AD-A232 821

MENTATION PAGE

DTIC FILE COPY

Form Approved
OMB No 0704-0188

MAR 27 1991

1b RESTRICTIVE MARKINGS

1. SECURITY CLASSIFICATION AUTHORITY

2. DECLASSIFICATION/DOWNGRADING SCHEDULE

3 DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4 PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
The Regents of the University of California

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
Cognitive Science Program
Office of Naval Research (Code 1142PT)

6c. ADDRESS (City, State, and ZIP Code)
iversity of California, Los Angeles
ffice of Contracts and Grants Administration
s Angeles, California 90024

7b. ADDRESS (City, State, and ZIP Code)
800 North Quincy Street
Arlington, VA 22217-5000

8a. NAME OF FUNDING/SPONSORING ORGANIZATION Defense Advanced Research Projects Agency

8b. OFFICE SYMBOL (If applicable)

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
N00014-86-K-0395

8c. ADDRESS (City, State, and ZIP Code)
1400 Wilson Boulevard
Arlington, VA 22209-2308

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61153N	RR04206	RR04206-0C	442c022

11. TITLE (Include Security Classification)
Knowledge Engineering Report: An Expert System for Selecting Reliability Index

12. PERSONAL AUTHOR(S)
Li, Zhongmin

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM 7/1/86 TO 4/30/88	April 1988	21

16. SUPPLEMENTARY NOTATION

COSATI CODES		
FIELD	GROUP	SUB-GROUP
12	05	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Artificial intelligence, expert system shells

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This paper was completed as part of the Artificial Intelligence Measurement System (AIMS) focused on expert system shells. It documents the knowledge encoded in Reliability Index Knowledge Base and describes the user's needs and representations strategy on the particular topic, using M-1 shell.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Dr. Susan Chipman

22b. TELEPHONE (Include Area Code)
(703) 696-4318

22c. OFFICE SYMBOL
ONR 1142CS

Form 1473, JUN 86 Previous editions are obsolete. SECURITY CLASSIFICATION OF THIS PAGE

S/N 0102-LF-014-6603

Project Report #17

KNOWLEDGE OF ENGINEERING REPORT: AN EXPERT SYSTEM FOR SELECTING RELIABILITY INDEX

Zhongmin Li

Department of Instructional Psychology and Technology
School of Education
University of Southern California

April 1988

Artificial Intelligence Measurement System
Contract Number N00014-86-K-0395

Principal Investigator: Eva L. Baker

Center for Technology Assessment
UCLA Center for the Study of Evaluation

This research report was supported by contract number N00014-86-K-0395 from the Defense Advanced Research Projects Agency (DARPA), administered by the Office of Naval Research (ONR), to the UCLA Center for the Study of Evaluation. However, the opinions expressed do not necessarily reflect the positions of DARPA or ONR, and no official endorsement by either organization should be inferred. Reproduction in whole or part is permitted for any purpose of the United States Government.

Approved for public release; distribution unlimited.

Table of Contents

Knowledge Engineering Report:
An Expert System for Selecting Reliability Index

Zhongmin Li
University of Southern California

1. Introduction

This report documents the knowledge encoded in Reliability Index Knowledge Base (RIKB). The knowledge is presented in terms of the conceptualizations of the judgmental knowledge used in various level of decision-makings during a consultation.

We used a declarative knowledge representation mechanism to encode the knowledge necessary for selecting reliability index. In declarative representation of knowledge, the basic constructs of a knowledge base are production rules and attribute-value pairs. In RIKB, attributes represent properties, and characteristics of reliability indexes that affect the decision-makings in selecting reliability indexes for a given Criterion-Referenced Test. The value specifies the specific nature of the attributes in a particular situation (decision-making point). For example, INTENDED-USE (of test score) is an attribute, and the value could be decision, description, or program-evaluation. In the following description, the attributes will be indicated by upper case, and the value will be in lower case. An attribute value can either derived from rules, or directly get from user input. The form rule-i, where i is a number, indicates which rule is used to determine the attribute value. The form question-i, where i is also a number, indicates which question will be asked to get the information for the attribute. Refer to the knowledge base list for the exact wording of the rules, and questions.

The following sections are organized based on the four phases in selecting reliability index: (a) determining reliability category, (b) determining reliability index, (c) determining reliability design, and (d) reporting consultation results.

2. Determining Reliability Category

This section describes the knowledge related to the determination of which reliability category is suitable for a given Criterion-Referenced Test. This phase

includes four tasks: (a) query how knowledgeable is the user, (b) query how important of the test-score use, (c) determining intended use of test score, (d) determining subcategory of test score use, and (e) determining reliability category. Although the information gathered in the first two tasks does not contribute to the reasoning in task (c), (d), and (e), they serves as front-end to the knowledge system. The information will be used later during the consultation. The information is collected up front to represent the natural flow of the consultation. Rule-3 controls the order of the four tasks. For the purpose of scoping, rule-3 serves the function of screening cases that can be handled by the system.

2.1. Determining How Knowledgeable is the User

A user's knowledge about measurement in general, and reliability index in specific provides important information regarding to how detail the expert system's recommendation should be, and what consultation mode the system will be in.[1] This information is represented by the USER_KNOWLEDGEABLE attribute, which can be either "yes" or "no". Figure 2.1 presents the decision tree for determining the USER_KNOWLEDGEABLE attribute.

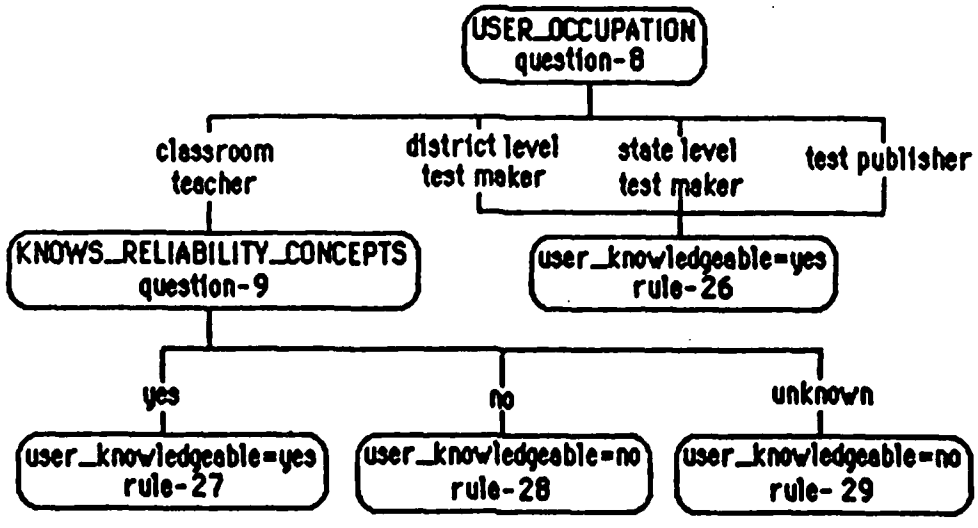


Figure 2.1. Determining USER_KNOWLEDGEALBE.

[1] In the first knowledge engineering session, Dr. Hambleton indicates that if a user is not knowledgeable about test measurement, he will recommend simple reliability index and related test design. Also, the consultation will be directed to tell the user what to do. For a knowledgeable user, he will recommend more statistically powerful indexes, and also provide several options to the user.

The decision is based on a taxonomy of possible users (USER_OCCUPATION attribute), which consists of four categories: (a) classroom teachers, (b) district level test maker, (c) state level test maker, and (d) test publisher.[1] In the current implementation, we treat category (b), (c), and (d) as if they are the same. It is assumed that a user from these three categories is knowledgeable about test measurement. If a user is a classroom teacher, the expert system uses question-9 (KNOWS_CONCEPT_OF_RELIABILITY attribute) to query that whether the user is knowledgeable about test measurements. The default value is that the user has little knowledge about test measurement (rule-29 will inform the user if the default value is used).

2.2. Query Importance of Test Score

The importance of test score use is judged based on what kinds of decisions will be made from the test results. Thus, the importance of score is represented by the IMPORTANCE_OF_RESULT attribute, and how the test results will be used is represented by the HOW_TEST_USED attribute. Figure 2.2 shows a decision table for determining the IMPORTANCE_OF_RESULT attribute.

HOW_TEST_USED question-1	day-to-day classroom management	forming instructional groups	longer-term decisions	credentialing exams
IMPORTANCE_OF_RESULT (rule used)	25 rule-4	60 rule-5	80 rule-6	95 rule-7

Figure 2.2. Determining IMPORTANCE_OF_RESULT.

This decision table provides a possible scale for judging the importance of test results.[2] In the "longer-term decisions" category, there are several sub-categories such as assigning students to special programs, assigning mid-term grades, and assigning final grades, etc. These information are included in the question-1.

---

[1]Berk (1984) classified three types of CRT practitioners: (a) classroom teachers, (b) district and state level test makers, and (c) test publisher. We separate item (b) to make the taxonomy contains four categories in anticipation that different treatments might be necessary for district and state level test makers.

[2]The taxonomy of how test score will be used is based on Hambleton's background nodes for the project (December 22, 1987, p.6). However, the notes only provides the order of the importance of test results. Therefore, the numbers in the table only serve the purpose of representing the order of importance.

2.3. Query Intended Score Use

The intended uses of test score (USE_CATEGORY) consists of three categories: (a) decision, (b) description, and (c) program evaluation (Hambleton, 1987). Figure 2.3 shows the decision tree used to determine the intended use of test result (USE_CATEGORY attribute).

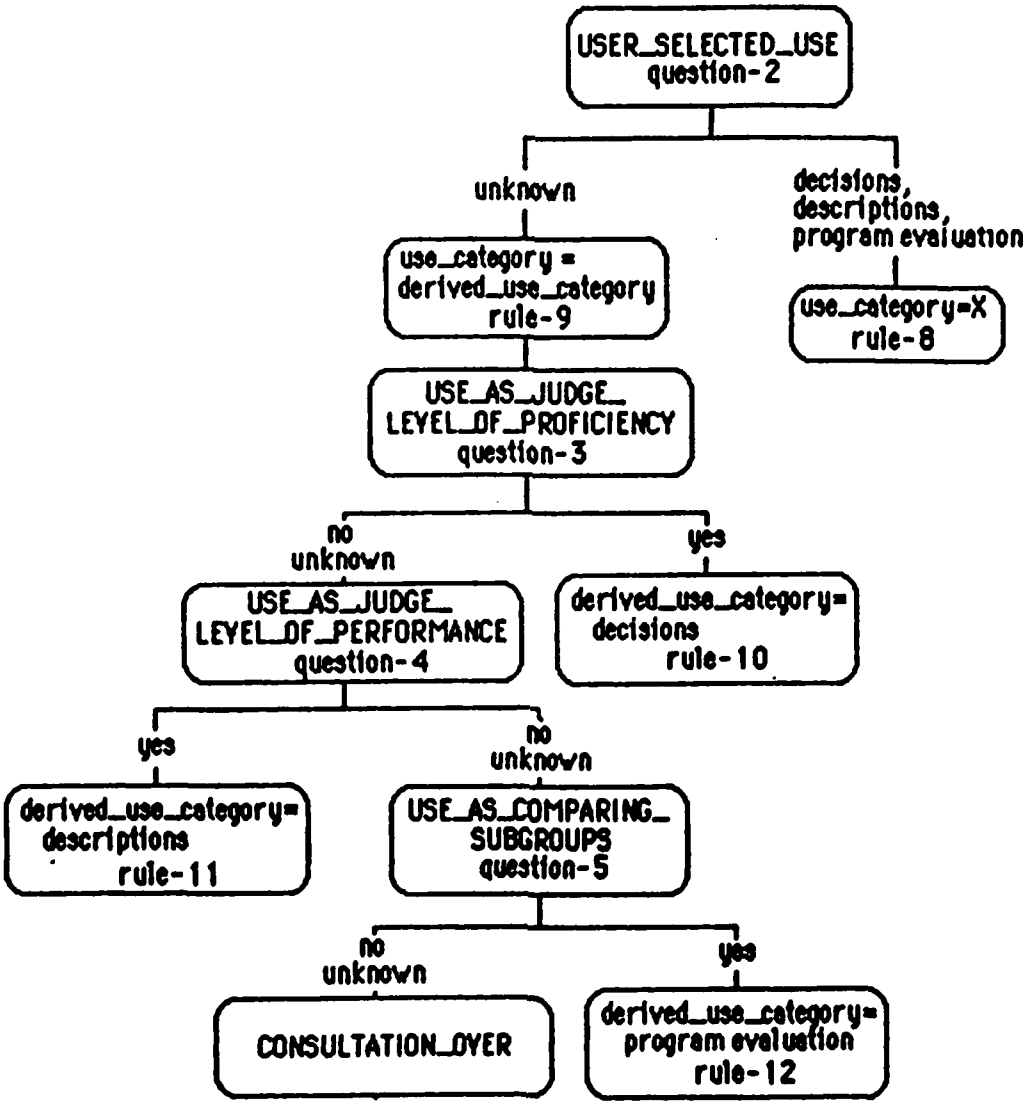


Figure 2.3 Determining USE_CATEGORY.

Although users generally know the intended use of test score, they may still have difficulty to select from the category due to unfamiliar with the definition of the category or this classification system. Thus, this piece of consultation is

designed to provide two levels of assistance to the users. At the first level, definitions of each category is presented when asking the users to select an intended test score use. If the users still have difficulty, they can type in "unknown", the system will enter a lower level query mode to provide more assistance.

In the current implementation, there is one question corresponding to each of the three test score use.[1] A confirmation for the users to a question will lead to the conclusion to the test score use corresponding to the question. The knowledge coded in rule-10 to rule-12 assures that there will be only one test score use[2] (USE_CATEGORY is a single value).

2.4. Determining Subcategory of Test Use

There are two different situations that reliability information is valued: (a) in the test development process, and (b) as one of the criteria used to evaluate an intended test use (Hambleton, 1988). USE_SUBCATEGORY is the attribute that represents this information. Figure 2.4 is the decision tree used to determining the value for the attribute.

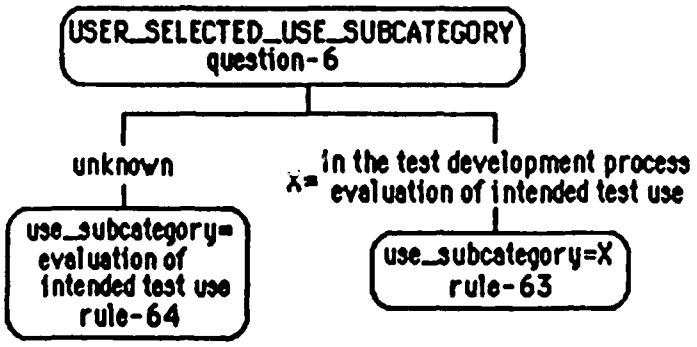


Figure 2.4. Determining USE_SUBCATEGORY.

Among the two subcategories of intended test use, the default one is for "evaluation of intended test use" because it represents most cases that reliability information is assessed. Also, current version of the system can consult on the "evaluation of intended test use" subcategory.[3] Therefore, if a user does not know the subcategory of intended test use, the system assumes that it is used for "evaluation of intended test use".

---

[1]The three questions for determining intended test use are elicited during first knowledge engineering session with Dr. Hambleton.

[2]Although it has been discussed in the knowledge engineering session that there might be multiple test score uses, the prototype expert system assumes that only one USE_CATEGORY is appropriate for a given test.

[3]It is not clear yet that how different it is between using the test score in test development process and using the score for evaluation of an intended test use. Since the concept of using the test score in test development process is relatively new, it is not included in the prototype system.

2.5. Determining Reliability Category

Berk (1984) categorizes reliability index for Criteria-Referenced Test into three categories: (a) threshold loss, (b) squared-error loss, and (c) domain score estimation. This categorization is used to limit the searching space for selecting reliability indexes. Hambleton (1988) proposes the test use category as (a) decisions, (b) descriptions, and (c) program evaluation. The relationship between the two categories is straight-forward. The descriptive use of test corresponds to domain score estimation category. The decision category corresponds to threshold loss and squared-error loss function categories. The correspondence for program evaluation use of test has not yet been elicited.[1]

Figure 2.5 shows a decision tree which concludes RELIABILITY_CATEGORY attribute.

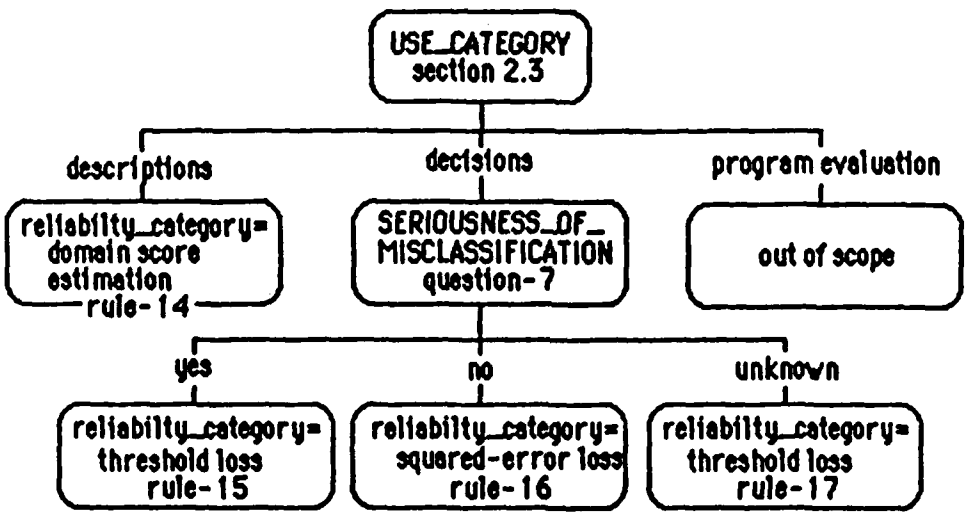


Figure 2.5 Determining RELIABILITY_CATEGORY.

Among the three reliability categories, the decision use of test requires special attention because it contains two reliability categories: threshold loss, and squared-error loss. The decision[2] is based on whether the losses associated with decision errors are equally serious (threshold loss) or not equally serious

[1]Since which reliability indexes are suitable for program evaluation use of test is still not clear, the prototype system will not handle the consultation on this category. Therefore, no further efforts are spent on eliciting related knowledge.

[2]In Berk (1984), three characteristics of reliability categories are provided (table 9.1, p. 236). They are (a) score interpretation, (b) type of decision or information required for decision, and (c) losses associated with decision errors. Each one of these can be used to distinguish the threshold loss or squared-error loss category. We used (c) in the system based on the knowledge elicited at second knowledge engineering session with Dr. Hambleton. Do we need to consider other two characteristics?

(squared-error loss). The default reliability category for decision category is threshold loss category because it is the most used reliability index.

3. Determining Reliability Index

This section describes the knowledge coded to determine reliability index suitable for a given test use. The decision is based on the reliability categories. Since "program evaluation" use of test score does not yet have a corresponding reliability category, it will not be considered in the decision-making included in this phase.

3.1. Reliability Index for Squared-Error Loss Function

The squared-error loss function deals with the consistency of measurements or test scores (Berk, 1984). The decision involved in determining reliability index for squared-error loss function is very simple. So far, there is only one index required for the function. That is "standard error of measurement".[1] Therefore, it is an one-to-one mapping[2] from intended use of test to reliability index category. This mapping is coded in rule-30.

3.2. Reliability Index for Threshold Loss Function

The threshold loss function focuses on the consistency of classification of students as masters and non-masters of an instructional objective based on a threshold or cut-off score. There are two types of reliability indexes: (a) decision consistency, and (b) kappa. Decision consistency estimate is relatively easy to compute, and interpret. It is basically recommended for use in every case. Kappa "provides estimate of level of agreement corrected for chance" (Hambleton, 1988), but it is harder to interpret than decision consistency. Therefore, it usually servers as an add-on reliability statistics to provide more information besides the decision consistency for important test. The decision regarding which index will be recommended depends on how important is the test, which is represented by HOW_TEST_USED attribute. As shown in figure 2.6, credentialing exams are very important, thus, both decision consistency and kappa are recommended. On the other hand, day-to-day classroom management and forming instructional groups

---

[1] Berk (1984) lists two indexes under the squared-error loss function category. Both of them belong to the general category, standard error of measurement.

[2] In second knowledge engineering session, we also discussed other indexes for squared-error loss function. However, none of them are very significantly used in practice. In this prototype system, only standard error measurement statistics for squared-error loss function will be considered.

are less important. Thus, only decision consistency is recommended for these two types of test use. The situation for longer-term decisions is little more complex. It involves two attributes: (a) HISTORICAL_DATA, and (b) USER_KNOWLEDGEABLE.

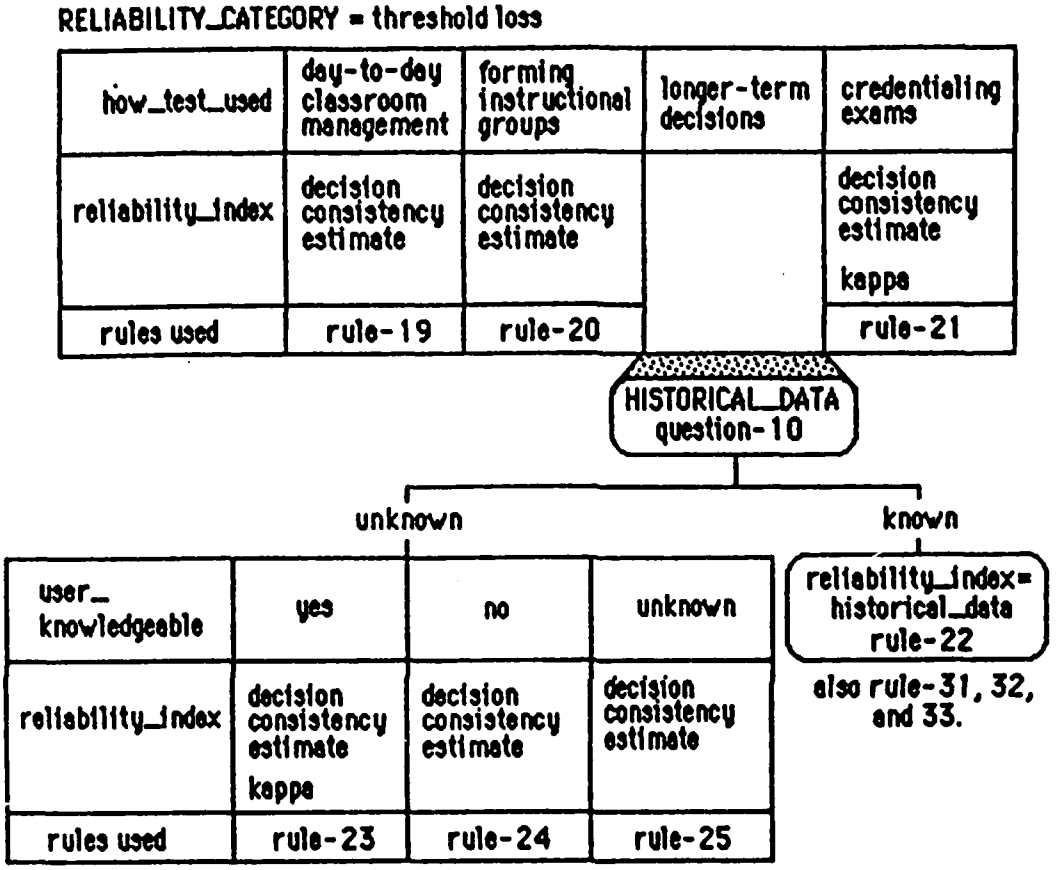


Figure 2.6. Selecting RELIABILITY_INDEX

The decision is that if the user knows that reliability data has been collected for the test, then the system recommends same indexes as the ones used before. Otherwise, information on users' knowledge about test measurement is used to determine reliability index. Basically, for a knowledgeable user, the system will recommend both decision consistency and kappa. Otherwise, only decision consistency estimate will be recommended.[1]

[1]These rules are elicited from second knowledge engineering session with Dr. Hambleton. It is still to simple here, and may not fully represent the knowledge used by a domain expert in decision-makings. More elicitation is needed for further expanding of the system.

4. Determining Reliability Design

This section documents the knowledge for selecting which reliability design is appropriate for a given test. Rule-34 is the control rule for determining reliability design.

4.1. Reliability Design, Number of Administrations, and Number of Forms

There four possible reliability designs: (a) test-retest with equivalent forms, (b) test-retest with same form, (c) single administration with one form, and (d) single administration with parallel forms.[1] Which reliability design is appropriate depends on the number of test administrations that a user is willing to give, and the number of test forms available. Figure 2.7 presents a decision table for recommending a reliability design.

		Number of Administration	
		one	more than one
Number of Form	one	single administration with same form rule-37	test-retest with same form rule-36
	> one	single administration with parallel forms rule-38	test-retest with equivalent forms rule-35

Figure 2.7. Determining Reliability Design.

In the figure, an entry "one" means that there will be only one form available or one administration possible. "More than one" means multiple forms are available, and two test administrations are possible. Assuming the information on number of administrations and forms are available, each of the four rules concludes about one particular reliability design. These rules make the relationships between reliability designs and number of test administration, and forms more clear for system's explanation purposes. Lengthy consultation might be required to determine number of possible test forms, and test administration, which will be described in later sections.

4.2. Determining Number of Test Administrations

If the reliability index selected is "standard error of measurement" , there is only one test administration required as coded in rule-39. This applies to both

[1]Hambleton (1988) provides three possible reliability designs for CRT test: (a) test-re-test with the same form, (b) test-re-test with equivalent forms, and (c) single administration. We further split item single administration into two designs based on the number of forms required for the design.

squared-error loss, and domain score estimate categories.[1] However, if the reliability index selected is "decision consistency estimate" or "kappa", then the number of possible test administrations could be either one or two. Thus, the number of test administrations required for "threshold loss" category depends on the reliability indices selected, the nature of materials to be tested, and other administrative considerations. The following discussions focus on the knowledge for determining possible number of test administrations.

Selecting the number of test administration consists of two steps: (a) query about the possibilities of multiple test administrations, and (b) asks user to confirm the multiple administration if it is recommended in step (a). The key attributes for step (a) is MULTIPLE_ADMINISTRATION_POSSIBLE, and for step (b) is USER_CONFIRMED_MULTIPLE_ADMINISTRATION.

Figure 2.8 shows the decision tree for determining whether multiple administration of test is possible.

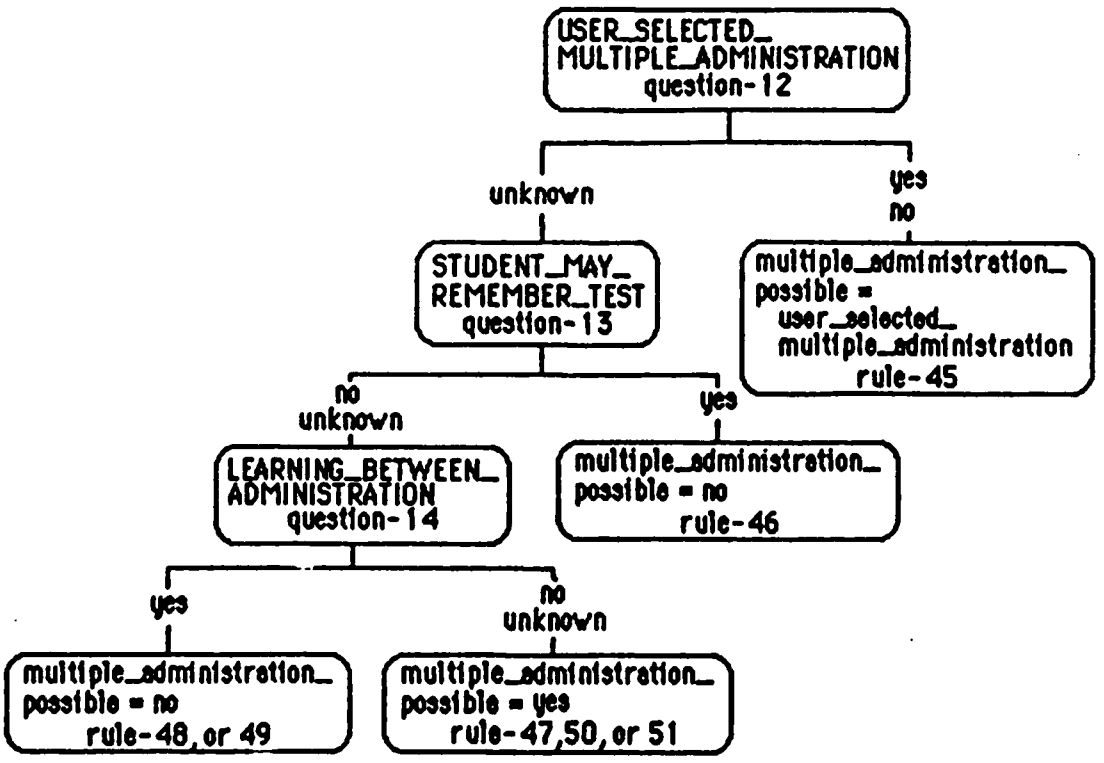


Figure 2.8. Determining MULTIPLE_ADMINISTRATION_POSSIBLE.

[1]Berk(1984) summarizes that there are two indices for squared-error loss category. Both are based on standard error of measurement (table 9.3, p.248-249). For domain score estimate categories, three of the six indices listed in table 9.4 (p.253-255) are labeled as standard error of measurement.

In this decision tree, system first asks the user to specify whether multiple administration of the test is possible (question-12). If the user answers "yes", or "no", then the value is passed to the MULTIPLE_ADMINISTRATION_POSSIBLE attribute. However, if the user answers "unknown", the decision will be based on the user's responses to two lower level questions: whether students may remember the test items (question-13) or whether learning may occur if the test is administered twice (question-14).[1] To both questions, an "unknown" response is assumed to be "no", but appropriate messages[2] will be displayed to inform system's assumption.

For Credentialing examinations, it is impossible to administer the tests more than once. This knowledge is implemented in rule-44. In this case, the decision tree demonstrated in figure 2.8 will not be invoked.

Once the information about whether it is possible to administer a test more than once, the information will be used in step (b) to determine the number of test administration for the test. Figure 2.9 shows the decision table.

RELIABILITY_CATEGORY = threshold loss

multiple_ administration_possible	yes	yes	yes	no
user_confirmed_ multiple_administration	yes	no	unknown	N/A
number_of_administration	more than one	one	one*	one
rules used	rule-40	rule-41	rule-42	rule-43

*Rule-42 contains the message that assumes no multiple test administration.

Figure 2.9. Determining NUMBER_OF_ADMINISTRATION.

To ensure the flexibility, the system asks users for confirmation when multiple test administrations are possible. The users may either confirm system's recommendations or reject them. If the users answer "unknown", then single administration is suggested.

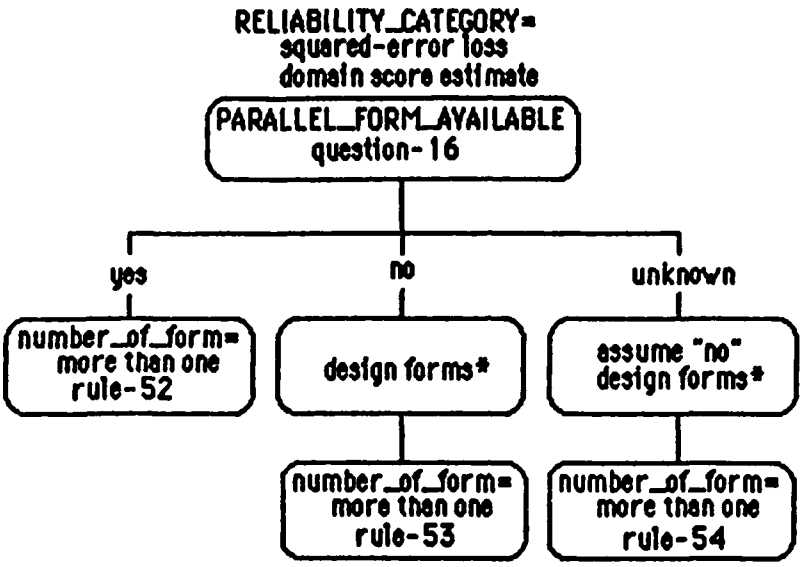
---

[1]The questions do not be limited to two. The code structure is flexible enough to include other knowledge necessary to help a user to determine whether multiple administration of test is possible in a given situation. The two questions coded in the current version of the system are from second knowledge engineering session with Dr. Hambleton.

[2]This is why some boxes in figure 2.8 contain more than one rule. Rule-49 and 51 contain the messages that inform the users about system's estimation when an "unknown" response is encountered.

4.3. Determining Number of Forms

Parallel forms are required if the reliability index category is either "squared-error loss" or "domain score estimate". Figure 2.9 shows a decision tree for determining number of forms for the two index categories mentioned.



*"design forms" may invoke the rules to help users design parallel forms[1]

Figure 2.9. Determining NUMBER_OF_FORM for "Squared-error Loss" or "Domain Score Estimate" Categories

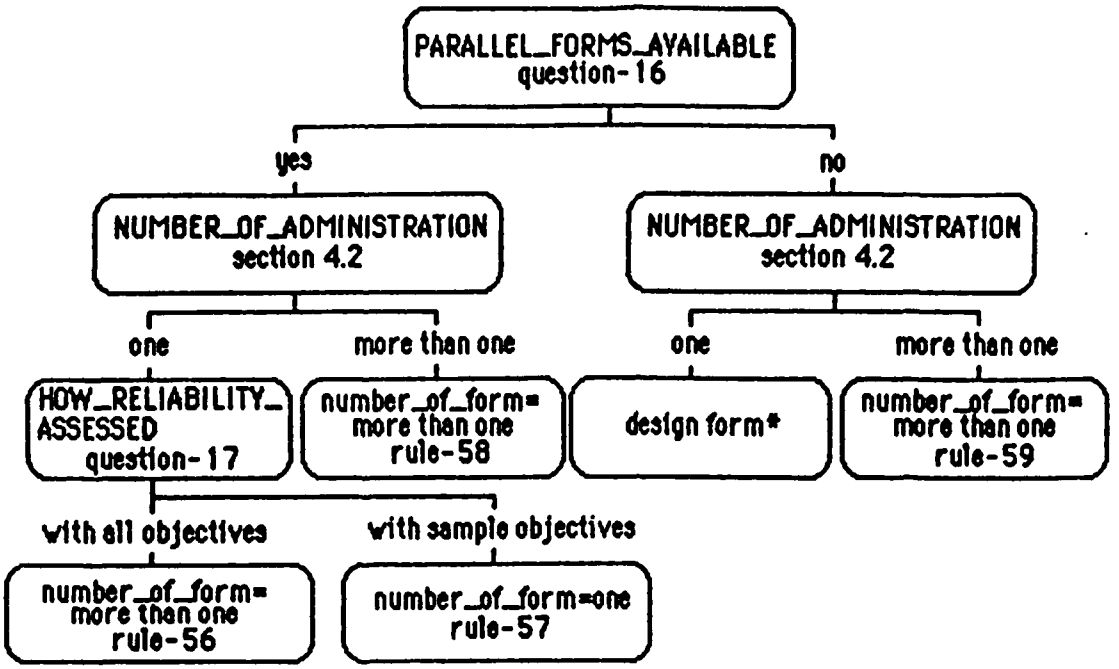
Rule-52, 53, and 54 all conclude multiple forms required. However, rule-53 and 54 allow further addition of knowledge regarding form design such as dividing an existing form, or recommending strategies for test length selection and items determination.

For "threshold loss" category, the matter becomes a bit complex because either simgle form or multiple forms can be used to accessing reliability indices. Figure 2.10 presents a decision tree for this category. The line of reasoning depends on whether parallel forms are available. "If parallel forms are available, the design for a reliability study is probably clear"[2]. However, if the parallel forms are available, but both forms cannot be administered, modifications are needed to the design (Hambleton, 1988, p. 6).

---

[1]This step is included in rule-53 or 54. It is designed for further expansion of the knowledge base to provide assistance in determining form length or test length. Right now, it does not do anything.

[2]I interpret this as if the parallel forms are available then recommend the use of the parallel forms. This interpretation is simplified, but more knowledge need to be elicited to make the reasoning more complex.



* It is not clear yet whether the system should assume single form here or query for possible parallel forms.

Figure 2.10. Determining NUMBER_OF_FORMS for "Threshold Loss" Category

> [Two possible designs are:] (a) creating several forms, where each form includes parallel items to measure a fraction of the objectives; in this way, the reliability of all objective scores can be assessed but with a fraction of the total examinee pool, [and] (b) assessing the reliability of a representative sampling of objectives and then generalizing the findings to describe all objectives. With this second design, single-administration estimates of reliability can be computed for the remaining objectives. (p. 6)

This information is represented by attribute HOW_RELIABILITY_ASSESSED (question-17). In the current implementation, if parallel forms are not available, and multiple administration is possible, rule-59 concludes to use simple form. Later on, we should replace this rule with another line of reasoning, where the system will invoke a discussion with the users about the possible of creating parallel forms from the exiting simple form.[1]

[1]We discussed this in both knowledge engineering session with Dr. Hambleton. However, there are still some knowledge needed to be elicited for extending rule-59.

5. Reporting Consultation Results

Hambleton (1988, p. 8) listed ten statistics that should be reported in any comprehensive reliability study: (a) reliability index, (b) cut-off score, (c) sample size, (d) descriptive information about the sample, (e) test length, (f) test score mean and standard deviation, (g) test score distribution, (h) if decisions are made on two occasions, the percent of examinees, and (i) standard errors. However, it is still not clear about the judgmental knowledge involved in determining which statistics must be reported, which statistics may be reported, and what is the relationships between reporting statistics and importance of test uses.

Besides, based on Berk 's chapter (1984) we intended to add some new items in the report such as the definition for each indices recommended, the initial sources of these indices, and the advantages and disadvantages of these indices.

Appendix A

Questions Generated by the Reliability Index Expert System

This appendix listed all the questions generated by the Reliability Index expert System. All questions in the system are asked in multiple-choice format. Therefore, possible responses to the questions are listed following the text of each question. Each entry is labeled by Question-i, where i is a number. The attribute name represented by the question is enclosed in the parenthesis after the label.

Question-1 (HOW_TEST_USED):

A Criteria Referenced Test may be used for many purposes. It can be used for day-to-day classroom management, forming instructional groups, longer-term decisions, and credentialing exams. Some examples of longer-term decisions are assigning students to special programs, assessing mid-term and final grades.

Which of the following categorized your use of the test?

1. day-to-day classroom management
2. forming instructional groups
3. longer-term decisions
4. credentialing exams

Question-2 (USER_SELECTED_USE):

There are three major categories of CRT score uses:
DESCRIPTIONS - We make statements such as the student is performing at an 80% level in the domain of content of interest. Such statements are often made for each objective measured in a CRT.
DECISIONS - We often desire to assign examinees to two or more mastery categories (e.g., pass/fail, mastery/non-mastery). The classifications may be used to ward diplomas, licenses, or certificates, or to monitor student performance an on objective based instructional programs.
PROGRAM EVALUATION - CRTs are often used in curriculum or program evaluation studies. Average scores on each objective or groups of objectives are reported for groups (and subgroups) of examinees administered the test.

Which of the following categorized your CRT score use?

1. descriptions
2. decisions
3. program evaluation
4. unknown

Question-3 (USE_AS_JUDGE_LEVEL_OF_PREFICIENCY):
Would you use the test results to judge the proficiency of individual students?

1. yes
2. no

Question-4 (USE_AS_JUDGE_LEVEL_OF_PERFORMANCE):
Would you use the test results to judgement the performance of individual students?

1. yes
2. no

Question-5 (USE_AS_COMPARING_SUBGROUP)
Would you use the test results to compare among subgroups?

1. yes
2. no

Question-6 (USER_SELECTED_USE_SUBCATEGORY):
There are two different situations where reliability information is valued:
In the test development process - reliability information collected during a field test can be invaluable in advising on desirable test lengths and in judging the soundness of the test items.
As one of the criteria used to evaluate an intended test use - reliability information influences the confidence that users have regarding the test scores and related decisions.

Which of the following situation is best applicable to your test?

1. in the test development process,
2. evaluation of intended test use
3. unknown

Question-7 (SERIOUSNESS_OF_MISCLASSIFICATION):

Are all misclassifications of test score approximately equal in their impact?

1. yes
2. no
3. unknown

Question-8 (USER_OCCUPATION):

Which of the following best describes your occupation as a test developer/user?

1. classroom teacher
2. district level test maker
3. state level test maker
4. test publisher

Question-9 (KNOWS_CONCEPT_OF_RELIABILITY):

Do you have the general knowledge of test reliability?

1. yes
2. no
3. unknown

Question-10 (HISTORICAL_TECHNICAL_DATA_EXISTS):

Were any technical data collected on the test before?

1. yes
2. no
3. unknown

Question-11 (USER_SPECIFIED_TECHNICAL_DATA):

What sort of technical data have been collected?

1. decision consistency estimate
2. kappa
3. both of the above

Question-12 (USER_SELECTED_MULTIPLE_ADMINISTRATION):

Is it possible to administer the test more than once?

1. yes
2. no
3. unknown

Question-13 (STUDENT_MAY_REMEMBER_TEST):

Will students be able to remember questions or aspects of the questions such as the passages, or diagrams in the test?

1. yes
2. no

Question-14: (LEARNING_BETWEEN_ADMINISTRATIONS):

If the test is administered twice, will some learning take place between the two administrations?

1. yes
2. no

Question-15 (USER_CONFIRMED_MULTIPLE_ADMINISTRATION):

From the information you give, I think that it is possible to administer the test twice. Since for single-administration, strong assumptions must be made in order to obtain a reliability estimate, I would recommend that you use a two administration design.

Do you think it is practical in your situation to administer the test twice?

1. yes
2. no

Question-16 (PARALLEL_FORM_AVAILABLE):

Are there parallel or equivalent forms available for the test?

1. yes
2. no
3. unknown

Question-17 (HOW_RELIABILITY_ASSESSED):

Since parallel forms cannot be administered, modifications are needed to the reliability design. Reliability can be assessed in two ways:
(1) creating several forms, where each form includes parallel items to measure a fraction of the objectives; in this way, the reliability of all objective scores can be assessed but with a fraction of the total examinee pools.
(2) assessing the reliability of a representative sampling of objectives and then generalizing the findings to describe all objectives. Thus, single-administration estimates of reliability can be computed for the remaining objectives.

How do you want the reliability be assessed?

1. with all objectives
2. with sample objectives