MODEL ACCREDITATION: A RATIONALE AND PROCESS FOR DETERMINING A NUMERICAL RATING

Saul I. Gass
College of Business and Management
University of Maryland
College Park, Maryland 20742

The Problem

Critical to the use of a mathematical model that describes a decision situation (the selection among competing alternatives) is the credibility or confidence that the decision maker has in the model and its ability to produce information that would be of value to the decision maker. In a previous paper [Gass and Joel 1981, p. 341-342], we noted the following:

> The role of model outputs in the decision process is based on the decision maker's understanding and evaluation of the total modeling process that has produced the outputs. Usually, the model outputs are modified and factored into an explicit or intuitive conceptual model of the decision maker. In an extreme case, the model can be allowed to define the decision. For decision makers, their confidence in a model is expressed by the influence the model's outputs had in the decision. ...

> Some may think that "confidence" is a quality of a model and a rough equivalent of validity. We emphasize model confidence not as an attribute of a model, but of the model user.

Thus, confidence will be considered from the point of view of the decision maker/user, rather than that of the analysts/developer, under the assumption that they differ. Model confidence is an expression of the user's total attitude toward the model and of the willingness to employ its results in making decisions.

...

> We take an extreme position by saying that a decision model without a designated user (which implies a specific use) has no basis upon which a confidence statement can be made, that is, the <u>a priori</u> confidence level is zero.

In [Gass and Joel 1981], we proposed a basic scoring scheme for determining the user's confidence level in a model. The process was based on attributes of a model which we felt were key to a model's confidence level. The attributes considered were the following:

-- Completeness and accuracy of underlying data.

-- Conceptual sufficiency of model specification.

-- Appropriateness of operating representation.

-- Appropriateness of embodied estimation methodologies.

-- Model sensitivity and stability.

-- Model performance compared to known outcomes.

-- Computer related model characteristics.

-- Any other model element or attribute which significantly influences the confidence in model results.

These attributes were summarized and transformed into seven model criteria against which the model was to be scored. The criteria considered were the following:

- -- model definition
- -- model structure
- -- model data
- -- computer model (program) verification
- -- model validation
- -- model usability
- -- model pedigree

To determine a confidence score, the decision maker, after defining and accepting the criteria, states a desired level of attainment (threshold values) for each criteria on a scale of one (low) to five (high). The scale represents the range of "not satisfying" to "fully satisfying" a criterion. Then, an independent model assessor or evaluator (not knowing the decision maker's desired levels of attainment for the criteria) reviews the model, its documentation and past use, its proposed use, and scores each criterion on the one to five scale. The decision maker is then presented with the evaluator's criteria scores and compares them against the desired threshold levels. The decision maker then makes a judgment call as to the confidence level that should be placed on the use of the model for the proposed use.

Although the above process is reasonably systematic, its implementation and final judgment call is open to discussion. For example, although the approach allows a decision maker to weight each criterion differently, the weights are not stated explicitly. Also, the sensitivity of the judgment call to changes in the implied weights and assessor's scores is not open to analysis. In addition, a final numerical score is not given. With this and other criticisms in mind, we propose an alternative approach for rating a model, as described below. Before we do so, we first discuss the concept of rating a model with a numerical score.

Rationale for a Numerical Score

The Military Operations Research Society (MORS) Working Group on Simulation Validation is currently investigating the desirability of developing a process for accreditating a simulation model, with accreditation defined as follows:

> Accreditation is the official determination that a computer model is acceptable for a specific use.

Accreditation is usually given with respect to a set of explicit standards. If the standards are fully met, the organization (e.g. university) or element (automobile mileage) is accredited. Comparable terms would be certified, credible, licensed. If the standards are not fully met, the item in question can receive limited and restricted accreditation. From this perspective, accrediting a model must be done with respect to the model's explicit specifications and the demonstration that the computer-based model does or does not meet the specifications. This demonstration is the purview of the model developers who must show that their work passes agreed to user/developer acceptance tests. If the modeling process was done properly and accompanied by appropriate documentation [see Gass 1984, 1987], accreditation of the model for its specified uses would follow naturally. However, it is universally recognized that the intersection between model specification, model development, model acceptance tests, and model documentation often leaves something to be desired.

Accreditation of a model must rely on a review of available documentation. Such a review, usually done by an independent third-party, is made against various criteria to determine the levels of accomplishment of the criteria. [See Gass 1983 for a discussion of independent model evaluation.] The review is also made with a specific user and use in mind, and should produce a report that gives guidance to the user on whether or not the model in question can be used with confidence for the designated use, that is, the model is or is not accredited for specific uses. In stating the criteria for accreditation, the user always has some implied weights that the user applies to the criteria in determining the accomplishment level associated with total criteria satisfaction. A linear, multiplicative weighting scheme is usually employed in such situations. That is, each criteria receives a level of satisfaction score and these scores are multiplied by their respective criteria weights. These products are then summed to produce a total satisfaction score [See Hwang and Yoon 1981 for a discussion of simple additive weighting].

The difficulty with this process is how to determine a set of numerical weights for the criteria and the numerical levels (scores) of satisfaction of the criteria. We usually impose the concept of consistency between the weights, where consistency means that when the criteria are compared to each other in a pairwise fashion, the weights are related by a transitivity relationship. However, it is recognized that in decision situations, when there are many criteria, with some of them subjective (e.g., prestige value of an automobile vs. maintenance cost), we are often inconsistent in our weighting. Inconsistent weights are in themselves not a bad thing, as long as we have some measure of inconsistency and can determine how sensitive the results are to any inconsistency.

The above discussion assumes that the results of the analysis (accreditation) is a numerical score for the model, when the model is to be used by a specific decision maker and use. There is some concern that such a number, which will be between 0 and 1, will be used incorrectly to place a value (a measure of worth) on the model or to compare models by these numbers. This is certainly a danger that has to be considered. But, to our mind, the value of developing a numerical score outweighs such dangers. First, by forcing the decision maker to determine criteria weights, we ensure careful thought to both the criteria used and the value of the criteria to the decision maker in approving accreditation. Second, with a set of numerical weights, sensitivity analyses can be made to demonstrate how the accreditation score changes with variations in the weights. Third, having numerical scores for criteria attainment will allow for sensitivity analyses to be done with respect to each criterion. And, fourth, although the assessor will submit a written report, scoring the attainment of the criteria numerically will cause the words in the report to have a specific interpretation and not subject to debate.

The accreditation score has no meaning by itself; it has to be combined with the written report, along with related sensitivity studies, so that the user can make a better judgment call as to whether to accredit the model.

We propose to use the Analytic Hierarchy Process (AHP) [Saaty 1980] to determine an accreditation score for a model, as described next.

Rating a Model Using the Analytic Hierarchy Process

The AHP was developed to resolve multi-criteria decision problems that have a small number of competing alternatives. It has been used successfully in a wide variety of decision situations [see the survey paper by Zahedi 1986]. The basic methodology of the AHP has been extended to include the rating of a large number of explicit alternatives. As described next, we use this ratings approach to determine a numerical accreditation rating for a model.

The basic scheme of the AHP is to cast the decision problem into a hierarchical structure which relates the goal of the problem (here, to determine an accreditation rating of a model) to the criteria (e.g., validation) and subcriteria (e.g., data validity) and the level of intensities (e.g., superior to poor) of each alternative with respect to the criteria and subcriteria. The criteria, subcriteria and intensities are given weights based on the AHP's procedure of pairwise comparisons (i.e., by eliciting answers from the decision maker to such questions as "How much more important is validation when compared to verification?"). The AHP methodology has been implemented in the software EXPERT CHOICE which we employ here.

As the MORS Working Group on Validation is currently investigating the elements and criteria that one would use to accredit a model, we illustrate our approach with a "strawman" set of criteria and note that the mechanics of the process do not depend on which criteria (and subcriteria) are used. However, critical to this and any other process of accreditation is a clear definition and understanding as to the meaning of the criteria.

In what follows, we assume some knowledge of the workings of the AHP and EXPERT CHOICE and illustrate the rating of a model by means of the AHP comparison matrices and the EXPERT CHOICE printouts. We use as criteria and subcriteria the following items:

CRITERIA AND THEIR SUBCRITERIA

** Specifications

** Verification

Mathematical Logic

Computer Code

** Validation

Theoretical Validity

Input Data Validity

Operational Validity

Face Validity

** Pedigree

Past Uses

Developers

** Configuration Management

** Usability

** Documentation

The basic hierarchy that has the goal of Accreditation Rating of Model is shown in Figure 1. The numbers in the boxes are the weights given to the criteria based on the decision maker's pairwise comparisons between the criteria. The pairwise comparison matrix is shown in Figure 2; the numbers reflect how the decision maker feels about the importance of each criteria with respect to the other criteria. For example, specifications versus verification is felt to be moderately more important; this converts to a value of 3.0 on the AHP ratio scale of 1 to 9. The AHP assumes that if specifications (the row heading) to verification (the column heading) has the value of 3.0, then verification to specifications has the reciprocal value of 1/3.0. Note that the diagonal elements are not filled in as they are naturally taken as 1.0 (equal) and the elements below the diagonal are not given as they are just the reciprocals of the numbers in the symmetric positions. Numbers in parentheses designate reciprocals. Thus, the (7.0) for usability to documentation means that documentation (the column heading) is more important than usability (the row heading) at a value of 7.0 (very strongly).

The EXPERT CHOICE software, using this pairwise comparison matrix, computes the normalized right-eigenvector associated with the maximum eigenvalue. Based on the theory of the AHP, these normalized values are the respective criteria weights. For our pairwise comparison matrix of Figure 2, these weights are shown at the bar graph part of Figure 2. If the pairwise comparisons were consistent, that is satisfied the transitivity relationship that element $a_{ij} = a_{ik}a_{kj}$, then the maximum eigenvalue would be equal to $\underline{n}$, the dimension of the comparison matrix. Deviation from this value gives a measure of inconsistency, which is here equal to 0.101, that is, a 10% deviation which is usually the upper limit that is acceptable. We perform similar pairwise comparisons for the subcriteria to determine their weights with respect to their criteria. Figure 3 shows the comparison matrix and the weights for the validity subcriteria; Figure 4 shows the subhierarchy and subcriteria weights for validation. Note that the subcriteria under verification and pedigree are taken to be of equal weight. In practice, all such comparison matrices are generated by the decision maker.

Figure 1 also shows for specifications, configuration management, usability, and documentation the intensity alternatives available to rate the respective criteria. For this application, we range them from superior to poor, but other interpretations and number of such alternatives can be used. Similar alternatives also are given to each of the subcriteria of verification, validation and pedigree. To rate a model, an assessor has to state the intensity level of each criteria or subcriteria, for example, the criterion of specifications is met in an average manner. These intensities also have to be weighted. The proper weights can be determined by the decision maker in a pairwise comparison manner or by stipulating absolute values. For discussion purposes, we have given absolute values for the intensity weights for all criteria and subcriteria as follows:

Superior	0.500
Above Average	0.300
Average	0.200

Below Average	0.100
Poor	0.000

These weights can be different for each criteria or subcriteria, as stipulated by the decision maker. The completed weighted hierarchy for our example is show by Figures 1, 4, 5, and 6.

Given a weighted hierarchy that corresponds to a decision maker's view of the criteria and their importance with respect to the proposed problem (model use) environment, the model is rated in the following manner. The assessor determines the levels of criteria or subcriteria intensities that are achieved for the model in question. These intensities are then entered into the EXPERT CHOICE ratings module which is a spreadsheet format for calculating the resulting total rating. This rating is just the sum of the products that result when the intensities are multiplied by the corresponding criteria and subcriteria weights. A set of ratings for fictitious models and intensity values are shown in Figure 7. The maximum rating (all superior) a model can get is 0.500, as shown in Figure 7.

Sensitivity analyses can be done along many dimensions. The criteria and subcriteria weights can be varied individually and/or collectively to determine how such changes impact the total rating. The intensity values for superior to poor can be changed and the impact measured. Additional criteria can be added or criteria removed with such changes reflected in the rating.

Validating the AHP Accreditation Model

The main reason for proposing a systematic method for determining a numerical accreditation rating is that such a process provides the decision maker, the assessor and others involved in the model development with an explicit model evaluation structure that enables them to have a consistent and focussed discussion about the use of the model. Again, we emphasize that the numerical rating by itself has no meaning. The process by which a numerical rating is obtained, here imbedded in the ratings approach of the AHP, imposes a modeling development and evaluation discipline on all concerned. The decision maker has to state the criteria and subcriteria of interest and determine their weights; the assessor has to translate the words of the assessment report into specific determinations for the criteria and subcriteria; and the developers (and their sponsors) have to employ a model development procedure that explicitly recognizes that future and alternative uses of their models depend on their producing models that can be rated.

We should give some concern as to its validity of the proposed AHP-based accreditation process; it is a model in itself. To address this concern, we propose the following research problems. - -

Standardized criteria and weights

Is it possible to identify models that are currently in use for which their is agreement as to their credibility and determine if the AHP numerical accreditation rating process does differentiate between these models? What we would like to find out is whether the AHP approach does or does not reflect what goes on in the real world of model use, and if there is a set of criteria and weights that would have general acceptance and wide applicability.

-- Can numerical accreditation ratings differentiate between models of a certain class

Is it possible to determine a set of criteria and weights that would apply to a set of models and across decision makers? For example, could this be done for weapons evaluation simulation models of a specific class, e.g., surface to air missiles?

-- Can decision makers really use the AHP approach

Experiments have to be made to determine if the proposed approach can work and has value to the decision maker.

References

Gass, S. I. 1981, "Concepts of Model Confidence," Computers and Operations Research, Vol. 8, No. 4, pp. 341-346.

Gass, S. I. 1983, "Decision-Aiding Models: Validation, Assessment, and Related Issues for Policy Analysis," Operations Research, Vol. 31, No. 3, pp. 603-631.

Gass, S. I. 1984, "Documenting a Computer-Based Model," Interfaces, Vol. 14, No. 3, pp. 84-93.

Gass, S. I. 1987, "Managing the Modeling Process: A Personal Perspective," European Journal of Operational Research, Vol. 31, No. 1, pp. 1-8.

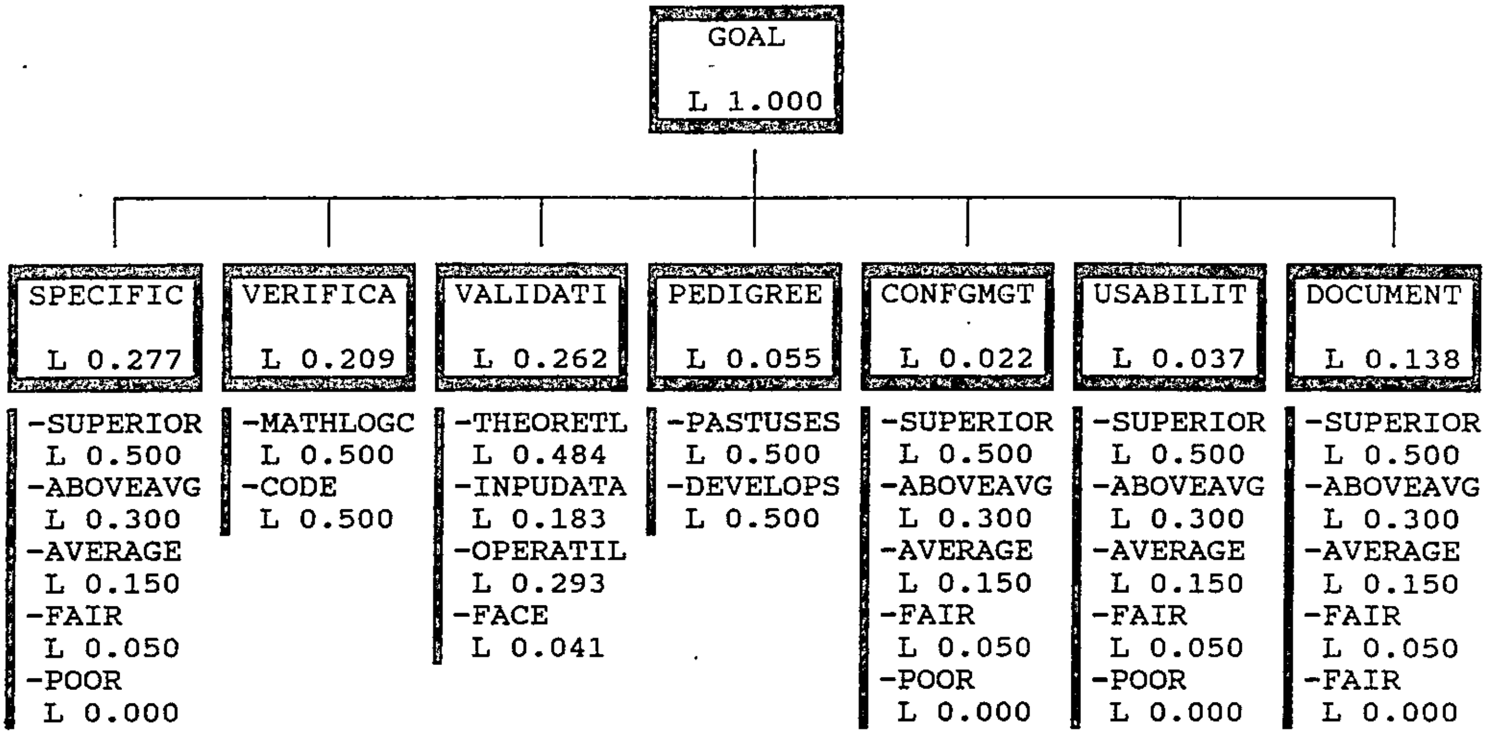
Hwang, C.-L. and Yoon, K. 1981, Multiple Attribute Decision Making, Springer-Verlag, New York.

Saaty, T. L. 1980, The Analytic Hierarchy Process, McGraw-Hill, New York.

Zahedi, F. 1985, "The Analytic Hierarchy Process - A Survey of the Method and Its Uses," Interfaces, Vol. 16, No. 4, pp. 96-108.

Figure 1

ACCREDITATION RATING OF MODEL



GOAL
L 1.000

SPECIFIC
L 0.277
-SUPERIOR
L 0.500
-ABOVEAVG
L 0.300
-AVERAGE
L 0.150
-FAIR
L 0.050
-POOR
L 0.000

VERIFICA
L 0.209
-MATHLOGC
L 0.500
-CODE
L 0.500

VALIDATI
L 0.262
-THEORETL
L 0.484
-INPUDATA
L 0.183
-OPERATIL
L 0.293
-FACE
L 0.041

PEDIGREE
L 0.055
-PASTUSES
L 0.500
-DEVELOPS
L 0.500

CONFGMGT
L 0.022
-SUPERIOR
L 0.500
-ABOVEAVG
L 0.300
-AVERAGE
L 0.150
-FAIR
L 0.050
-POOR
L 0.000

USABILIT
L 0.037
-SUPERIOR
L 0.500
-ABOVEAVG
L 0.300
-AVERAGE
L 0.150
-FAIR
L 0.050
-POOR
L 0.000

DOCUMENT
L 0.138
-SUPERIOR
L 0.500
-ABOVEAVG
L 0.300
-AVERAGE
L 0.150
-FAIR
L 0.050
-FAIR
L 0.000

ABOVEAVG	---	
AVERAGE	---	
CODE	---	DEBUGGING AND TESTING OF CODE
CONFGMGT	---	CONFIGURATION MANAGEMENT, CONTROLS
DEVELOPS	---	MODEL DEVELOPERS
DOCUMENT	---	AVAILABLE MANUALS AND THEIR CONTENTS
FACE	---	EXPERT OVERVIEW, NO REAL WORLD
FAIR	---	
INPUDATA	---	INPUT DATA AND PARAMETERS
MATHLOGC	---	MATH/LOGIC FROM SPECS TO CODE
OPERATIL	---	OPERATIONAL OUTPUTS VS. REAL WORLD
PASTUSES	---	PAST USES OF MODEL
PEDIGREE	---	ANTECEDENTS, PAST USES, DEVELOPERS
POOR	---	
SPECIFIC	---	SPECIFICATIONS, PROBLEM DEFINITION
SUPERIOR	---	
THEORETL	---	THEORY, ASSUMPTIONS, ALGORITHMS
USABILIT	---	RESOURCES, TRANSFERABILITY, MAINTENANCE
VALIDATI	---	VALIDATION THEORY, DATA, REAL WORLD
VERIFICA	---	VERIFICATION OF MATH/LOGIC AND CODE
L	---	LOCAL PRIORITY: PRIORITY RELATIVE TO PARENT

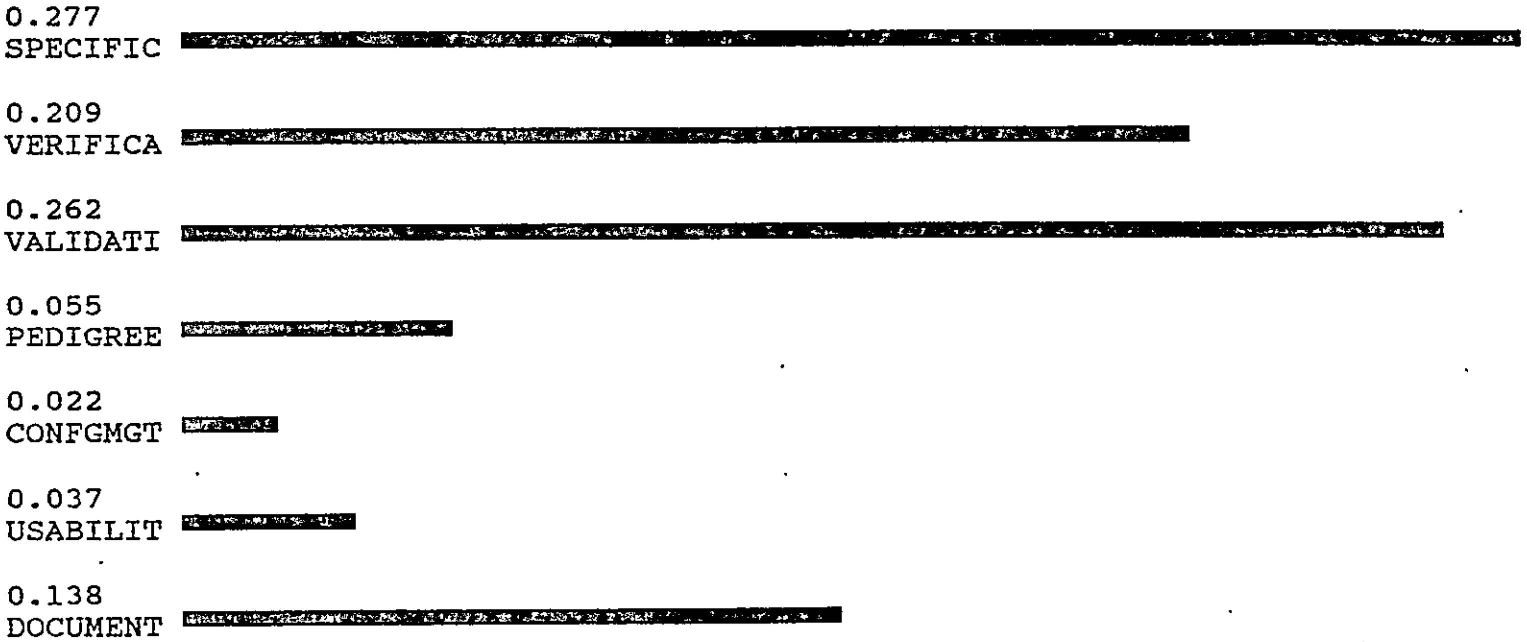
Figure 2

JUDGMENTS AND PRIORITIES WITH RESPECT TO GOAL

	SPECIFIC	VERIFICA	VALIDATI	PEDIGREE	CONFGMGT	USABILIT	DOCUMENT
SPECIFIC		3.0	1.0	7.0	7.0	5.0	2.0
VERIFICA			1.0	5.0	7.0	6.0	3.0
VALIDATI				6.0	7.0	7.0	4.0
PEDIGREE					5.0	3.0	(5.0)
CONFGMGT						(4.0)	(6.0)
USABILIT							(7.0)
DOCUMENT							

Matrix entry indicates that ROW element is ___
1 EQUALLY 3 MODERATELY 5 STRONGLY 7 VERY STRONGLY 9 EXTREMELY
more IMPORTANT than COLUMN element
unless enclosed in parenthesis.

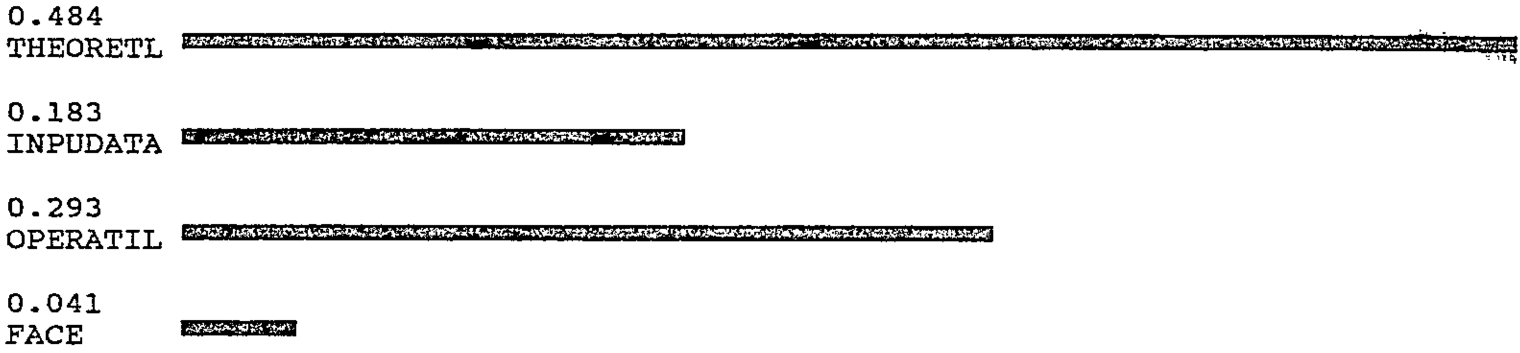
SPECIFIC :SPECIFICATIONS, PROBLEM DEFINITION
VERIFICA :VERIFICATION OF MATH/LOGIC AND CODE
VALIDATI :VALIDATION THEORY, DATA, REAL WORLD
PEDIGREE :ANTECEDENTS, PAST USES, DEVELOPERS
CONFGMGT :CONFIGURATION MANAGEMENT, CONTROLS
USABILIT :RESOURCES, TRANSFERABILITY, MAINTENANCE
DOCUMENT :AVAILABLE MANUALS AND THEIR CONTENTS



INCONSISTENCY RATIO = 0.101

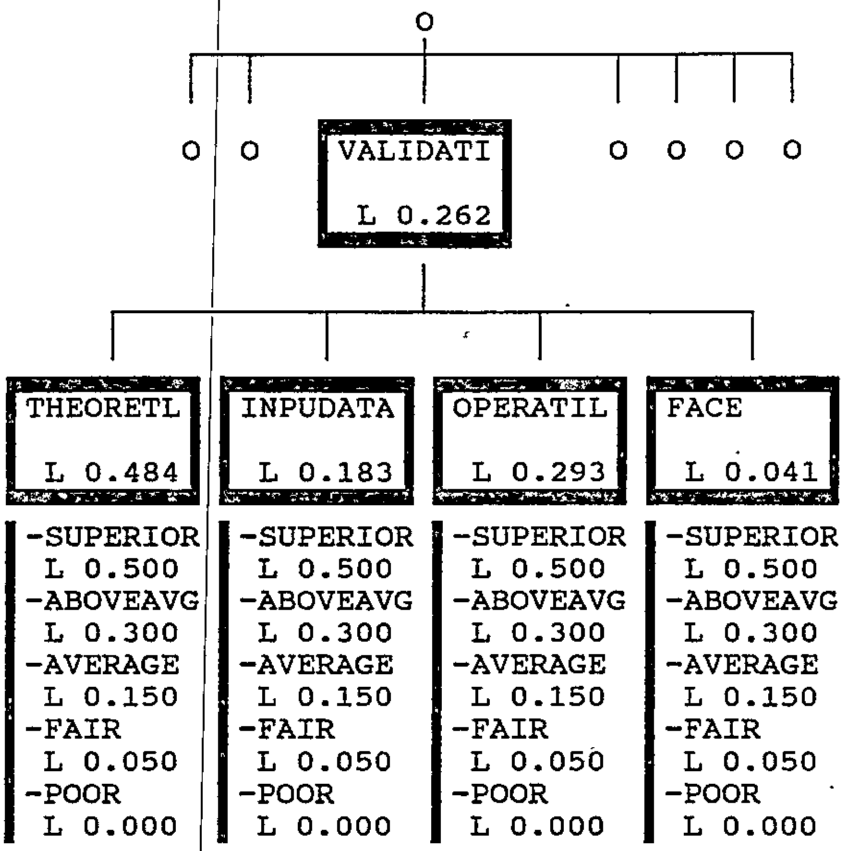
Figure 3

```
                 JUDGMENTS AND PRIORITIES WITH RESPECT TO
                              GOAL > VALIDATI

          THEORETL  INPUDATA  OPERATIL  FACE
 THEORETL           5.0       1.0       9.0
 INPUDATA                     1.0       5.0
 OPERATIL                               7.0
 FACE

Matrix entry indicates that ROW element is ___
  1 EQUALLY   3 MODERATELY   5 STRONGLY   7 VERY STRONGLY   9 EXTREMELY
more IMPORTANT than COLUMN element
    unless enclosed in parenthesis.


THEORETL :THEORY, ASSUMPTIONS, ALGORITHMS
INPUDATA :INPUT DATA AND PARAMETERS
OPERATIL :OPERATIONAL OUTPUTS VS. REAL WORLD
FACE     :EXPERT OVERVIEW, NO REAL WORLD
```



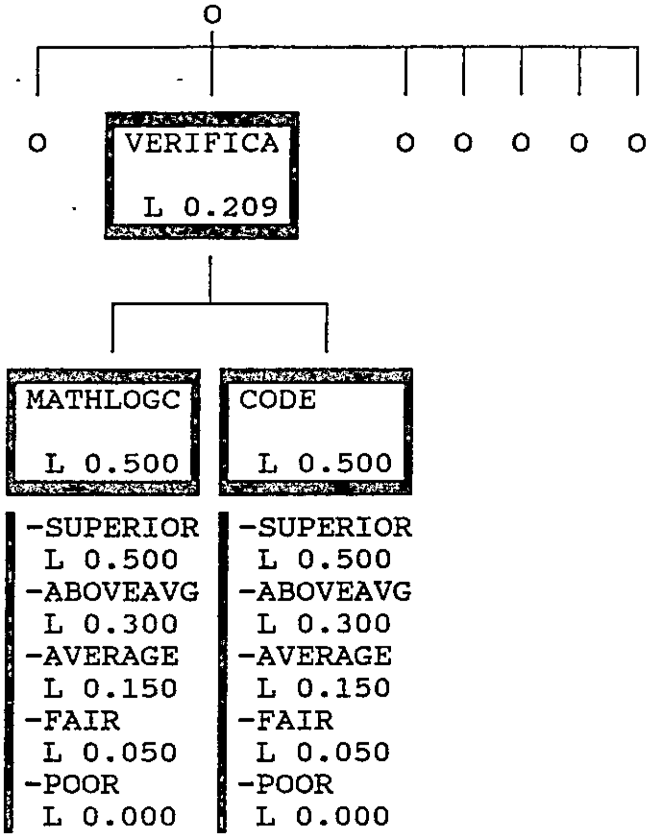
INCONSISTENCY RATIO = 0.093

Figure 4



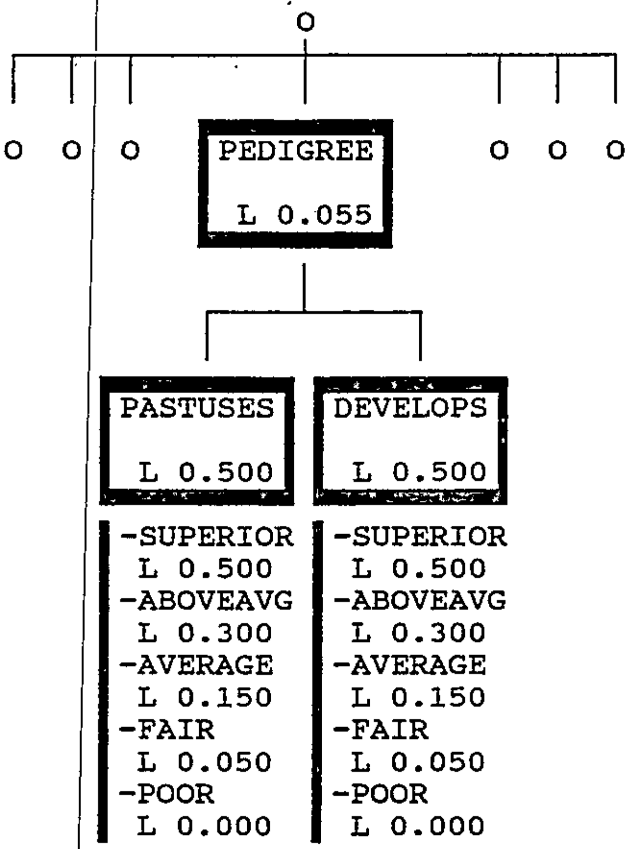
ABOVEAVG	---	
AVERAGE	---	
FACE	---	EXPERT OVERVIEW, NO REAL WORLD
FAIR	---	
INPUDATA	---	INPUT DATA AND PARAMETERS
OPERATIL	---	OPERATIONAL OUTPUTS VS. REAL WORLD
POOR	---	
SUPERIOR	---	
THEORETL	---	THEORY, ASSUMPTIONS, ALGORITHMS
VALIDATI	---	VALIDATION THEORY, DATA, REAL WORLD
L	---	LOCAL PRIORITY: PRIORITY RELATIVE TO PARENT

Figure 5



ABOVEAVG	---	
AVERAGE	---	
CODE	---	DEBUGGING AND TESTING OF CODE
FAIR	---	
MATHLOGC	---	MATH/LOGIC FROM SPECS TO CODE
POOR	---	
SUPERIOR	---	
VERIFICA	---	VERIFICATION OF MATH/LOGIC AND CODE
L	---	LOCAL PRIORITY: PRIORITY RELATIVE TO PARENT

Figure 6



ABOVEAVG	---	
AVERAGE	---	
DEVELOPS	---	MODEL DEVELOPERS
FAIR	---	
PASTUSES	---	PAST USES OF MODEL
PEDIGREE	---	ANTECEDENTS, PAST USES, DEVELOPERS
POOR	---	
SUPERIOR	---	
L	---	LOCAL PRIORITY: PRIORITY RELATIVE TO PARENT

Figure 7

	SPECIFIC	VERIFICA	VERIFICA	VALIDATI	VALIDATI	VALIDATI	VALIDATI	PEDIGREE	PEDIGREE	CONFGMGT
	.	MATHLOGC	CODE	THEORETL	INPUDATA	OPERATIL	FACE	PASTUSES	DEVELOPS	.
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
ALTERNATIVES	.2773	.1043	.1043	.1267	.0478	.0766	.0106	.0275	.0275	.0220
1 MODEL A	SUPERIOR	ABOVEAVG	AVERAGE	SUPERIOR	FAIR	AVERAGE	ABOVEAVG	SUPERIOR	FAIR	FAIR
2 MODEL B	ABOVEAVG	SUPERIOR	FAIR	AVERAGE	ABOVEAVG	POOR	ABOVEAVG	SUPERIOR	SUPERIOR	AVERAGE
3 MODEL C	SUPERIOR	AVERAGE	ABOVEAVG	FAIR	POOR	ABOVEAVG	ABOVEAVG	AVERAGE	ABOVEAVG	SUPERIOR
4 MODEL MAXIMUM	SUPERIOR	SUPERIOR	SUPERIOR	SUPERIOR	SUPERIOR	SUPERIOR	SUPERIOR	SUPERIOR	SUPERIOR	SUPERIOR

	USABILIT	DOCUMENT	
	.	.	
	.	.	
	.	.	
ALTERNATIVES	.0373	.1382	TOTAL
1 MODEL A	AVERAGE	ABOVEAVG	0.329
2 MODEL B	FAIR	ABOVEAVG	0.251
3 MODEL C	AVERAGE	ABOVEAVG	0.288
4 MODEL MAXIMUM	SUPERIOR	SUPERIOR	0.500