. Tables 2 and 3 summarize the score of agreement among district engineers, and the score of agreement between district engineers and the state engineer, respectively.

In absolute terms, BREXS did not consistently show a high level of agreement with the experts at both local and state levels. However, considering the relatively low level of agreement among experts in the domain, BREXS showed an acceptable level of agreement

with human experts at both levels. It also showed a higher level of agreement with the state expert than the local experts. This means that the level of expertise of BREXS lies between that of the local experts and the state expert.

TABLE 2. Score of Agreement among District Engineers

Test Site (1)	Number of Test Cases (2)	Average Score of Overall Answers (3)	Average Score of Top Recommended Rail (4)
Bryan	15	26.4	33.2
Houston	19	28.74	32.42
San Antonio	14	37.35	44.57
	Average	30.8	36.73

TABLE 3. Score of Agreement between District Engineers and State Engineer

Test Site (1)	Number of Test Cases (2)	Average Score of Overall Answers (3)	Average Score of Top Recommended Rail (4)
Bryan	9	34.2	56.44
Houston	15	28.53	25.32
San Antonio	10	52.55	65.7
	Average	38.4	49.1

6.3 Results of Turing Test

Fifteen cases were used for the Turing test. With the case data, three sets of solutions were presented to the state expert. Among these three, one set of solutions was generated by BREXS, while the other two were solutions recommended by district engineers. Out of 15 cases, the state expert correctly identified BREXS twice. Out of the remaining 13 cases, 8 were incorrectly identified. In the remaining 5 cases, the state expert correctly identified BREXS once with uncertainty. Results of the Turing test are summarized in Table 4. The state expert correctly identified BREXS in only 20% of the total number of cases.

TABLE 4. Results of Turing Test

Action Taken by Test Subject (1)	Number of Test Cases (2)	Percentage (3)
Identified BREXS	2	13.1%
Identified BREXS with uncertainty	1	6.67%
Chose human expert	8	53.33%
Chose human expert with uncertainty	4	26.67%
Total	15	100%

While the number of test cases was relatively small, the results lead to the conclusion that BREXS's advice could not usually be distinguished from that of district engineers. This finding is consistent to the conclusion drawn above concerning the level of expertise that BREXS exhibits.

6.4 Results of Sensitivity Analysis

Combinations of input parameters were input to BREXS, and changes in the output were observed. Due to a very large number of possible combinations and time constraints, a limited number of test cases were used to test the system. The test was also simplified by recognition of the fact that no changes in the final output occur for several ranges of many of the input parameters. For example, design speed, which ranges from 30 mph to 60 mph, affects output only when changed in increments of 10 mph.

Results of the sensitivity analysis reveal that BREXS exhibits an appropriate level of sensitivity. It is more sensitive to important parameters such as ADT, percent trucks, type of understructure, visibility requirements, and the existing bridge type. Contrariwise, it is less sensitive to relatively unimportant parameters such as hydraulics, and type of construction. In comparison with human experts, knowledge engineers subjectively rate BREXS as shows more sensitivity.

7. CONCLUSION

Despite the fact that BREXS performs in an acceptable manner, some weaknesses were discovered during system validation. In some cases BREXS fails to discern small qualitative differences between the candidate rails. The overall ratings are not precise enough to reflect these discernible differences. In one case, BREXS rates the overall performance of T501 and T502 equally and recommends both as the topmost rails. However, according to case data in which a slightly longer bridge is indicated (150 ft), T502, which has small openings for precipitation drainage, is to be slightly preferred over T501 which has no opening. Despite the small margin of difference, T502 is obviously a better performing rail and should be ranked higher. Culprit of this weakness is possibly due to lack of uncertainty management of the system. Revision of the decision-making mechanism is planned.

The validation process outlined above was embedded into the development cycle of BREXS. Approximately 30% of the total effort was spent on validation of BREXS. Results of validation show that BREXS performs within the acceptable level set forth by the development team. BREXS exhibits a level of expertise that is comparable to experienced district engineers. In addition to the formal validation described earlier, several informal field tests have been carried out. Results of these test are consistent with the previous discussion.

Given the critical decisions that an expert system is called on to make, validation is important for the development of an expert system. While the task requires a great amount of effort and time, it is essential to consider validation an integral part of the development life cycle.

REFERENCES

Bellman, D., and Zadeh, L. A. (1970) Decision-making in a Fuzzy Environment, **Management Science, Vol. 17**, No 4, B141-B164.

Benefit to Cost Analysis Program, Vol. 1, Reference Manual, Publication No. FHWA-TS 88, Turner-Fairbanks Highway Research Center, McLean, VA.

Buchanan, B. et al. (1983) "Constructing an Expert System, **Building Expert Systems**, F. Hayes-Roth, D. Waterman, and D. Lenat, eds., Addison-Wesley, Reading, MA.

Guide Specifications for Bridge Railings (1989) American Association of State Highway and Transportation Officials (AASHTO), Washington, DC.

Mitri, W. (1991) **Validation of Expert Systems with Multiple Experts**, thesis presented to Texas A&M University, College Station, Texas, in partial fulfillment of the requirements for the degree of Master of Science.

Premthamkorn, P. (1993) **Expert System for Bridge Rail Selection**, dissertation presented to Texas A&M University, College Station, Texas, in partial fulfillment of requirements for the degree of Doctor of Philosophy.

Roschke, P.N., and Premthamkorn, P. (1990) Expert System for Bridge Rail Retrofit and Design: Architecture and Knowledge Acquisition, **OECD Workshop on Knowledge-based Expert Systems in Transportation, Vol. 1,** Technical Research Center of Finland (VTT) and Organization for Economic Co-operation and Development (OECD), Espoo, Finland.

Roschke, P.N., and Premthamkorn, P. (1992) Expert System for Bridge Rail Retrofit and Design: System Integration and Validation, **2nd OECD Workshop on Knowledge-based Expert Systems in Transportation, Vol. 1,** Transportation Development Centre, Montreal, Canada.

Roschke, P.N., and Premthamkorn, P., Mitri, W.A., and Wang, B. (1991) Expert System for Bridge Rail Design, Report No. FHWA/TX-91/1240-2F, Texas Transportation Institute, College Station, TX.

Turing, A. (1963) Computer machinery and intelligence, **Computers and Thought**, Feigenbaum, E., and Feldman, J., editors, McGraw-Hill, New York, NY.

Yager, R. R. (1981) Concepts, Theory and Techniques: a New Methodology for Ordinal Multiobjective Decisions Based on Fuzzy Sets, **Decision Science**, American Institute for Decision Sciences, Atlanta, GA.