A COMPARISON OF LEARNING HIERARCHY VALIDATION TECHNIQUES USING SIMULATED DATA

by

Ronald D. Owston

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the Ph.D. degree in Education

UNIVERSITY OF OTTAWA
OTTAWA, CANADA, 1978

ACKNOWLEDGEMENTS

This thesis was prepared under the supervision of Professors Matin Cooper, Marvin Boss and Jean Paul Dionne. Professor Cooper acted as committee chairman.

The writer is indebted to the Canada Council for its generous financial support of this research.

CURRICULUM STUDIORUM

Ronald D. Owston was born July 17, 1945, in Toronto, Ontario. He received the Bachelor of Science degree from Sir George Williams University in 1967, the Bachelor of Education degree in 1972 from the University of New Brunswick, and the Master of Education degree in 1973 from the University of New Brunswick. The title of his thesis was Small Group Problem Solving in High School Mathematics.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES

INTRODUCTION

A number of techniques have been suggested in the literature for testing the validity of learning hierarchies. These techniques may be criticized for not fully taking into account problems associated with measurement error and skill difficulty, and, with one exception, for not having known sampling distributions.

In order to overcome these problems an alternative validation technique based upon the method of maximum likelihood is proposed in this report. This technique is then compared to existing techniques using simulated data. The purpose of the comparison is to determine the decision error which may be expected with each technique on various populations and sample sizes, and to determine which technique is the most accurate overall.

This investigation should be particularly important in providing an empirical assessment of learning hierarchy techniques. Presently there is a paucity of information on the effectiveness of these techniques, hence the researcher faces a dilemma when choosing a technique to validate a hierarchy.

In the first part of this report the literature on learning hierarchy theory and validation methodology is reviewed. The strengths and weaknesses of existing validation techniques are then discussed. From this discussion the need for an alternative validation technique is shown and such a technique is proposed. This part is then concluded with a statement of the problem and the research hypothesis.

The experimental design is described next. This description includes the method used for generating simulated data by computer, the selection of population parameters and sample sizes and the procedure by which the hypothesis was tested.

The results of the investigation are presented and discussed following the experimental design. First the results of the test of the hypothesis are reported and then each technique is discussed in detail and specific recommendations on its utility are made.

Finally several implications for future research are given, and in an appendix, the computer programs used in the research are listed and the raw data are tabulated.

CHAPTER I

BACKGROUND OF THE STUDY

The concept of a learning hierarchy was first introduced over a decade ago by Robert Gagné. Since then numerous claims have been made about the usefulness of learning hierarchies in the field of education. Despite these claims, little progress has been made in applying learning hierarchy theory to practice. Part of the reason for this appears to be the lack of an entirely satisfactory technique for testing the validity of hypothesized hierarchies, and a dearth of information about the effectiveness of available validation techniques. The aim of this investigation is to empirically examine several of the most often used and theoretically most accurate techniques available with the hope of providing the researcher with specific recommendations when these techniques may be expected to correctly detect valid and invalid hierarchies.

The first chapter begins with a review of learning hierarchy theory. Following this, the problems encountered in testing the validity of learning hierarchies are discussed, and then available validation techniques are critically reviewed. The chapter concludes with a suggestion for an alternative validation technique and a statement of the problem and hypothesis.

1. Learning Hierarchy Theory.

Basic to the understanding of learning hierarchies is the concept of positive transfer. Positive transfer implies that learning or doing a new activity is made easier by the learning of a previous activity[1]. For instance, suppose two groups of individuals A and B, are known to be equivalent in their ability to perform a task Y. Before attempting to learn task Y, members of one group, say group A, learn another task - - task X. Positive transfer is said to occur from task X to task Y if members of group A learn task Y more readily than those of group B. That is, positive transfer occurs if members of group A require less time or fewer trials to successfully perform task Y than do members of group B.

Gagné[2] states that the fundamental unit of a learning hierarchy is positive transfer between pairs of tasks. A learning hierarchy is a network of tasks arranged in a manner such that a substantial amount of positive transfer occurs from tasks in a lower position in the network to connected ones in a higher position. Figure 1 shows a typical arrangement of tasks in a learning hierarchy. In this hierarchy,

---

1 R. M. Gagné and E.A. Fleishmah, <u>Psychology and Human Performance</u>, New York, Holt, 1959, P.162.

2 R. M. Gagné, <u>The Conditions of Learning</u>, 2nd ed., New York, Holt, Rinehart and Winston, 1970, p.239.

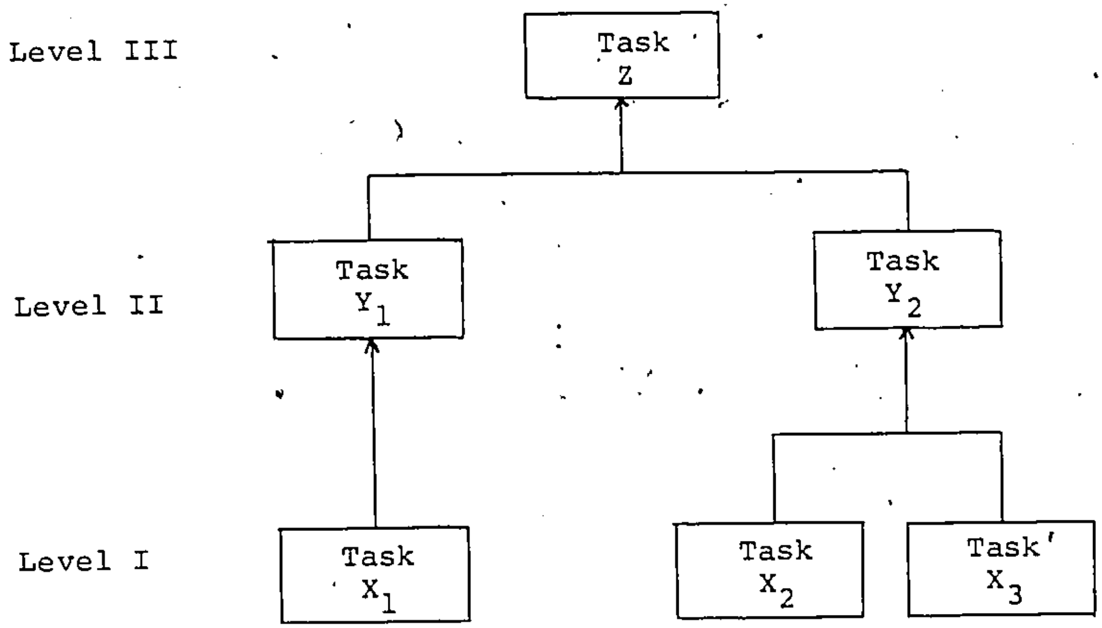

Figure 1 : Typical Arragement of Six Tasks in a Learning Hierarchy

there are six tasks labelled X_1, X_2, X_3, Y_1, Y_2, and Z. Arrows indicate the direction in which positive transfer occurs between tasks. When positive transfer does not occur, the corresponding connection does not appear in the network.

Positive transfer occurs from the tasks at level I in Figure 1 to connected ones at level II. The tasks at level II then transfer positively to the tasks at level III. The task at level III is called the terminal task since all other tasks transfer positively to it; the tasks below the terminal task are subordinate tasks. Thus it is implied that the terminal task is more readily learned if the learner possesses all necessary subordinate tasks, or conversely, failure to learn the terminal task may be due to failure to learn one or more of the essential subordinate tasks. Evidence to support this idea comes from a variety of sources which are comprehensively documented by Walbesser and Eisenberg[3].

Learning hierarchies are characterized by three important features. The first concerns the nature of the tasks in a hierarchy. Gagné[4] describes the tasks as acquiring intellectual

---

3 H. H. Walbesser and T. A. Eisenberg, A Review of Research on Behavioral Objectives and Learning Hierarchies, S.M.E.A.C. Report, Columbus, Ohio, ERIC Information Analysis Center, 1972.

4 R. M. Gagné, "Learning and Instructional Sequence", in F. N. Kerlinger, Ed., Review of Research in Education, Vol.1, Itasca, Ill., Peacock, 1973, p.21.

skills, each of which is stated in behavioral terms. An example of a task is "given two single digit integers and the instruction to add, the learner solves to find the sum". Equally important as what the tasks are is what they are not. The tasks are not descriptions of verbal information, cognitive strategies, motivational factors, or performance sets. While Gagné acknowledges that the latter variables contribute to the learning of the intellectual skills in a hierarchy, his intention is not to describe them specifically in a learning hierarchy. Thus when the elements of a hierarchy are being derived, emphasis must be placed upon what the individual is <u>able to do</u>, and statements about what the individual knows must be avoided[5].

Another feature of learning hierarchies is the role they play in the act of learning. Complex learning, such as school learning, is composed of three distinct classes of stimuli, each of which must be present at the time of learning otherwise some reduction of learning ease or efficiency may be expected to occur[6]. They are (1) integral stimuli - -

---

5 R. M. Gagné, "Learning Hierarchies", <u>Educational Psychologist</u>, 1968, 6, 1-9

6 R. M. Gagné, "Learning and Instructional Sequence", in F. N. Kerlinger, Ed., <u>Review of Research in Education</u>, Vol. 1, Itasca, Ill., Peacock, 1973, p.21.

stimuli that enter directly into learning such as the printed text; (2) activating stimuli -- stimuli that make active previously acquired states or processes such as the teacher commanding "Remember that..."; and (3) stimuli from recall -- stimuli that arise internally from recalled capabilities. Each of these classes of stimuli is in turn made up of various sub-classes of stimuli. Of particular importance in the discussion of learning hierarchies are the subclasses of stimuli from recall. The subclasses of stimuli from recall are stimuli provided by recall of intellectual skills, recall of information, and recall of cognitive strategies. When the learning objective is the acquisition of a new intellectual skill, the stimulus provided by the recall of intellectual skills is a crucial component of the total stimulus situation for efficient learning. The theoretical position of Gagné[7] is that at the time of learning, the learner must recall a hierarchy of intellectual skills which mediate positive transfer to the new intellectual skill. Thus the role of learning hierarchies in the act of learning is that they constitute an important internal component

---

7 <u>Idem, ibid.</u>, p.22.

of the total stimulus situation.

The third feature of learning hierarchies has to do with their uniqueness. Gagné[8] argues that learning hierarchies do not necessarily represent a unique or most efficient route to acquiring a new skill for any given learner. Indeed, some learners may possess all the skills in the hierarchy before instruction; others may be able to omit a subordinate skill and learn a higher order skill through a problem-solving approach; while still others may possibly combine several atypical subordinate skills which yield an unexpected source of positive transfer. What a learning hierarchy represents then is "the most probable expectation of greatest positive transfer for an entire sample of learners concerning whom we know nothing more than what specifically relevant skills they start with"[9].

Thus learning hierarchy theory provides a model which helps to explain the acquisition of intellectual skills. The practical uses of the model are next discussed.

---

8 R. M. Gagné, "Learning Hierarchies", Educational Psychologist, 1968, 6, 1-9.

9 Idem, ibid.

2. Uses of Learning Hierarchies.

The nature of learning hierarchies suggests that they would be of great use in designing instructional sequences, either for single lessons or for entire curricula. Heimer[10] states that even though learning hierarchies are widely used in the development of presentation sequences, their role is not yet clear. Theoretically a learning hierarchy is not intended to be a description of an entire instructional sequence. Rather it should indicate to the curriculum designer that among other stimuli, a certain set of prerequisite intellectual skills must be available to the learner for recall at the time new learning occurs. Learning hierarchies can, however, provide a rough guide for sequencing instruction in the sense that they indicate the intellectual skills that need to be successively mastered by the learner. Support for this statement comes from a number of sources. Briggs[11], for example, concludes after completing a most comprehensive review of research on instructional sequences, that arranging instruction according to the hierarchical nature of the topic is a powerful

---

10 R. T. Heimer, "Conditions of Learning in Mathematics", Review of Educational Research, 1969, 39, 493-508.

11 L. J. Briggs, Sequencing Instruction in Relation to Hierarchies of Competence, Pittsburg, Pa., American Institute for Research, 1969, p. 114.

influence on the outcome of an instructional unit. More recently, Brown[12] compared the performance of groups of tenth- and eleventh- grade students after completing a programmed lesson in mathematics. The control group followed the lesson in a scrambled order and the experimental group followed the same lesson in hierarchical sequence. Brown found that the experimental group required less time to complete the lesson, and made fewer errors during the lesson and on the post-test.

Another use of learning hierarchies is outlined by Glaser and Resnick[13]. They suggest that learning hierarchies can provide a basis for tailored testing; that is, they can be used to design tests which locate an individual's weaknesses in a particular skill area. This is accomplished first by deriving a hierarchy for the skill area of interest and estimating the position of an individual on the hierarchy. Then if each sequential skill in the hierarchy defines a test exercise, a decision tree results for determining the next test exercise for an individual. If an individual "fails" a test exercise, he should be tested at a subordinate level to determine whether

---

12 J. L. Brown, "Effects of Logical and Scrambled Sequences in Mathematical Materials on Learning Programmed Instruction Materials", Journal of Educational Psychology, 1970, 61, 41-45.

13 R. Glaser and L. Resnick, "Instructional Psychology", Annual Review of Psychology, 1972, 23, 207-276.

he needs more instruction for that skill or is lacking subordinate skills. On the other hand, if an individual "passes" a test exercise he should be tested at the next superordinate level. The advantages of this method of testing are that it is efficient timewise, boredom and frustration resulting from testing individuals at inappropriate levels is reduced, and it is particularly suitable for computer based testing.

Learning hierarchies have also been used in mastery learning[14] -- learning where a high degree of proficiency is required in certain fundamental skills necessary for later learning; and they have been used for developing sequential achievement tests[15] -- tests which are intended to measure achievement of an ordered series of skills.

Common to the above mentioned uses of learning hierarchies as well as any other uses is that a valid hierarchy for the skill area under consideration must be derived. The derivation of learning hierarchies is considered in the next section.

---

14 B. S. Bloom, J. J. Hastings, and G. G. Madaus, Handbook of Formative and Summative Evaluation of Student Learning, New York, McGraw-Hill, 1971, p.27.

15 R. C. Cox and G. T. Graham, "The Development of a Sequentially Scaled Achievement Test", Journal of Educational Measurement, 1966, 3, 147-150.

3. Derivation of Learning Hierarchies.

The usual method of deriving learning hierarchies is to hypothesize them on an *a priori* basis through the process known as task analysis, and then to subject them to empirical validiation. Most researchers have adopted this procedure as the most suitable even though it has been criticized for relying heavily on the analyst's knowledge of the learners and his ability to specify prerequisite skills accurately[16]. Other procedures, such as the one Bart and Airasian[17] suggest, require an *a posteriori* examination of the data to identify prerequisite relationships among skills. Nevertheless, the *a posteriori* procedures necessitate a prior analysis of the skill area in order to determine what data to collect. Thus, the difference between the two methods of deriving learning hierarchies is essentially a matter of emphasis on the prior logical analysis of the skill area.

The derivation of a learning hierarchy by task analysis is illustrated below because the *a priori* method is the most widely used.

---

16 G. M. Huntley, "Skill Hierarchies and Their Implications for Testing and Teaching", Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1974, p.3.

17 P. W. Airasian and W. M. Bart, "The Uses of Ordering Theory in the Identification and Specification of Instructional Sequences", Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1974, p. 1-14.

Task analysis, according to Gagné, requires starting with the desired terminal skill and asking the question, "What must the learner have to be able to do in order that he can successfully attain performance of the skill provided he is given only instructions?"[18]. The question is then applied successively to the subordinate skills identified by the answer until the very basic skills required by the problem are reached. Figure 2 illustrates a learning hierarchy for a pre-kindergarten topic in mathematics derived from task analysis by Resnick[19]. In this hierarchy the task to be learned by the child is the ability to demonstrate "halves" of a set of objects as two equivalent subsets, and "thirds" as three equivalent subsets. In order for the child to successfully perform this task, Resnick suggests that he must be able to identify two equivalent subsets as "halves" of a total, and three equivalent subsets as "thirds" of a total; also the child must be able to construct equivalent subsets by distributing objects, one to each of two or three subsets. These tasks are further analyzed until the basic prerequisites of identifying the number of subsets of a group of objects as 2 or 3, and identifying one object of a set are obtained.

---

18 R. M. Gagné, op. cit., p.l.

19 L. B. Resnick, "Design of an Early Learning Curriculum", Working Paper 16, Pittsburg, University of Pittsburg Learning and Research Center, 1967.

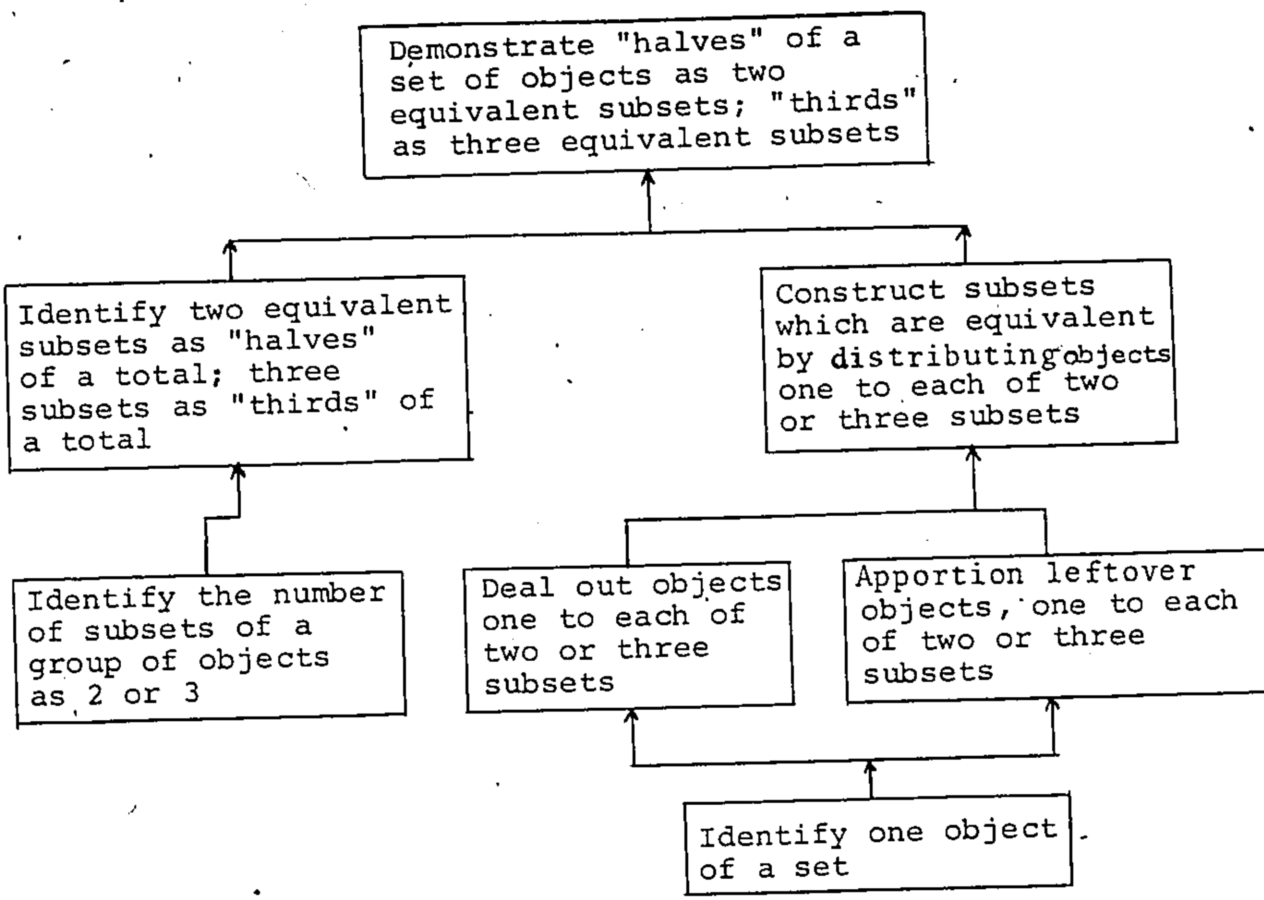

Figure 2: An Example of Pre-Kindergarten Mathematics Learning Hierarchy Derived by Resnick (1967)

It is important to note that in task analysis the question one asks in order to identify prerequisite skills relates to what the individual can do -- not to what the individual knows. Clearly, in Resnick's example, the tasks are concerned with what the child can do as he is asked "to deal out","to apportion", and "to construct".

4. Validation Methodology.

Once a learning hierarchy has been derived, there is no guarantee that it is valid. That is, even though a particular ordering of skills is obtained from the above concepts, one does not know whether the ordering represents a learning hierarchy as some skills may have been overlooked, while others may be superfluous. Thus empirical evidence must be obtained to support or validate the hypothesized hierarchy.

Anderson[20] states that the critical test of the validity of a learning hierarchy must rest upon the demonstration of positive transfer from subordinate to superordinate skills in the hypothesized hierarchy. To this end, two general methods of validating learning hierarchies have evolved. White and

---

20 R. C. Anderson, "Educational Psychology", Annual Review of Psychology, 1967, 18, 129-164.

Gagné [21] refer to them as the "positive" and "negative" methods.

The positive method is based upon the classical transfer model. The method requires examination of each pair of adjacent skills in the hypothesized hierarchy and comparison of the performance of subjects with and without instruction on the subordinate skill(s). If the performance of subjects who received instruction on the subordinate skill(s) is superior on a criterion variable to those who did not, the hierarchy is said to be valid; otherwise it is not valid.

The main drawback to using the positive method of validation is the large number of groups and subjects required to validate even the simplest hierarchy. Uprichard[22] adapted this method in an investigation with pre-school children on the sequence of learning the concepts equivalence, greater than, and less than. Since Uprichard found no obvious hierarchy to hypothesize, he taught the concepts in the 3! or 6 possible sequences to as many different groups in order to determine the optimum sequence for learning these concepts. However, if a hierarchy can be hypothesized, fewer groups are

---

21 R. T. White and R. M. Gagné, "Past and Future Research on Learning Hierarchies", Educational Psychologist, 1974, 11, 19-28.

22 A. E. Uprichard, "The Effect of Sequence on the Learning of Three Set Relations: An Experiment with Pre-Schoolers", Arithmetic Teacher, 1970, 17, 597-604.

required. White[23] shows in this case that 2(k-1) experimental groups are required to validate simple hierarchies of k elements. For instance, suppose the simple hierarchy shown in Figure 3 is to be validated by the positive method. Since this hierarchy has three elements, four experimental groups are necessary. These groups are shown in Figure 4. At each connection in the hierarchy two experimental groups are required-one to be taught only the superordinate skill, and the other to be taught the two skills in the hypothesized order. For complex hierarchies, that is hierarchies with a number of branches and elements, even more than 2(k-1) experimental groups may be required to validate hypothesized hierarchies by the positive method. White calculated that seventy-four experimental groups would have been necessary to validate a twenty element hierarchy hypothesized by Gagné in one of his earlier studies.

In addition to the disadvantage of requiring a large number of groups, the positive method of validation is generally very time consuming and costly. The result is that the positive method is rather cumbersome; hence it is seldom used since most investigators are concerned with hierarchies of more than just a few skills.

---

23 R. T. White, "Research into Learning Hierarchies", Review of Educational Research, 1973, 43, 361-374.

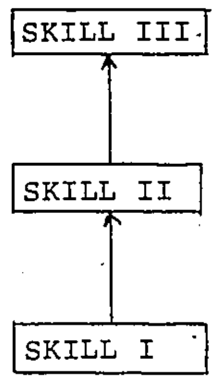

Figure 3: A Simple Hierarchy of Three Skills

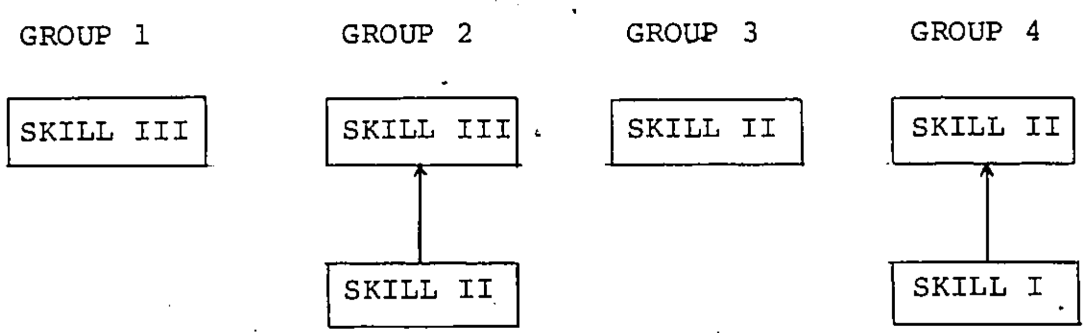

Figure 4: Experimental Groups Required to Validate a Three-Skilled Hierarchy by the Positive Method

The negative method of validation is based upon the premise that learners cannot acquire a particular skill unless they possess all of its relevant subordinate skills. Unlike the positive method, the negative one requires only one group of subjects to test the validity of a hierarchy. White[24] suggests the steps to be followed with this method. The first step of validation is to teach and test the possession of each skill in the hypothesized hierarchy, starting from the bottom and working upwards. The test data collected for each pair of adjacent skills are then analyzed for skill dependency with any one of a number of validation techniques which are reviewed later in this chapter. If a pair of skills is found to be independent, the connection is removed. Thus the method only removes invalid connections and does not uncover new ones. Examples of this method are common in the literature as it is a relatively straightforward and less costly way to validate a hierarchy with many skills.

Pedagogically, learning hierarchies validated by the negative method do not appear to be quite as useful as

---

24 R. T. White, "A Model for the Validation of Learning Hierarchies", Journal of Research in Science Teaching, 1974, 11, 1-3.

positively validated hierarchies. Glaser and Resnick[25] point out that negatively validated hierarchies do not establish sufficient conditions for positive transfer while positively validated ones do. That is, the negative method cannot be used to determine all of the skills which must be acquired before a specific new skill can be learned. Negatively validated hierarchies do, however, provide necessary conditions for positive transfer as they indicate a minimum set of skills which the learner must possess to learn the next skill. Furthermore, negatively validated hierarchies can be used to suggest to the researcher which skills must be included in studies using the positive method. The exact nature of the relationship between negatively and positively validated hierarchies is not clear at present. White and Gagné[26] call for a continuation of studies using both methods as well as for research to establish a correspondence between the two methods.

The general focus of this study is on the negative method of validation as the researcher is able to obtain much

---

25 R. Glaser and L. B. Resnick, <u>op. cit.</u>, pp.207-276.

26 R. T. White and R. M. Gagné, <u>op. cit.</u>

more information about the validity of a hierarchy per experimental group with it than with the positive method. For this reason the negative method will likely continue to be the most frequently used method; hence information about it is of practical value in validating learning hierarchies. More specifically, skill dependency validation techniques used with the negative method are investigated in this study as they are the most essential part of the method.

In order to examine these validation techniques it is first necessary to discuss the nature of the relationship between two hierarchical skills so that a common frame of reference may be obtained.

5. Nature of the Relationship Between Hierarchical Skills.

To clarify the discussion of the nature of the relationship between two hierarchical skills and subsequently to discuss validation techniques, a uniform system of notation is first adopted.

Let a hierarchical connection be defined as the relationship between two intellectual skills, skill I and skill II, such that possession of skill I is hypothesized to be prerequisite to possession of skill II. For any given population four groups may then be distinguished:

1. those who possess only skill I,
2. those who possess only skill II,
3. those who possess both skill I and skill II, and
4. those who possess neither skill.

The statistics f_{10}, f_{01}, f_{11}, and f_{00} are defined to be the number of members of a sample of size N from this population observed in groups (1) to (4) respectively, where the first subscript indicates possession or non-possession of skill I and the second subscript indicates possession or non-possession of skill II. In Figure 5 a summary of the different groups and the notation for the number of members of the sample in them is presented.

A hierarchical connection is defined to be valid if the proportion of the population possessing skill II without also possessing skill I is equal to zero. And conversely, a hierarchical connection is defined to be not valid if the proportion of the population possessing skill II without also possessing skill I is greater than zero. Hence in terms of the notation in Figure 5, a hierarchical connection is valid if the expected frequency in the 01 cell is zero for the population under consideration; otherwise, if the expected frequency in the 01 cell is greater than zero, the connection is not valid.

		SKILL II		
		Do not possess (0)	Possess (1)	
SKILL I	Possess (1)	f_{10}	f_{11}	$(f_{10} + f_{11})$
	Do not possess (0)	f_{00}	f_{01}	$(f_{00} + f_{01})$
		$(f_{10} + f_{00})$	$(f_{11} + f_{01})$	

Figure 5: Notation Used to Describe the Number of Members of the Sample Observed in Each of the Four Possible Groups Involved in a Hierarchical Connection.

When an attempt is made to demonstrate that a hierarchical connection is valid by testing a sample of learners to determine their possession of skills I and II and observing the value of f_{O1}, problems of measurement error and sampling error arise. Because of these errors, the observed distribution of cell frequencies in Figure 5 is only an approximation of the true distribution. Therefore, even though the observed f_{O1} may be zero the hierarchical connection is not necessarily valid, and similarly, if the observed f_{O1} is found to be greater than zero the connection is not necessarily invalid. The problem of sampling error is discussed when particular validation techniques are described later in this chapter. As far as measurement error is concerned, two kinds may exist: an individual may answer a skill-testing question correctly without truly possessing the skill, or an individual may answer a question incorrectly while truly possessing the skill. The first kind of measurement error will be referred to as "guessing", the second as "blundering". The effects of guessing and blundering on f_{O1} and indeed on all cell frequencies in Figure 5 are very complex and cannot be readily predicted. There are at least three reasons why this is so. First, both guessing and blundering are likely to occur simultaneously with different probabilities of occurence. Second, the probability of committing these errors is not necessarily the same for both skill I and skill II. Third, the effect of either of the errors on the cell frequencies of Figure 5 depends upon the proportion of the sample size each cell frequency comprises. That is, if f_{ij}/N is close to unity, the effect of measurement error on f_{ij} is different than it would be if f_{ij}/N was close to 0.5.

If a hierarchical connection is valid and there is no measurement error, skill I and skill II are maximally correlated for any given set of marginal frequencies in Figure 5. This is demonstrated by comparison of the expression for the phi correlation coefficient to the expression for the maximum possible value of phi when f_{01} is equal to zero. The phi correlation coefficient[27] for skill I and skill II is estimated by

$$\phi = \frac{(f_{00}f_{11} - f_{10}f_{01})}{\sqrt{(f_{10} + f_{00})\ (f_{11} + f_{01})\ (f_{00} + f_{01})\ (f_{10} + f_{11})}} \qquad (1)$$

Since f_{01} is equal to zero when the skills are hierarchical

$$\phi = \frac{f_{00}f_{11}}{\sqrt{(f_{10} + f_{00})\ f_{11}f_{00}\ (f_{10} + f_{11})}} \qquad (2)$$

27 J. P. Guilford and B. Fruchter, Fundamental Statistics in Psychology and Education, 5th. Ed., New York, McGraw-Hill, 1973, p. 306.

The maximum possible correlation[28] for any given set of marginal frequencies is

$$\phi_{max} = \sqrt{\frac{(f_{00} + f_{01})\ (f_{11} + f_{01})}{(f_{10} + f_{11})\ (f_{10} + f_{00})}} \quad (3)$$

When f_{01} is equal to zero,

$$\phi_{max} = \sqrt{\frac{f_{00}f_{11}}{(f_{10} + f_{11})\ (f_{10} + f_{00})}} \quad (4)$$

which may be shown to be algebraically equivalent to (2). Thus a valid connection implies that the two skills are maximally correlated. Maximum correlation, however, does not necessarily imply that a hierarchical connection is valid. This can be seen from (3) because maximum correlation for a given set of marginal totals is dependent upon only the marginal totals, hence maximum correlation may occur when any one of the four cell frequencies f_{10}, f_{00}, f_{01}, or f_{11} are zero.

On the other hand, when there is measurement error and the connection is valid, there is no certainty that the two skills are maximally correlated for a given set of marginal totals as the correlation coefficient may be suppressed due to guessing in the 01 cell of Figure 5.

---

28 Idem, ibid., p. 309.

Furthermore when there is measurement error, maximum correlation does not necessarily imply that the connection is valid for the same reason as stated above in the case without measurement error.

The relationship between the difficulties of skill I and skill II in a hierarchical connection is also of interest. Skill difficulty (d_i) may be defined as the proportion of individuals of the population possessing a skill. Hence by definition, the difficulty of skill I is

$$d_I = \frac{f_{10} + f_{11}}{N} \tag{5}$$

and the difficulty of skill II is

$$d_{II} = \frac{f_{11} + f_{01}}{N} \tag{6}$$

where $N = f_{00} + f_{11} + f_{01} + f_{10}$.

If a hierarchical connection is valid and there is no measurement error, skill I must necessarily be as easy as or easier than skill II (i.e. $d_I \geq d_{II}$). This can be proven quite readily because when the connection is valid, f_{01} is equal to zero by definition; thus

$$d_{II} = \frac{f_{11}}{N} \qquad (7)$$

By substituting (7) into (5),

$$d_I = d_{II} + \frac{f_{10}}{N} \qquad (8)$$

since $f_{10} \geqq 0$,

$$d_I \geqq d_{II} \qquad (9)$$

However, skill I being as easy as or easier than skill II does not necessarily imply that the connection is valid. For instance, if

$$d_I \geqq d_{II} \qquad (10)$$

then by substitution,

$$(f_{10} + f_{11}) \geqq (f_{11} + f_{01}) \qquad (11)$$

hence,

$$f_{10} \geqq f_{01} \qquad (12)$$

Thus when skill I is easier than skill II, the only condition that needs to be satisfied is that f_{10} be greater or equal to f_{01}, and there is no requirement that f_{01} be equal to zero.

In the case where a hierarchical connection is valid and there is measurement error, skill I does not necessarily have to be observed to be as easy as or easier than skill II. This is apparent because individuals could blunder on a question testing possession of skill I, which would then make the skill appear more difficult than it actually is. Nor does skill I being as easy as or easier than skill II imply that the connection is valid for the same reason as the case when there is no measurement error. Therefore, as with correlation, skill difficulty is a necessary condition for the validity of a connection only when there is no measurement error.

There is an obvious relationship between skill difficulty and skill intercorrelation. So far it has been shown that when there is no measurement error and a connection is valid (1) skill I and skill II are maximally correlated for given marginal totals, and (2) skill I is as easy as or easier than skill II. Since skill difficulty is derived from the marginal totals, the correlation between skill I and skill II is a function of the difficulty of the skills. An inspection of the formula for ϕ_{max} given in (3) together with the definitions of skill difficulty given in (5) and (6) shows that ϕ_{max} is equal to unity when d_I is equal to d_{II}. However, when d_I is greater than d_{II}, ϕ_{max} is always less than unity. Therefore, one can conclude more precisely that when a connection is valid, skill I and II are maximally correlated, and the degree of correlation is dependent upon how close in difficulty the two skills are.

In summary, a uniform system of notation for describing skills in a hierarchical connection has been defined. Following that, the effects of measurement error, skill intercorrelation, and skill difficulty on a pair of hierarchically related skills were discussed. Next, the techniques available for testing the validity of learning hierarchies are reviewed.

6. Validation Techniques.

A number of techniques are proposed in the literature for testing the validity of learning hierarchies. These techniques can be classified as probabilistic or deterministic according to the theoretical model upon which they are based. Furthermore, the techniques can be subdivided into those used for testing the validity of each connection in a hierarchy, and those used for testing the validity of linear sequences of connections.

Regardless of the technique used, the researcher faces the possibility of committing two types of error when testing the validity of a hierarchy. He may conclude that the hypothesized hierarchy is valid when it actually is not valid, or he may conclude that the hierarchy is not valid when it actually is valid. In this study, the former type of error is referred to as type V-I error (concluding that hierarchy is _valid_ when it actually is _invalid_). The latter is referred to a type I-V error (concluding that a hierarchy is _invalid_ when it actually is _valid_). These errors are summarized in Figure 6. From the point of view of "scientific conservativeness", type V-I error is more serious than type I-V error. Scientific conservativeness implies that it is preferable to make no claim to knowledge than to make a false claim. Thus to make a mistake in claiming that a hierarchy is valid is more serious than to make a mistake in claiming that a hierarchy is not valid.

Now, the various validation techniques available are defined, illustrated with an example, critically reviewed, and their overall merits assessed.

a) Proportion of Positive Transfer (PPT). - The "proportion of positive transfer" index was originally proposed by Gagné and Paradise[29] as a deterministic measure

---

29 R. M. Gagné and N. E. Paradise, "Abilities and Learning Sets in Knowledge Acquisition", _Psychological Monographs_, 1961, 75, Whole Number 518.

Indicated State of Connection	True State of Connection: Valid	True State of Connection: Invalid
Invalid	Error (type I-V)	Correct Decision
Valid	Correct Decision	Error (type V-I)

Figure 6 : Types of Errors Possible when the Validity of a Connection in a Hierarchy is Tested.

of the validity of a hierarchical connection. They define the index as the proportion of the population which supports the hypothesized hierarchy. Symbolically the index can be written as,

$$PPT = \frac{f_{11} + f_{00}}{f_{11} + f_{00} + f_{01}} \tag{13}$$

Implicit in the definition of PPT by Gagné and Paradise is that f_{10} does not yield information to support the hierarchy because subjects may not have received sufficient instruction to learn skill II and hence cannot be included in the ratio.

PPT has a possible range of values of from 0 to 1, with closeness to 1 being indictive of a valid connection. Gagné and Paradise[30] state that the "chance" value of the index is generally in the range of 0.25 to 0.50, however, no reason was offered why this is the case. Presumably what the authors mean by the chance value is the calculated value of PPT when the skills are independent. Furthermore, Gagné and Paradise do not suggest a criterion value for PPT, so one must assume that any value exceeding the chance value would be

---

30 R. M. Gagné and N. E. Paradise, op. cit.

evidence of a valid hierarchy. Most subsequent validation studies which used PPT cited values of 0.90 or greater as the criterion for a valid hierarchy and this has become the commonly accepted criterion[31].

To illustrate the calculation of the ratio suppose the data in Table I are observed in a validation experiment. Thus the value of PPT is found to be,

$$PPT = \frac{45 + 15}{45 + 15 + 5} = 0.92 \qquad (14)$$

Since PPT is greater than the criterion, the hierarchical connection is assumed to be valid.

Phillips[32] investigated the efficacy of PPT. Two tests were administered to a large sample of elementary school children of varying ability. One test had items ordered according to a hypothesized hierarchy, while the other had the same items randomly ordered. The value of PPT was calculated for each pair of adjacent items on both tests, and at almost every connection in both hierarchies values of 0.90 or greater were found. A closer examination of the index suggests some reasons why this may be so.

---

31 H. H. Walbesser and T. A. Eisenberg, <u>op.cit.</u>

32 E. R. Phillips, "Development of Optimal Instructional Sequences", Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1974, p.4.

Table I

Hypothetical Observed Frequencies Used in Illustrative Examples of Deterministic Validation Techniques

SKILL II

SKILL I	Do not possess (0)	Possess (1)	
Possess (1)	$f_{10} = 35$	$f_{11} = 45$	80
Do not Possess (0)	$f_{00} = 15$	$f_{01} = 5$	20
	50	50	100

First, PPT is directly influenced by the difficulty of the skills in a connection. If the skills are easy, the calculated value of PPT will be closer to unity than if the skills are difficult because of the influence of f_{11} on PPT. Thus if the connection is valid there is a lesser probability of committing type I-V error with PPT when the skills are easy than when they are difficult. On the other hand, if a connection is invalid, there is a lesser probability of committing type V-I error when the skills are difficult than when they are easy.

Another reason why PPT may be misleading is that it can take values greater than the criterion even when the skills are independent. White[33] illustrates this characteristic of PPT by comparing for each connection the observed value of PPT in a hierarchy validated by Gagné, to the value of PPT expected when the skills are independent. In almost half of the connections White found the expected value of PPT for independent skills greater than 0.90, and in one case the expected value even exceeded the observed value of PPT yet the connection was nevertheless shown to be valid. White concludes that PPT is little more than a measure of the correlation between two skills. It was shown in the previous section that skill intercorrelation

33 R. T. White, "Indexes Used in Testing the Validity of Learning Hierarchies", Journal of Research in Science Teaching, 1974, 11, p.61-66.

is a necessary but not sufficient indicator of the validity of a connection.

Besides these weaknesses, PPT lacks a known sampling distribution. Thus a subjective judgement has to be made when generalizing from the sample to the population. Furthermore, it does not take into account measurement errors which are bound to occur in any behavioral research.

In general then, PPT appears to be a technique of very little utility for hierarchy validation. The next technique proposed by Walbesser was intended to overcome the faults of PPT.

b) Walbesser Ratios (WR). - Walbesser[34] suggests three ratios for testing the validity of hierarchical connections; the consistency ratio, the adequacy ratio, and the completeness ratio. The ratios range in value from 0 to 1, and all three ratios must exceed an arbitrarily set criterion of 0.85 for a connection to be valid.

The consistency ratio is intended to test the implication that possession of skill II implies possession of skill I. Symbolically, the ratio is defined as

---

34 H. H. Walbesser, An Evaluation Model and its Application, American Association for the Advancement of Science, Washington, D.C. 1968.

$$\frac{f_{11}}{f_{11} + f_{01}} \tag{15}$$

In effect, the consistency ratio compares the number of subjects who possess both skills in the connection, to the total number who possess the superordinate. The ratio will be close to 1 if relatively few subjects are observed to possess only the superordinate skill, and it will be close to zero if a relatively large number of subjects are observed to possess only the superordinate skill.

The adequacy ratio is defined as

$$\frac{f_{11}}{f_{11} + f_{10}} \tag{16}$$

This ratio is intended to test the implication that acquisition of skill I implies acquisition of skill II provided appropriate instruction in skill II is given. This is accomplished by comparing the number of subjects possessing both skills to the total number possessing the subordinate skill. The ratio will be close to unity if relatively few subjects possess the subordinate skill only, whereas it will be close to zero if a relatively large number of subjects possess only the subordinate skill.

The third ratio, the completeness ratio, is intended to compare the number of subjects possessing both skills in a connection relative to that who possess neither or both skills. It is defined as

$$\frac{f_{11}}{f_{11} + f_{00}} \tag{17}$$

The ratio is close to unity if relatively few subjects possess neither of the skills in the connection, whereas it is close to zero if a relatively large number of subjects possess neither skill.

Calculation of the Walbesser ratios can be illustrated from the sample data in Table I, as follows:

$$\text{consistency ratio} = \frac{45}{45 + 5} = 0.90 \tag{18}$$

$$\text{adequacy ratio} = \frac{45}{45 + 35} = 0.56 \tag{19}$$

$$\text{completeness ratio} = \frac{45}{45 + 15} = 0.75 \tag{20}$$

Since two of the three ratios do not exceed the 0.85 criterion, the hypothesis that skill I is subordinate to skill II must be rejected with the sample data.

The above example demonstrates a difficulty likely to be experienced quite often when attempting to validate a hierarchy with ŴR -- the criterion of 0.85 for the ratios is very stringent. It can be shown that f_{10}, f_{01}, and f_{00} all have to be quite small relative to f_{11} in order to validate a connection. For instance, the following condition must be satisfied for the consistency ratio for a valid connection:

$$\frac{f_{11}}{f_{11} + f_{01}} \geqq 0.85 \qquad (21)$$

Therefore,

$$f_{11} \geqq 0.85\ f_{11} + 0.85\ f_{01} \qquad (22)$$

so that

$$f_{11} \geqq 5.67\ f_{01} \qquad (23)$$

Similarly, it can be shown that f_{11} also has to be at least 5.67 times greater in magnitude than each of f_{00} and f_{10} by use of the completeness and adequacy ratios. Hence for a connection to be valid f_{11} must be approximately six times greater in magnitude than each of f_{01}, f_{00}, and f_{10}. This may occur at the bottom of a hierarchy where most individuals would be expected to possess both the skills in a connection. However, as progress is made up the hierarchy, more and more individuals are likely to "drop out" as the skills become increasingly more complex and difficult.

That is, f_{11} is likely to decrease in magnitude relative to the other three frequencies. As a result, the probability of a connection being validated would tend to decrease as the skills become more difficult. In theory at the uppermost levels of a hierarchy it is possible that one or more of f_{00}, f_{10} and f_{01} may be greater than f_{11} and the connection still be valid - a situation in which WR would be useless. The above discussion suggests that with WR the probability of committing type V-I error decreases as the skills become more difficult if the connection is not valid, and if the connection is valid, the probability of type I-V error increases as the skills become more difficult.

At least three possibilities appear to exist for minimizing skill difficulty troubles with WR. One is to validate the hierarchy using a more able group of learners so as to ensure that the criteria of the adequacy and completeness ratios are satisfied. Another possibility is to use WR only in situations where the skills have just been taught. The third possibility is to lower the criteria for the adequacy and completeness ratios. Kane, McDaniel, and Phillips[35] suggest the later possibility to

---

35 R. B. Kane, E. D. McDaniel, and E. R. Phillips, "Analyzing Learning Hierarchies Relative to Transfer Relationships within Arithmetic", U.S.O.E. Project No. 9-E-128, Purdue University, 1971, p. 28.

reduce skill difficulty problems with WR. They recommend that the adequacy ratio criterion be lowered to 0.70, and the criterion of the completeness ratio be lowered to 0.50 and the consistency ratio be left unchanged at 0.85. No reasons were given for selecting these criteria, so presumably they were chosen on the basis of experience with the ratios in validating learning hierarchies. Further research is necessary to determine the usefulness of the Kane, McDaniel, and Phillips suggestion.

Another difficulty with WR is that, like PPT, it is a deterministic technique which does not take into account measurement error. Also, the ratios lack a known sampling distribution; thus a subjective judgement is required when making inferences beyond the sample of learners with whom the hierarchy is validated.

The three troubles with WR - their sensitivity to skill difficulty, their susceptibility to measurement error, and their lack of sampling distribution - suggest that the technique is of little use for testing the validity of hierarchical connections. The technique has been used quite often in validation studies and appears to be regarded as an acceptable measure of the validity of a connection, an anomaly which warrants further investigation.

c) The Difference Ratio (DR). - The difference ratio is a deterministic index developed by Durell[36] to test the validity of hierarchical connections. Durell established four a priori criteria to guide the development of the technique. The criteria were that the technique would (1) make use of all the cells in Table I so as to utilize the maximum amount of information available about the relationship between a pair of skills; (2) rely heavily on f_{11} and f_{01} since they contribute the most important information for acceptance or rejection of the hypothesis of skill dependency respectively; (3) require f_{11}/N to be equal to or greater than 0.50 before the hypothesis of skill dependency can be supported; and (4) cause the hypothesis of skill dependency to be rejected if f_{01}/N is greater than 0.25. The two last criteria were arbitrarily determined.

The technique yields an index ranging in value from +1 to -1, with a value of 0.50 or greater required before a connection can be considered valid. Originally the formula for DR was presented as

$$DR = \frac{f_{11}}{f_{11} + f_{01} + f_{10} + f_{00} + |f_{10} - f_{00}|} - P\{f_{01}\} \qquad (24)$$

---

36 A. B. Durell, "A Comparison of Measures for Validating Learning Hierarchies", Unpublished Doctoral Dissertation, University of Toronto,1973.

where the right hand term represents the proportion of (0,1) responses. For the sake of simplicity the formula can be reduced to

$$DR = \frac{f_{11}}{N + |f_{10} - f_{00}|} - \frac{f_{01}}{N} \qquad (25)$$

since f_{11}, f_{01}, f_{10}, and f_{00} sum to N, the sample size.

Using the data in Table I as an illustration, DR is calculated as follows:

$$DR = \frac{45}{100 + |35 - 15|} - \frac{5}{100} = 0.325 \qquad (26)$$

In this case, the hypothesis of skill dependency is not supported since the calculated value of DR is less than the criterion of 0.50.

Again, like PPT and WR, DR appears to be sensitive to skill difficulty. When f_{11} is large relative to each of f_{00}, f_{10}, and f_{01}, the value of DR tends to be greater than when f_{11} is small relative to each of f_{00}, f_{10}, and f_{01}, regardless of whether or not the connection is valid. The first situation would be one when the skills in a connection are easy, the latter when the skills are difficult. Thus DR would tend to validate connections more readily when the skills in a connection are easy than when they are difficult. As a result if a connection is not valid, the probability of committing type V-I error increases as both of the skills become easier. Whereas, if a connection is valid, the probability of type I-V error increases as both skills become more difficult.

Besides the effects of skill difficulty several other criticisms may be made of DR. The first concerns one of the original criteria used by Durell to develop the ratio. DR was developed so that as much as 25% of the sample can be observed to possess skill II without skill I and the connection still can be considered valid. No justification for choosing this criterion is offered by Durell, and indeed, no justification seems plausible which would allow such a great proportion of the sample to contradict the hypothesis of skill dependency. The proportion of the sample allowed to contradict the hypothesis can, however, be reduced by raising the criterion for the ratio above 0.50. By doing so, the probability of type V-I error, the most serious kind of decision error, would be decreased.

Other criticisms of DR are that, like PPT and WR, it does not have a known sampling distribution and it does not take into account measurement errors.

The above criticisms suggest that theoretically the technique is weak. Not enough empirical research evidence is available to determine whether or not this is also true in practice.

d) Guttman Scalogram Analysis (GSA). - The Guttman[37] scalogram analysis technique was applied to learning hierarchy validation by Resnick and Wang[38]. Essentially GSA is a method of determining to what extent the response patterns of dichotomously scored (0 or 1) test items form a linear ordering. When the method is applied to validating a learning hierarchy, an attempt is made to establish if the test items measuring the skills under consideration form a linear ordering in the manner hypothesized. From the order of the test items it is then inferred that the skills were acquired in that sequence.

Unlike the validation techniques discussed to this point, GSA is typically used for validating linear sequences of more than two skills. Figure 7 illustrates a perfect linear ordering of a hierarchy of five skills for a group of six learners. From this illustration it can be seen that possession of a given skill implies possession of all subordinate skills. However, rarely is such a perfect scaling obtained in experimental situations. The extent to which the data can be arranged in the

---

37 L. Guttman, "A Basis for Scaling Qualitative Data", American Sociological Review, 1944, 9, p. 139-150.

38 L. B. Resnick and M. C. Wang, "Approaches to the Validation of Learning Hierarchies", Proceedings of the Eighteenth Annual Regional Conference on Testing Problems, Princeton, N.J., Educational Testing Service, 1969.

manner shown in Figure 7 is expressed as the coefficient of reproducibility (Rep) which ranges in value from 0 to 1.

The coefficient of reproducibility is given by the formula

$$\text{Rep} = 1 - \frac{\text{total number of errors}}{\text{total number of responses}} \tag{27}$$

where an error is defined as an inversion in the order in which the skills are acquired as compared to the order in which they are hypothesized to be acquired. Figure 8 illustrates a linear ordering with one error - learner number 3 failed the item testing possession of skill II, but passed that testing skill III. The total number of responses possible is obtained by multiplying the number of learners by the number of items. Hence, the maximum number of errors possible is equal to the number of responses. From formula (27) it can be seen that when there are no errors Rep is equal to 1, when the maximum number of errors occurs Rep is equal to 0.

Calculation of Rep can be illustrated with the data in Figure 8 as follows:

$$\text{Rep} = 1 - \frac{1}{30} = 0.967 \tag{28}$$

		Skill				
		I	II	III	IV	V
Learner	1	1	1	1	1	1
	2	1	1	1	1	O
	3	1	1	1	O	O
	4	1	1	O	O	O
	5	1	O	O	O	O
	6	O	O	O	O	O

Figure 7: A Perfect Linear Ordering of Five Skills for a Group of Six Learners. For Each Learner-Skill Combination a 1 Indicates that the Learner Possesses the Skill; a O Indicates that he does not Possess the Skill.

		Skill				
		I	II	III	IV	V
	1	1	1	1	1	1
	2	1	1	1	1	0
Learner	3	1	0	1	0	0
	4	1	1	0	0	0
	5	1	0	0	0	0
	6	0	0	0	0	0

Figure 8: A Linear Ordering of Five Skills for Six Learners with One Error. The Error is Indicated by the Performance of Learner 3 Who Passed the Item Testing Skill III But Not That Testing Skill II.

No sampling distribution has been found for Rep. A Rep of 0.90 or greater is generally considered to be sufficient for a linear ordering of items as this represents 10% or fewer errors. Thus according to the Guttman technique the data in Figure 8 indicate a valid linear hierarchy. Torgenson[39] points out, however, that care must be taken when evaluating Rep as it has a tendency to depend upon the values of the item marginal totals consequently the expected value of Rep is high even when the items are independent.

Several weaknesses with GSA as a hierarchy validation technique are apparent. First, like the other deterministic techniques discussed, Rep is affected by measurement errors. For instance in the hierarchy illustrated in Figure 8, there is no available method of estimating whether learner 3 blundered on the item testing skill II, correctly guessed the answer to the item testing skill III or that it is his true state. The exact effect of measurement errors on Rep is not readily determinable as the matter is complicated by the additional variable of position of the skill-testing question in the hierarchy. Generally speaking, however, both guessing

---

39 W. S. Torgenson, Theory and Methods of Scaling, New York, Wiley, 1958.

and blundering are likely to cause an "error" as defined in formula (27). The effect of an error on Rep is to decrease its calculated value. Thus if the hierarchy is valid, type I-V error would tend to increase as measurement errors increase since Rep would tend to be lower than its true value. For the same reason, if the hierarchy is not valid, type V-I error would tend to increase as measurement errors increase.

A second weakness of Rep is that learners who correctly answer all or none of the skill testing items tend to inflate the calculated value of the coefficient. This is because such learners contribute no additional information about the sequential order of the skills, yet they increase the total number of item responses thus decreasing the ratio of errors to responses in the formula for Rep. This in turn increases the calculated value of Rep over what it would be had such learner responses not been part of the data. For example had learners 1 and 6 not been included in the data in Figure 8, Rep would be 0.95 instead of 0.967 as calculated in (28), since there would be a total of only 20 responses instead of 30.

Another weakness of GSA is that one invalid connection, in an otherwise valid hierarchy, may lead to the rejection of the entire hierarchy. This is because Rep is inversely proportional to the total number of errors made by all subjects, even if the errors are committed on the same connection. As an example, the connection between skill I and II in Figure 7 may be considered. If this connection were invalid, learners 1 to 5 inclusive would have "0" under the skill I column. The result would be five errors; thus

$$\text{Rep} = 1 - \frac{5}{30} = 0.83 \qquad (29)$$

and the entire hierarchy would be considered invalid since Rep is less than the criterion value of 0.90. In this case the connections between the remaining skills are valid, yet the hierarchy was rejected because of an invalid connection between skills I and II. This weakness could be dealt with by eliminating the connection and re-calculating Rep; however, theoretically there would be no grounds for doing so since the purpose of applying Rep to the original data would be to validate a particular postulated hierarchy.

A third weakness with GSA is that it has very limited applicability since the technique can be used only to validate linear hierarchies. According to Airasian and Bart:

> Most prior research in the areas of cognitive development, tests and measurement, and curriculum development has demonstrated that groups of tasks are much more likely to form non-linear, branched hierarchical networks than they are to form linear networks. As a consequence, reproducible Guttman scales exceeding six or seven tasks are rare. While non-linear networks of tasks or behaviors do not manifest the conceptual simplicity of linear networks, it is likely that the former represent a more accurate representation of the prerequisite relationships existing among tasks or behaviors.[40]

While it it possible to apply GSA to linear branches in a non-linear hierarchy, the results would not prove very meaningful, and the single advantage of GSA- that it provides an overall index of the validity of a many-skilled hierarchy - would be lost.

In summary, GSA is a deterministic technique which has been applied to the validation of learning hierarchies. The technique yields a single index, Rep, which gives an indication of the overall validity of a linear hierarchy. Its usefulness, however, is minimal as the technique does not have a known sampling distribution, does not take into account measurement errors, permits one invalid connection in a hierarchy to cause rejection of the entire hierarchy, is inflated by learners who

---

40 P. W. Airasian and W. M. Bart, "Validating a Priori Instructional Hierarchies", Journal of Education Measurement, 1975, 12, p. 166.

correctly answer all or none of the test items, and is applicable only to linear hierarchies.

e) White and Clark Technique (WC). - Unlike the validation techniques discussed thus far, the White and Clark[41] technique is based upon a probabilistic model. WC is used to test the validity of hierarchical relationships hypothesized between pairs of skills. This is accomplished first by use of two questions for testing subjects for possession of each of the skills. The assumption is made that the skills have been so precisely defined that, if it were not for errors of measurement, both skill-testing questions per skill would be answered either correctly or incorrectly by all subjects. Also, for each skill the probabilities of random errors are assumed to be equal, yet independent, for the two questions. An estimate is then obtained of the number of subjects expected at a particular confidence level to answer two skill II questions correctly and no skill I questions correctly. In order to do this the assumption is made that the skills are hierarchical. The hierarchical connection is concluded to be valid if the observed value of

---

41 R. T. White and R. M. Clark, "A Test of Inclusion Which Allows for Errors of Measurement", Psychometrika, 1973,38, 77-86.

f_{O2} is less than or equal to the upper limit of the confidence interval of the expected value of f_{O2}; otherwise, the connection is concluded to be not valid.

Before illustrating the details of the technique, the probabilistic model used by WC is described.

The model is given in Table II. White and Clark derive it by dividing the population into four mutually exclusive and exhaustive groups - subjects who possess only skill I, subjects who possess only skill II, subjects who possess neither skill, and subjects who possess both skills. The proportion of the population belonging in each of these groups is defined to be P_I, P_{II}, P_O, and P_B respectively. Next, the probability of an individual answering a skill I question correctly who truly possess the skill is defined as θ_a (i.e. the probability of not blundering on a skill I question), and who does not possess the skill as θ_b (i.e. the probability of guessing the answer correctly to a skill I question). The corresponding probabilities for skill II are defined as θ_c and θ_d. Then the probabilities, P_{ij} where i, j = 0, 1, 2, of a member of a randomly selected sample being classified into one of the nine cells in Table II are found in terms of the proportions P and the probabilities θ. These are obtained by summing over the four possible groups, the probability of an individual being a member of each group multiplied by the conditional probability of members of the

TABLE II

Probabilities (p_{ij}) of Answering Correctly the Various Possible Combinations of Numbers of Questions Testing Skills I and II Respectively (White and Clark Model)

Number of Skill I Test Questions Answered Correctly	Number of Skill II Test Questions Answered Correctly: 0	1	2
2	$p_{20} =$ $P_0\theta_b^2(1-\theta_d)^2$ $+ P_I\theta_a^2(1-\theta_d)^2$ $+ P_{II}\theta_b^2(1-\theta_c)^2$ $+ P_B\theta_a^2(1-\theta_c)^2$	$p_{21} =$ $2P_0\theta_b^2\theta_d(1-\theta_d)$ $+2P_I\theta_a^2\theta_d(1-\theta_d)$ $+2P_{II}\theta_b^2\theta_c(1-\theta_c)$ $+2P_B\theta_a^2\theta_c(1-\theta_c)$	$p_{22} =$ $P_0\theta_b^2\theta_d^2$ $+ P_I\theta_a^2\theta_d^2$ $+ P_{II}\theta_b^2\theta_c^2$ $+ P_B\theta_a^2\theta_c^2$
1	$p_{10} =$ $2P_0\theta_b(1-\theta_b)(1-\theta_d)^2$ $+2P_I\theta_a(1-\theta_a)(1-\theta_d)^2$ $+2P_{II}\theta_b(1-\theta_b)(1-\theta_c)^2$ $+2P_B\theta_a(1-\theta_a)(1-\theta_c)^2$	$p_{11} =$ $4P_0\theta_b(1-\theta_b)\theta_d(1-\theta_d)$ $+4P_I\theta_a(1-\theta_a)\theta_d(1-\theta_d)$ $+4P_{II}\theta_b(1-\theta_b)\theta_c(1-\theta_c)$ $+4P_B\theta_a(1-\theta_a)\theta_c(1-\theta_c)$	$p_{12} =$ $2P_0\theta_b(1-\theta_b)\theta_d^2$ $+2P_I\theta_a(1-\theta_a)\theta_d^2$ $+2P_{II}\theta_b(1-\theta_b)\theta_c^2$ $+2P_B\theta_a(1-\theta_a)\theta_c^2$
0	$p_{00} =$ $P_0(1-\theta_b)^2(1-\theta_d)^2$ $+ P_I(1-\theta_a)^2(1-\theta_d)^2$ $+ P_{II}(1-\theta_b)^2(1-\theta_c)^2$ $+ P_B(1-\theta_a)^2(1-\theta_c)^2$	$p_{01} =$ $2P_0(1-\theta_b)^2\theta_d(1-\theta_d)$ $+2P_I(1-\theta_a)^2\theta_d(1-\theta_d)$ $+2P_{II}(1-\theta_b)^2\theta_c(1-\theta_c)$ $+2P_B(1-\theta_a)^2\theta_c(1-\theta_c)$	$p_{02} =$ $P_0(1-\theta_b)^2\theta_d^2$ $+ P_I(1-\theta_a)^2\theta_d^2$ $+ P_{II}(1-\theta_b)^2\theta_c^2$ $+ P_B(1-\theta_a)^2\theta_c^2$

group answering the relevant number of questions correctly.

WC may be illustrated with the sample data in Table III. The statistical null hypothesis tested by WC is that the proportion of the population possessing only skill II is equal to zero; the alternative hypothesis that the same proportion is greater than zero. Symbolically, the null and alternative hypotheses are respectively,

$$H_O : P_{II} = O$$

and

$$H_A : P_{II} > O$$

In order to estimate p_{O2} and subsequently to test the null hypothesis, estimates of the parameters in the model must be obtained under the assumption that P_{II} is equal to zero. The parameters are estimated from the marginal frequencies a, b, c, d, e, and f which are respectively the number of subjects answering 2, 1, and 0 skill I questions correctly and 2, 1, and 0 skill II questions correctly. To arrive at estimates from the marginal frequencies White and Clark found it necessary to assume that $\theta_b = 0$ and $\theta_c = 1$. This is equivalent to saying that there is no chance of an individual guessing the answer to a skill I question correctly, and no chance of an individual blundering on a skill II question.

Table III

Hypothetical Observed Frequencies Used in Illustrative Example of the White and Clark Technique[1]

Number of Skill I Test Questions Answered Correctly	Number of Skill II Test Questions Answered Correctly: 0	1	2	
2	$f_{20} = 31$	$f_{21} = 19$	$f_{22} = 95$	a = 145
1	$f_{10} = 4$	$f_{11} = 0$	$f_{12} = 4$	b = 8
0	$f_{00} = 11$	$f_{01} = 3$	$f_{02} = 1$	c = 15
	f = 46	e = 22	d =100	N = 168

[1]Marginal frequencies for 2, 1 and 0 questions answered correctly testing skills I and II are defined as a, b, c, d, e, and f respectively; the sample size is defined by N.

When the appropriate marginal frequencies in Table III are substituted into the White and Clark formulae the P's and θ's are estimated to be:

$$\hat{\theta}_a = \frac{2a}{2a + b} = \frac{2 \times 145}{2 \times 145 + 8} = 0.9731 \qquad (30)$$

$$\hat{\theta}_d = \frac{e}{e + 2f} = \frac{22}{22 + 2 \times 46} = 0.1929 \qquad (31)$$

$$\hat{P}_B = 1 - \frac{(e + 2f)^2}{4f\ N} = 1 - \frac{(22 + 2 \times 46)^2}{4 \times 46 \times 168} = 0.5796 \qquad (32)$$

$$\hat{P}_I = \frac{(2a + b)^2}{4aN} - \hat{P}_B = \frac{(2 \times 145 + 8)^2}{4 \times 145 \times 168} - 0.5796 = 0.3318 \qquad (33)$$

$$\hat{P}_O = 1 - (\hat{P}_B + \hat{P}_I) = 1 - (0.5796 + 0.3318) = 0.0886 \qquad (34)$$

The above estimates are then substituted into the expression for p_{O2} in the model in Table II, and $\hat{p}_{O2}$ is found to be 0.0023. It should be noted that p_{O2} is estimated under the assumption that the null hypothesis is true, therefore $\hat{P}_{II}$ is assigned the value of zero. According to White and Clark, f_{O2} - the observed frequency in the O2 cell of Table II - is a binomial random variable with parameters (N, p_{O2}). Using the binomial expansion of $(N = 168, \hat{p}_{O2} = 0.0023)$, the cumulative probability of $f_{O2} = 0, 1,$ or 2 is greater than 0.95. Thus f_{O2} would not be expected to be greater than 2 at the 95% confidence level under the null hypothesis. However in Table III, f_{O2} was observed to be 1 so the null hypothesis must not be rejected, and it must be concluded that it is reasonable to assume that skill I is prerequisite to skill II.

While WC appears to be a more reliable validation technique than the preceding ones since it partially takes into account measurement errors and provides a sampling distribution, several weaknesses in the technique are evident. The first concerns the parameter estimation procedure used by WC. The White and Clark model is based upon seven independent parameters. Under the null hypothesis P_{II} is equal to zero, therefore six parameters must be estimated. White and Clark chose to estimate the parameters by marginal frequencies, and since only four of these frequencies are independent it was necessary to make assumptions about the values of two parameters. As mentioned previously θ_b and θ_c were assumed to be zero and unity respectively. An examination of the expression for p_{O2} in Table II shows that the result of this assumption is that p_{O2} tends to be over-estimated when the population values of θ_b and θ_c are not zero and unity. Since the null hypothesis is tested by determining if the observed frequency in the O2 cell of Table II is greater than that expected by chance, over-estimating p_{O2} makes the null hypothesis less likely to be rejected. The result of the above assumptions therefore is to increase the probability of committing type V-I error. This weakness, however, could be overcome by obtaining parameter estimates from the cell frequencies in Table II. Since there are eight independent cell frequencies all of the parameters could be estimated from them, and the need to make assumptions about the values of two of the parameters would be eliminated.

Another weakness of WC is that large samples are required in order to reduce the probability of type V-I and type I-V errors to acceptable levels. Typically, White and Clark[42] found that with 10% measurement error, that is $\theta_a = 0.90$, $\theta_b = 0.10$, $\theta_c = 0.90$, and $\theta_d = 0.10$, and with $P_0 = 0.45$, $P_I = 0.175$, $P_{II} = 0.025$, and $P_B = 0.35$, a sample of size 500 is required to make the probability of type I-V error 5% and the probability of type V-I error 6%. Decision errors such as these would be acceptable to most researchers. However, as sample size decreases, the risk of decision error increases to unacceptable levels with this technique. Under the same conditions as above but with a sample size of 100, White and Clark found that with a 5% probability of type I-V error, the

---

42 R. T. White and R. M. Clark, op.cit., p.83.

probability of type V-I error increases to 60%. Also, the risk of decision error decreases as measurement error decreases for a given sample size. For example with 5% measurement error and the same parameters as before, the probability of type I-V error is 5% and type V-I is 0.5% for a sample of size 500, for a sample of size 100 the corresponding decisions errors were 5% and 35%.

The above situation could be improved by basing the significance test on all nine cells of Table II rather than on only one as done in WC. White and Clark argued that it was sufficient to use only the O2 cell for the significance test as that cell exhibits the greatest proportional change of all nine cells for changes in P_{II}. While it is true that the O2 cell is the most sensitive to changes in P_{II}, there is little doubt that a significance test based upon all of the information available would be superior to one based upon only partial information. Recently, White and Linke[43] suggested this improvement to WC but did not work out a general procedure to follow.

---

43 R. T. White and R. D. Linke, "Alternative Procedures for the White and Clark Test of Inclusion", Unpublished Paper, Monash University, 1974, p.6-9.

From the discussion thus far, WC appears to be the most satisfactory existing technique for testing the validity of learning hierarchy connections. Its superiority is due to the use of a probabilistic model which takes into account some measurement errors and provides a known sampling distribution. However, the technique may be criticized since it uses a method of parameter estimation which requires assumptions to be made about the values of two of the measurement error parameters, and also since the significance test is based upon only partial information. These criticisms imply the need for a theoretically more rigorous validation technique and indicate the manner in which it may be developed.

7. An Alternative Validation Technique.

The technique suggested as an alternative is based upon the original White and Clark model given in Table II; however, parameter estimates are obtained from cell frequencies by the method of maximum likelihood. From the maximum likelihood estimate of the proportion of the population possessing only the superordinate skill ($\hat{P}_{II}$), a confidence interval is then established about $\hat{P}_{II}$. If it is found that this confidence interval includes the value zero, the conclusion is that skill I is a prerequisite to skill II; whereas, if it is found that the interval does not include zero, the conclusion is that skill I is not prerequisite to skill II.

Before the details of the suggested technique are discussed, the theory of the method of maximum likelihood applied to the estimation of several parameters simultaneously is outlined to provide a theoretical rationale for the technique.

a) The Method of Maximum Likelihood. - Four general methods of parameter estimation could be used to estimate the parameters in the model - (1) the method of least squares, (2) the method of moments, (3) the method of Bayes, and (4) the method of maximum likelihood. Of these four methods, the least desirable one to use is the method of least squares because, unlike the other three methods, variance estimates of the parameters are not readily obtainable so the accuracy of the parameter estimates is not known. Therefore, a choice had to be made among the three remaining methods.

According to Mood and Graybill[44], moment, Bayesian and maximum likelihood estimators are squared error consistent and simple consistent under very general conditions. That is,

---

44 A. M. Mood, and F. A. Graybill, Introduction to the Theory of Statistics, 2nd ed., New York, McGraw-Hill, 1963, pp. 161-197.

the estimators become increasingly "better" as sample size increases. More precisely, if $\hat{\theta}_1, \ldots, \hat{\theta}_n, \ldots,$ is a sequence of estimators of θ, the sequence is squared error consistent if

$$\lim_{n\to\infty} E \{(\hat{\theta}_n - \theta)^2\} = 0 \tag{35}$$

for all θ in some interval. Furthermore, the sequence is simple consistent if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P(\theta - \varepsilon < \hat{\theta}_n < \theta + \varepsilon) = 1 \tag{36}$$

for all θ in some interval. Bayesian and maximum likelihood estimators, however, are superior to moment estimators because in addition to being squared error consistent and simple consistent they are asymptotically efficient and best asymptotically normal (BAN) estimators. The choice of which method to use for the technique developed in this study was therefore narrowed to a choice between the method of Bayes and maximum likelihood. Finally, the method of maximum likelihood was chosen because maximum likelihood estimators are sufficient whenever sufficient estimators exist, and unlike the method of Bayes, the method of maximum likelihood does not require any assumptions about the density functions of the parameters being estimated.

Now, the theory of the method of maximum likelihood applied to the estimation of several parameters simultaneously is reviewed. The review is based chiefly upon the work of Fisher[45], Kendall[46], Kempthorne[47], and Rao[48].

Assume that q parameters are to be estimated from n independent trials. The sample space may be partitioned into k non-overlapping events, say $E_1, \ldots, E_k$. Let x_i be the number of outcomes of E_i and let p_i = Prob (E_i) where $\sum_i^k p_i = 1$. Also assume that each p_i is defined as a function of the parameters $\theta_1, \ldots, \theta_q$ where $q < k$. Hence the probability of observing f_i outcomes of E_i in any order in n independent trials is given by the multinomial probability distribution function:

$$\text{Prob } \{x_1 = f_1, \ldots, x_f = f_k\} \doteq \frac{N!}{\prod_{i=1}^{k} f_i!} \prod_{i=1}^{k} p_i^{f_i} \quad (37)$$

where $0 \leq f_i \leq N$ and $\sum_1^k f_i = N$

The expression in (37) is known as the likelihood function (L)

---

45 R. A. Fisher, Statistical Methods for Research Workers, 8th Ed., New York, Stechert, 1941, pp. 293-325.

46 M. G. Kendal, The Advanced Theory of Statistics, Vol. 2, London, Griffin, 1948, pp. 1-45.

47 O. Kempthorne, An Introduction to Genetic Statistics, New York, Wiley, 1966, pp. 172-177.

48 C. R. Rao, Linear Statistical Inference and its Applications, New York, Wiley, 1968, pp. 302-309.

and this function is maximized by the method of maximum likelihood. One may accomplish the maximization of L by (1) finding the first partial derivatives of L with respect to each of the q parameters, (2) equating each of the first derivatives to zero, (3) solving the resulting system of q equations simultaneously and (4) checking to ensure that each of the second partial derivatives of L with respect to each of the q parameters is negative. The solutions to the system of equations would then be the maximum likelihood estimates of the parameters. In practice, however, the system of equations is often too complex to solve by ordinary methods, hence an iterative technique known as the method of scoring is often used to solve the equations.

With the method of scoring, the complexity of the calculations is often reduced by maximizing the logarithm of the likelihood function rather than the function itself. This is possible since the values of θ_q which maximize L also maximize log L because log L is a monotomic increasing function of L. Therefore, when the logarithm of expression (37) is taken, the following is obtained;

$$\log L = C + \sum_{i=1}^{k} \log p_i^{f_i} \qquad (38)$$

where C is a constant given by:

$$C = \log \frac{N!}{\prod_{i=1}^{k} f_i!} \tag{39}$$

When the partial derivatives of the logarithm of the likelihood function are taken with respect of the q parameters, and the resulting functions then equated to zero, a system of equations of the following form is obtained:

$$\frac{\partial \log L}{\partial \theta_i} = 0 \qquad \text{where } i = 1, \ldots, q. \tag{40}$$

Now any function, say f $(\theta_1, \ldots \theta_k)$, may be approximated at a point $(\theta_1 = \theta_1^o, \ldots, \theta_k = \theta_k^o)$ by a Taylor series in the following way provided the function is continuous and differentiable in an interval containing the point:

$$f(\theta_1, \ldots, \theta_k) \simeq f(\theta_1^o, \ldots, \theta_k^o) + \delta\theta_1 \frac{\partial f}{\partial \theta_1}\Big|_{(\theta_1^o, \ldots, \theta_k^o)} + \ldots + \delta\theta_k \frac{\partial f}{\partial \theta_k}\Big|_{(\theta_1^o, \ldots, \theta_k^o)} \tag{41}$$

where $\delta\theta_i = \theta_i - \theta_i^o$ for $1 \leqq i \leqq k$, and $\frac{\partial f}{\partial \theta_i}\Big|_{(\theta_1^o, \ldots, \theta_k^o)}$

means the partial derivative of the function f with respect to θ_i evaluated at the point $(\theta_1^o, \ldots, \theta_k^o)$. This type of approximation corresponds to a linear approximation of a function, and is known as a linearization of a function. Thus the function in (40) may be approximated in a similar manner:

$$\frac{\partial \log L}{\partial \theta_i} \simeq \frac{\partial \log L}{\partial \theta_i}\bigg|_{(\theta_1^o,\ldots,\theta_q^o)} + \delta\theta_1 \frac{\partial^2 \log L}{\partial\theta_1 \partial\theta_i}\bigg|_{(\theta_1^o,\ldots,\theta_q^o)} + \delta\theta_2 \frac{\partial^2 \log L}{\partial\theta_2 \partial\theta_i}\bigg|_{(\theta_1^o,\ldots,\theta_q^o)} + \ldots + \delta\theta_q \frac{\partial^2 \log L}{\partial\theta_q \partial\theta_i}\bigg|_{(\theta_1^o,\ldots,\theta_q^o)} \quad (42)$$

where $i = 1, \ldots, q$.

The expression $\frac{\partial \log L}{\partial \theta_i}\bigg|_{(\theta_1^o, \ldots, \theta_q^o)}$ is called the i^{th} efficient score and will be denoted S_i^o. Now if I_{ij}^o is defined as $-\frac{\partial^2 \log L}{\partial\theta_i \partial\theta_j}\bigg|_{(\theta_1^o,\ldots,\theta_q^o)}$ for $i, j = 1, \ldots, q$, the system in (40) may be written using the approximations in (42) as follows:

$$\begin{array}{ccccccc} I_{11}^o\,\delta\theta_1 & + & I_{12}^o\,\delta\theta_2 & + \ldots\ldots + & I_{1q}^o\,\delta\theta_q & = & S_1^o \\ \cdot & & \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot & & \cdot \\ I_{q1}^o\,\delta\theta_1 & + & I_{q2}^o\,\delta\theta_2 & + \ldots\ldots + & I_{qq}^o\,\delta\theta_q & = & S_q^o \end{array} \quad (43)$$

or more simply in matrix form as,

$$\begin{bmatrix} I^o_{11} & I^o_{12} & \cdots & I^o_{1q} \\ I^o_{21} & I^o_{22} & \cdots & . \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ I^o_{q1} & . & \cdots & I^o_{qq} \end{bmatrix} \begin{bmatrix} \delta\theta_1 \\ \delta\theta_2 \\ . \\ . \\ . \\ \delta\theta_q \end{bmatrix} = \begin{bmatrix} S^o_1 \\ S^o_2 \\ . \\ . \\ . \\ S^o_q \end{bmatrix} \qquad (44)$$

The matrix on the left side of (44) containing the elements I^o_{ij} is known as the information matrix. The elements of the information matrix and the elements of the vector of efficient scores are obtained by evaluating the appropriate expressions with approximations θ^o_1 of θ_i. The deviations, $\delta\theta_i$, of these approximations of θ_i from the maximum likelihood estimates of θ_i, can be solved for as follows:

$$\begin{bmatrix} \delta\theta_1 \\ \delta\theta_2 \\ . \\ . \\ . \\ \delta\theta_q \end{bmatrix} = \begin{bmatrix} I^o_{11} & I^o_{12} & \cdots\cdot & I^o_{1q} \\ I^o_{21} & I^o_{22} & \cdots\cdot & . \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ I^o_{q1} & . & & I^o_{qq} \end{bmatrix}^{-1} \begin{bmatrix} S^o_1 \\ S^o_2 \\ . \\ . \\ . \\ S^o_q \end{bmatrix} \qquad (45)$$

where the determinant of the information matrix is non-zero.

Thus by starting with approximations for θ_i, additive corrections $\delta\theta_i$ can be found for θ_i, and then new approximations θ_i' of θ_i can be calculated as in (46) below.

$$\begin{bmatrix} \theta_1' \\ \theta_2' \\ \cdot \\ \cdot \\ \cdot \\ \theta_q' \end{bmatrix} = \begin{bmatrix} \theta_1^o \\ \theta_2^o \\ \cdot \\ \cdot \\ \cdot \\ \theta_q^o \end{bmatrix} + \begin{bmatrix} \delta\theta_1 \\ \delta\theta_2 \\ \cdot \\ \cdot \\ \cdot \\ \delta\theta_q \end{bmatrix} \tag{46}$$

The new approximations, θ_i', are substituted back into (42), then after the information matrix and efficient score vector are recomputed, new additive corrections are found. The whole process is then repeated as many times as necessary for the additive corrections, $\delta\theta_i$, to be zero, or very close to zero, so that stable values for the parameter estimates are obtained. Care must be taken, however, that the additive corrections do in fact converge towards zero; convergence can be ensured by using efficient estimators of the parameters as the initial approximations in the iterative process. When the process is completed, the final corrected values for the parameters are good approximations of the maximum likelihood estimates of the parameters. In fact, approximations as good as necessary can be obtained since the process is controlled.

The application of the method of maximum likelihood to obtain estimates of the parameters in the White and Clark model is next discussed.

b) Maximum Likelihood Parameter Estimation in the White and Clark Model. The likelihood function, L, for the White and Clark model is:

$$L = \frac{N!}{\prod_{i,j=0}^{2} f_{ij}!} \prod_{i,j=0}^{2} p_{ij}^{f_{ij}} \tag{47}$$

where f_{ij} is the observed frequency in the ij^{th} cell of Table II, and p_{ij} is the probability of a member of a randomly selected sample being classified in the ij^{th} cell. Important to note is that the f_{ij} are treated as constants once they are observed and hence (47) is a multinomial function. In the model it can be seen that each p_{ij} is a function of θ_a, θ_b, θ_c, θ_d, P_I, P_{II}, P_O, and P_B, however, only the first seven of these parameters need to be estimated since $P_B = 1 - (P_O + P_I + P_{II})$.

Next, the logarithm of the likelihood function in (47) is found to be:

$$\log L = c + \sum_{i,j=0}^{2} f_{ij} \log p_{ij} \tag{48}$$

For simplicity let $(\theta_1, \ldots, \theta_7) = (P_O, P_I, P_{II}, \theta_a, \theta_b, \theta_c, \theta_d)$. The efficient scores are then obtained by evaluating the following expression at approximations, $\theta_1^o, \ldots, \theta_7^o$, of the parameters in the model and at particular observed cell frequencies;[49]

$$S_u^o = \frac{\delta \log L}{\delta \theta_u}\bigg|_{(\theta_1^o, \ldots, \theta_7^o)} = \sum_{i,j=0}^{2} \frac{f_{ij}}{p_{ij}} \frac{\delta p_{ij}}{\delta \theta u}\bigg|_{(\theta_1^o, \ldots, \theta_7^o)} \qquad (49)$$

where $u = 1, \ldots, 7$, and θ_u is the u^{th} coordinate of $(\theta_1, \ldots, \theta_7)$. Also, the elements of the information matrix can be found[50] by,

$$I_{ut}^o = N \sum_{i,j=0}^{2} \frac{1}{p_{ij}} \frac{\partial p_{ij}}{\partial \theta_u} \frac{\partial p_{ij}}{\partial \theta_t}\bigg|_{(\theta_1^o, \ldots, \theta_7^o)} \qquad (50)$$

where $u, t = 1, \ldots, 7$.

At this point the application of the method of maximum likelihood to estimate the parameters in the White and Clark model would best be shown by an example. Therefore, using the sample data in Table III, approximations of the parameters in

---

49 Note that the derivative of the logarithm of a function is found by multiplying the reciprocal of the function by the derivation of the function, eg.

$$\frac{\partial \log Y}{\partial X} = \frac{1}{Y} \frac{\partial Y}{\partial X}$$

50 C. R. Rao, op.cit., p. 305.

the model may be obtained by the White and Clark method discussed previously. These approximations are summarized in Table IV. The efficient scores are then obtained by substituting the parameter approximations and observed cell frequencies into (49):

$$\begin{bmatrix} S_1^o \\ S_2^o \\ S_3^o \\ S_4^o \\ S_5^o \\ S_6^o \\ S_7^o \end{bmatrix} = \begin{bmatrix} 0.08784 \\ 1.66100 \\ 100.84000 \\ {}^{-}3.11200 \\ 12.05000 \\ {}^{-}19.10900 \\ 3.35800 \end{bmatrix} \tag{51}$$

Similarly, the elements of the information matrix are found by evaluating (50):

$$\left[I_{ut}^{o} \right] = \begin{bmatrix} 2163 & 273.2 & 1928 & {}^{-}93.75 & {}^{-}335.3 & {}^{-}328.9 & {}^{-}44.69 \\ & 736.2 & 241.7 & 0.9705 & 84.58 & {}^{-}903.0 & {}^{-}100.8 \\ & & 45290 & {}^{-}1440 & {}^{-}299.2 & {}^{-}291.1 & 1508 \\ & & & 11155 & {}^{-}1063 & {}^{-}1.168 & 47.97 \\ & & & & 352.5 & {}^{-}101.8 & {}^{-}9.965 \\ & & & & & 2.563 & {}^{-}724.4 \\ & & & & & & 682.1 \end{bmatrix} \tag{52}$$

Since the information matrix is always symmetrical about the diagonal, only the diagonal and one triangular part need to be calculated.

TABLE IV

Parameter Estimates by White and Clark Method for Data of Table III

u	θ_u	$\hat{\theta}_u^o$
1	P_O	0.0886
2	P_I	0.3318
3	P_{II}	0.0000
4	θ_a	0.9731
5	θ_b	0.0000
6	θ_c	1.0000
7	θ_d	0.1929

The product of the inverse of the information matrix and the vector of efficient scores yields the following vector of corrections for the initial approximations of the parameters.

$$\begin{bmatrix} \delta P_O \\ \delta P_I \\ \delta P_{II} \\ \delta \theta a \\ \delta \theta b \\ \delta \theta c \\ \delta \theta d \end{bmatrix} = \begin{bmatrix} 0.0073 \\ -0.0967 \\ 0.0055 \\ 0.0068 \\ 0.0694 \\ -0.0611 \\ -0.0844 \end{bmatrix} \tag{53}$$

Corrected estimates of the parameters are then found by adding the vector of corrections in (53) to the initial approximations given in Table IV:

$$\begin{bmatrix} \hat{P}_O' \\ \hat{P}_I' \\ \hat{P}_{II}' \\ \hat{\theta}_a' \\ \hat{\theta}_b' \\ \hat{\theta}_c' \\ \hat{\theta}_d' \end{bmatrix} = \begin{bmatrix} \hat{P}_O + \delta P_O \\ \hat{P}_I + \delta P_I \\ \hat{P}_{II} + \delta P_{II} \\ \hat{\theta}_a + \delta \theta_a \\ \hat{\theta}_b + \delta \theta_b \\ \hat{\theta}_c + \delta \theta_c \\ \hat{\theta}_d + \delta \theta_d \end{bmatrix} = \begin{bmatrix} 0.0959 \\ 0.2351 \\ 0.0055 \\ 0.9799 \\ 0.0694 \\ 0.9378 \\ 0.1085 \end{bmatrix} \tag{54}$$

The corrected value of $\hat{P}_B$ is determined as follows:

$$\hat{P}_B' = 1 - (\hat{P}_O' + \hat{P}_{II}' + \hat{P}_I') = 0.6635 \tag{55}$$

Since the corrections in (53) were not very close to zero, the whole process is repeated using the estimates in (54) and (55) to evaluate a new set of efficient scores and a new information matrix. In the present example a total of eight iterations were necessary for each of the elements of the vector of corrections to be less than 0.0001. Table V summarizes the results of eleven iterations.

The last row of Table V gives the final maximum likelihood estimates for the model based upon the data in Table III, while the first row (iteration zero) gives the White and Clark estimates which were used as initial approximations for the maximum likelihood method. There is a noticeable discrepancy between the estimates obtained by the two methods - in some cases the estimates differ by as much as 0.10.

Next, a procedure is suggested for testing the hypothesis of skill dependency using the maximum likelihood estimate of P_{II}, its variance, and the fact that the distribution of maximum likelihood estimators is asymptotically normal.

TABLE V

Estimates of Parameters in White and Clark Model Obtained from Data in Table III by the Method of Scoring Using Eleven Iterations

Iteration Number	$\hat{P}_O$	$\hat{P}_I$	$\hat{P}_{II}$	$\hat{P}_B$	$\hat{\theta}_a$	$\hat{\theta}_b$	$\hat{\theta}_c$	$\hat{\theta}_d$
0	.0886	.3318	.0000	.5796	.9731	.0000	1.0000	.1929
1	.0959	.2351	.0055	.6635	.9799	.0694	.9389	.1085
2	.1006	.2351	.0068	.6584	.9829	.0910	.9432	.0988
3	.0996	.2308	.0070	.6624	.9824	.0866	.9406	.0938
4	.1001	.2316	.0070	.6611	.9827	.0886	.9414	.0954
5	.0996	.2312	.0070	.6617	.9826	.0878	.9410	.0947
6	.1000	.2314	.0070	.6615	.9827	.0881	.9412	.0949
7	.1000	.2314	.0070	.6616	.9827	.0880	.9411	.0949
8	.1000	.2314	.0070	.6615	.9827	.0881	.9412	.0949
9	.1000	.2314	.0070	.6615	.9827	.0881	.9412	.0949
10	.1000	.2314	.0070	.6615	.9827	.0881	.9412	.0949

c) Testing the Hypothesis of Skill Dependency. – Elandt – Johnson[51] demonstrates that the asymptotic variances and covariances of the maximum likelihood estimates of the parameters $\theta_1, \ldots, \theta_q$ are given by the matrix

$$I_{ij}^{-1} = -\frac{1}{E\left(\frac{\partial^2 \log L}{\partial\theta_i \partial\theta_j}\right)} \quad i,j = 1,\ldots,q \text{ and } \det I_{ij} \neq 0 \tag{56}$$

which is the inverse of the information matrix evaluated with the maximum likelihood estimates of the parameters. Thus the estimated variance of $\hat{\theta}_i$ is the $(i, i)^{th}$ element of the final information matrix after inversion:

$$\operatorname{var}(\hat{\theta}_i) = -\left[\frac{E(\partial^2 \log L)}{\partial\theta_i^2}\right]^{-1} = (i, i)^{th} \text{ element } (I_{ut})^{-1} \tag{57}$$

where I_{ut} is evaluated at the maximum likelihood estimate of the parameters. Since maximum likelihood estimators are approximately normally distributed for large samples, a confidence interval about θ_i may be established[52]:

$$\hat{\theta}_i - z\sqrt{\operatorname{var}(\hat{\theta}_i)} < \theta_i < \hat{\theta}_i + z\sqrt{\operatorname{var}(\hat{\theta}_i)} \tag{58}$$

where z is the normal deviate such that there is a $(1 - \alpha)$ probability between $-z$ and z.

---

51 R. C. Elandt-Johnson, Probability Models and Statistical Methods in Statistics, New York, Wiley, 1971, p. 306

52 O. Kempthorne, op.cit., p. 172.

The above information may now be applied to test the hypothesis of skill dependency - that skill I is prerequisite to skill II - in the example discussed in the last section. From the inverse of the information matrix[53], the estimated variance of $\hat{P}_{II}$ was found to be 7.451×10^{-5}; also, from Table V, $\hat{P}_{II}$ is 0.007. If one substitutes into (58) $\hat{P}_{II}$ for θ_i, the estimated variance of $\hat{P}_{II}$ for $var(\hat{\theta}_i)$, and 1.96 for z, the following 95% confidence interval can be established:

$$-0.00994 < P_{II} < 0.02394 \qquad (59)$$

For practical purposes the above interval may be regarded as equivalent to $0 \leq P_{II} < 0.02394$. Since this interval includes the value zero, it may be concluded that skill I is a prerequisite to skill II. However, if the interval had not included zero, the conclusion would have been that skill I is not prerequisite to skill II.

Herewith, an alternative probabilistic technique is available to test the validity of hierarchical connections. This alternative technique will be referred to as MAX, and will include the entire process of obtaining a maximum likelihood estimate for P_{II} from cell frequencies and establishing a confidence interval about $\hat{P}_{II}$ to determine if it is significantly different from zero.

---

53 The information matrix was inverted using the MINV subroutine of the <u>System 360 Scientific Subroutine Package</u>, Version III, Programmers Manual h20-0205-3, International Business Machines Corporation, White Plains, N.Y.,1968.

8. Summary and Hypothesis.

This chapter began with a brief discussion of learning hierarchies and their important influences on learning and instruction. Any application of learning hierarchies, however, must necessarily rest upon empirically validated hierarchies. When testing the validity of hypothesized learning hierarchies, the researcher faces the possibility of erring by concluding that a connection in a hierarchy is valid when it actually is not valid, or by concluding that the connection is not valid when it actually is valid. These types of errors were defined as type V-I and type I-V respectively.

Existing validation techniques were next reviewed. The techniques PPT, WR, and DR compare observed test item response patterns to arbitrarily determined criteria in order to test the validity of connections hypothesized to be hierarchical; the technique GSA is used to determine if a set of skills form a linear ordering; and WC uses a probabilistic model for which parameter estimates are obtained to test the hypothesis that all subjects who possess the superordinate skill also posses the subordinate skill. All of these techniques with the exception of WC were criticized on the basis that they do not take into account measurement errors, do not have a known sampling distribution, and are very dependent upon skill difficulty. Also, WC was criticized for the generally weak method of parameter estimation employed which fails to make use of all of the information available from the data and for requiring assumptions about the value of two measurement error parameters. In response to the criticisms made about existing techniques, the technique MAX was suggested as an alternative, since it uses the superior maximum likelihood method of parameter estimation and does not require assumptions about the value of any of the measurement error parameters.

Throughout the review of existing validation techniques it was remarked that there is a distinct lack of information available to guide the researcher on the risk of committing decision errors when using these techniques. From this lack of information together with the development of an alternative validation technique, the problem of the present investigation arose. The problem of this investigation is to determine under what specific conditions certain selected validation techniques may be expected to yield the least amount of decision error, and to determine which technique is the most accurate overall for testing the validity of hierarchical connections. The investigation will be limited to only four validation techniques - WR, DR, WC, and MAX.

PPT is not included in this study because the findings of available research suggest that the technique is of little or no value for testing the validity of connections. Nor is GSA included since the technique is applicable essentially only to linear hierarchies which are seldom hypothesized.

From the review of validation techniques in this chapter, three factors emerged as determiners of the accuracy of validation techniques. These factors are skill difficulty, measurement error, and sample size. Because of their important influence on the accuracy of validation techniques, the research hypothesis of this study is based upon them. Before presenting the hypothesis, a brief summary of the effects of skill difficulty, measurement error, and sample size on the accuracy of WR, DR, WC and MAX is presented as a rationale for the hypothesis.

Both WR and DR depend greatly upon skill difficulty for making a correct decision about a hierarchical connection. When f_{11} is at least 5.67 times as great in magnitude as each of f_{10}, f_{01}, and f_{00}, the probability of type V-I error for WR will be close to unity and the probability of type I-V error will be close to zero. However, when f_{11} is less than 5.67 times the magnitude of each of f_{10}, f_{01}, and f_{00}, the probabilities of decision error will be opposite. The decision errors for DR will be the same as those of WR when f_{11} is not less than

50% of the sample size, and when f_{11} is equal to or greater than 50% of the sample size respectively. On the other hand, the effect of skill difficulty on the accuracy of MAX and WC is less clear. Undoubtedly the effect will be less than it is on WR and DR because MAX and WC are based upon probabilistic models, and likely skill difficulty will affect the accuracy of MAX less than that of WC because of the theoretically superior method of parameter estimation used in MAX.

The techniques WR and DR do not take into account measurement errors at all. While WC and MAX do take measurement errors into account, MAX should be the least affected due to its maximum likelihood parameter estimation procedure which uses cell frequencies.

Since WR and DR are deterministic techniques, the effect of sample size on their accuracy is indirect and not entirely clear. Generally, it can be said that the larger the sample the more representative of the population it will be, and hence, the more accurate the techniques should be. However, since MAX and WC are theoretically superior to WR and DR they should be more accurate for any given sample size. When MAX and WC are compared, MAX should be the more accurate for the same reason as stated above.

Thus from a consideration of the characteristics on the four selected validation techniques, it was hypothesized that MAX is a more accurate validation technique than either WC, WR or DR over various sample sizes, probabilities of measurement error, and skill difficulties likely to be encountered by the researcher.

In the next chapter, the method by which this hypothesis was tested is explicated.

CHAPTER II

EXPERIMENTAL DESIGN

The research hypothesis stated in the preceding chapter was investigated through the use of simulated data. For this investigation, simulation offered two important advantages over empirical methods using live data. First, simulation allowed a comparison of validation techniques on hierarchies of known validity. If validation techniques were compared using live data seldom, if ever, would the true state of validity of a particular hierarchy be known, thus making it difficult to say with certainty whether or not the techniques yielded correct decisions. The second advantage of simulation was that validation techniques could be compared over a number of different combinations of sample sizes, skill difficulties and measurement errors with little effort and expense as compared to using live subjects.

Thus, in this chapter an outline of the method used to generate the simulated data is first given. Following this, the rationale for selecting certain sample sizes, skill difficulties and measurement errors on which to study the validation techniques is presented. The chapter is then concluded with a description of the procedure by which the research hypothesis was tested.

1. Data Generation.

Simulated data were obtained through a computer generation procedure. This procedure yielded random samples of a specified size which simulated subjects' responses to two test questions for each skill in a hierarchical connection. Since the procedure was controlled, the difficulty of the skills and the probability of the occurrence of measurement error could also be specified.

The procedure used to generate the data is illustrated in Figure 9 and is described below in seven phases. In order to clarify the description, the definitions of the P_i's and θ_i's should be recalled. The P_i's consist of P_O, P_I, P_{II} and P_B which are defined respectively as the proportions of the population possessing neither skill I nor skill II, only skill I, only skill II, and both skills I and II. The θ_i's, consist of θ_a, θ_b, θ_c and θ_d which are defined respectively as the probabilities of answering a skill I question correctly while truly possessing skill I, of answering a skill I question correctly without truly possessing skill I, of answering skill II question correctly while truly possessing skill II, and of answering a skill II question correctly without truly possessing skill II. In the text, the θ_i's are referred to as the parameters associated with measurement error; θ_a and θ_c are probabilities

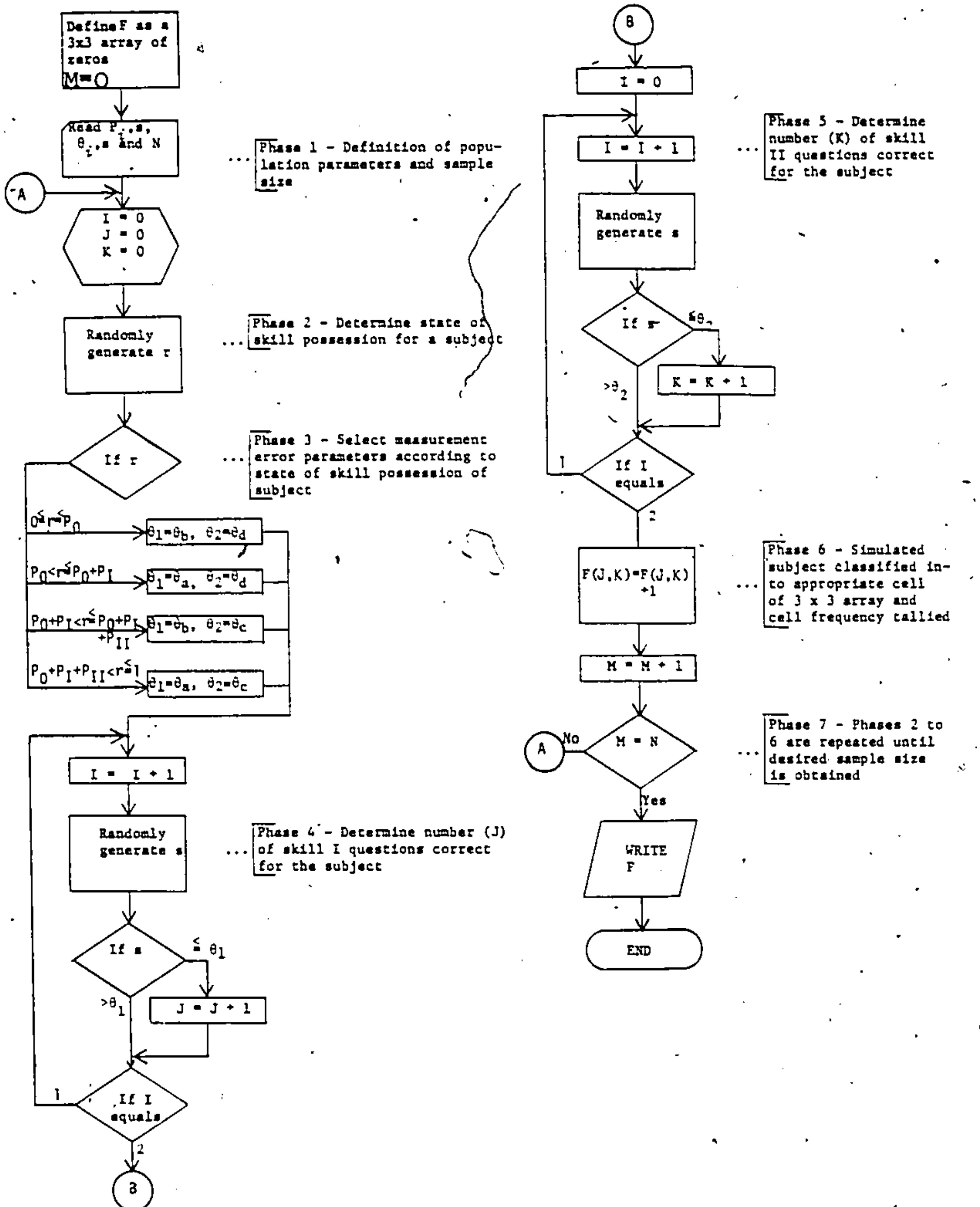

Figure 9: Flow Chart for Random Generation of Subjects' Responses to Two Questions Per Skill for a Sample of Size N.

of not blundering, θ_b and θ_d are probabilities of guessing.

Phase 1

This is the definition phase. It is in this phase that the P_i's and θ_i's are assigned specific values according this characteristics of the population being studied. If P_{II} is assigned a value of zero the skills are defined as hierarchical; otherwise if P_{II} is assigned a value greater than zero the skills are defined as not hierarchical. The difficulty of skill I is $(P_I + P_B)$ and the difficulty of skill II is $(P_{II} + P_B)$. Since the P_i's and θ_i's are proportions, they are assigned values in the range from zero to unity inclusive. Also in this phase the size of the random sample, N, to be generated is specified.

Phase 2

The state of skill possession of each simulated subject is determined in this phase according to the probability of his possessing only skill I, only skill II, both skill I and II, or neither skill as defined in phase 1. This is done by first randomly generating a real number, r, such that $0 \leq r \leq 1$. The state of skill possession is then determined by the following algorithm:

if $0 \leq r \leq P_O$, the subject possesses neither skill;

if $P_O < r \leq (P_O + P_I)$, the subject possesses only skill I;

if $(P_O + P_I) < r \leq (P_O + P_I + P_{II})$, the subject possesses only skill II;

if $(P_O + P_I + P_{II}) < r \leq 1$, the subject possesses both skill I and II.

This algorithm allows for the random samples generated to be representative of the population in terms of the proportion of subjects possessing the various combinations of skills I and II.

Phase 3

In this phase the appropriate parameters associated with measurement error are selected to correspond with the state of skill possession of the simulated subject as determined in Phase 2. That is, if the subject possesses neither skill he could answer a skill-testing question correctly only by guessing, hence the appropriate parameters would be θ_b and θ_d according to definition. Similarly, if the subject possesses only skill I he could blunder on this skill and guess the answer correctly to a skill II question; if the subject possesses only skill II, he could guess the answer to a skill I question correctly and blunder on skill II; and if the subject possesses both skills he could blunder on either one of them. In these latter situations the appropriate pairs of parameters to be selected are θ_a and θ_d, θ_b and θ_c, and θ_a and θ_c respectively. These parameters are designated as θ_1 and θ_2 in Figure 9.

Phase 4

In Phase 4 the number of questions answered correctly by a simulated subject for skill I is determined. This is accomplished by generating a random real number, s, such that $0 \leq s \leq 1$. If s is less than or equal to the probability of answering a skill I question correctly as specified in phase 3, θ_a or θ_b, a subject answers the first skill I question correctly; otherwise, he answers the question incorrectly.

By the same method as described above it is determined whether or not a subject answers the second skill I question correctly.

Phase 5

Phase 4 is repeated to determine whether a simulated subject answers 0, 1, or 2 skill II questions correctly. The appropriate measurement error parameters in this case are θ_c and θ_d.

Phase 6

In this phase a simulated subject is classified into the cell of the 3 x 3 contigency table corresponding to whether he answers 0, 1 or 2 of the skill I questions correctly and 0, 1 or 2 of the skill II questions correctly.

Phase 7

Phases 2 through 6 are repeated N - 1 more times in order to obtain the required size of random sample. Each time Phases 2 through 6 are repeated, the number of subjects in each of the cells of the 3 x 3 table is tallied.

The random numbers required of the data generation were obtained by the I.B.M. computer subroutine RANDU[1] on an I.B.M. model 360/65 machine. RANDU supplies pseudo-random numbers by the power residue method, and will generate 2^{39} numbers before repeating a sequence[2]. According to the technical staff at I.B.M.[3], the power residue method of generating random numbers has been extensively tested and found to be superior in many ways to other arithmetic methods. The RANDU subroutine is initiated by a random six digit integer called the "seed". Thereafter, the seed for RANDU for each successive random number is the random number previously generated. The use of this method prevents the duplication of a sequence of random numbers which, otherwise, would invalidate the data generation process.

A FORTRAN IV computer program was written to generate simulated data by the procedure described in this section. A listing of the program appears in Appendix A.

---

1 System/360 Scientific Subroutine Package, Version III Programmers Manual h20-0205-3, International Business Machines Corporation, White Plains, N.Y.,1968, p.77.

2 System/360 Scientific Subroutine Package, op.cit.

3 Random Number Generation and Testing, Technical Publication no. GC20-8011-0, International Business Machines Corporation, White Plains, N.Y., 1969, p.9.

2. Specification of Parameters Values and Sample Sizes.

The first phase of the random sample generation process just outlined required specification of values for the population parameters P_i and θ_i, and the sample size. The selection of these values required a compromise between choosing a sufficient number of values to simulate almost every set of conditions an experimenter might encounter when testing the validity of hierarchies, and choosing a number which the time and resources available for this study would permit. With this as a basic consideration, the rationale for selecting parameter values and sample sizes is outlined starting first with the selection of values for the P_i's.

In the first chapter after the validation techniques were examined, skill difficulty was found to be related to the risk of decision error for most validation techniques. Thus, for comparing validation techniques, it seemed logical to select values of P_i which represent a variety of levels of skill difficulty. Consequently, four levels of difficulty for skill I (d_I), the subordinate skill, were selected: 0.95, 0.70, 0.40, and 0.10. Four values appeared to be appropriate in order to avoid obtaining an excessive number of populations to study. Furthermore these values provided a reasonable range of skill difficulty, from easy (0.95) to difficult (0.10) and are approximately evenly spaced in the interval ranging from 0 to 1. It was considered necessary to have this range of skill difficulty because typically at the bottom of hierarchies skills tend to be fairly easy, while at the top skills tend to be complex and difficult; hence it is of interest to know the performance of validation techniques at a wide range of difficulty levels.

After the skill I difficulty levels were decided upon, the skill II difficulty levels (d_{II}) were selected. This selection was based upon two criteria. First, in a valid hierarchy, adjacent skills are usually close in difficulty with the subordinate skill being less difficult than the superordinate. Second, hierarchical skills are always positively correlated to a moderate extent. An analysis of data on validated hierarchies in mathematics collected by Gagné and Paradise[4] showed that hierarchical skills were inter-correlated 0.68 typically. While not every connection in the Gagné and Paradise hierarchy is likely to be valid, it is reasonable to assume that a number of connections are in fact valid because of the logical nature of the skills involved. Consequently, the Gagné and Paradise hierarchy provided a useful model to follow in this study. Thus for each of the four levels of skill I difficulty, a corresponding skill II difficulty was computed so that skill I and skill II would have a correlation

---

4 R. M. Gagné and N. E. Paradise, "Abilities and Learning Sets in Knowledge Acquisition", Psychological Monographs, 1961, Whole No. 518.

of approximately 0.68 assuming a valid connection. The computed skill II difficulties were 0.90, 0.50, 0.25, and 0.05 respectively. The pairs of skill I, skill II difficulties were defined as difficulty levels 1 to 4. In summary, difficulty levels 1 to 4 are (.95, .90), (.70, .50), (.40, .25) and (.10, .05) respectively.

For each of the four difficulty levels the corresponding P_i were computed for valid connections. These were found by using the fact that when a connection is valid P_{II} is zero; therefore $P_B = d_{II}$, $P_I = (d_I - P_B)$ and $P_O = 1 - (P_I + P_B)$. The resulting sets of values of P for the four levels of skill difficulty for valid connections were (.05, .05, .00, .90), (.30, .20, .00, .50), (.60, .15, .00, .25), and (.90, .05, .00, .05) respectively for (P_O, P_I, P_{II}, P_B). These sets of values were defined as P_i sets 1, 2, 3, and 4.

Next, for the purpose of comparing the performance of validation techniques on invalid connections in hierarchies, eight more P_i sets were defined. Sets 5 to 8 inclusive were defined as those having the same levels of skill difficulty as sets 1 to 4, however, the P_i were chosen so that skills I and II would be independent (ϕ = 0.00). Sets 9 to 12 inclusive were also defined as having the same level of skill difficulty as sets 1 to 4 but the P_i were chosen so that the skills would be positively correlated with ϕ = 0.25 and not hierarchical.

A correlation of 0.25 was used because it yielded values of P_{II} which were approximately mid-way between zero and the value for independent skills, thus providing four sets of P_i values to test how well the validation techniques could detect invalid connections. Furthermore, a correlation of this degree is typical of what might be expected among achievement test items which are homogeneous in content but not necessarily hierarchically related[5]. The resulting twelve sets of P_i used for the experimental populations are shown in Table VI.

After the P_i were selected, the parameters associated with measurement error, θ_i, were then chosen. The same probabilities of blundering and guessing on skill I questions were used for skill II questions in order to limit the number of combinations of levels of measurement error (i.e. $1-\theta_a = 1 - \theta_c$ and $\theta_b = \theta_c$). Four levels of measurement error were used in the generation of random samples. When these four levels were combined with the twelve sets of P_i, forty-eight experimental populations resulted - a number sufficient to represent the variety of hierarchies researchers are likely to encounter.

The four levels of measurement error were obtained by first setting $(1 - \theta_a)$ and $(1 - \theta_c)$, the probabilities of

---

5 F. M. Lord and M. R. Novick, Statistical Theories of Mental Test Scores, Reading, Mass., Addison Wesley, 1968, p. 337.

TABLE VI

Sets of P_i Used for Experimental Populations in Computer Generation of Random Samples[a]

Set Number	P_O	P_I	P_{II}	P_B
1	.050	.050	.000	.900
2	.300	.200	.000	.500
3	.600	.150	.000	.250
4	.900	.050	.000	.050
5	.005	.095	.045	.855
6	.150	.350	.150	.350
7	.450	.300	.150	.100
8	.855	.095	.045	.005
9	.020	.080	.030	.870
10	.210	.290	.090	.410
11	.500	.250	.100	.150
12	.870	.080	.030	.020

a Sets 1 to 4 represent populations with the connection valid (ϕ = .68) and sets 5 to 12 represent populations with the connection not valid (ϕ = .00 for sets 5 to 8; ϕ = .25 for sets 9 to 12).

blundering on a skill I and skill II question both at 0.05. The reasons for choosing this value were that rarely with carefully constructed test items would the probability of blundering on a skill-testing question be expected to be greater in practice, and by making these parameters constant, the guessing parameters, which are more likely to vary, could take on several values without resulting in an unmanageable number of combinations of values. Next, the values of θ_b and θ_d, the probabilities of answering a skill I and skill II question correctly while not possessing the skill, were selected as 0.05, 0.10, 0.20, and 0.25. The first two values were chosen so as to simulate situations whereby subjects have to supply the answer to open-ended test items, but perhaps because of testwiseness, there is a chance that some subjects may be able to guess the answer correctly on a skill-testing item. The 0.20 and 0.25 values were chosen to simulate situations where multiple choice test items are used; there would be a 0.20 probability of guessing on an item with five possible choices, and a 0.25 probability of guessing on an item with four possible choices. Finally, the selected values of the measurement error parameters were combined in the order ($1-\theta_a$, θ_b, $1-\theta_c$, θ_d) to produce the four levels of measurement error. Levels 1 to 4 were respectively defined as (.05, .05, .05, .05), (.05, .10, .05, .10), (.05, .20, .05, .20) and (.05, .25, .05, .25).

The result of the parameter values selected was that 48 different populations could be simulated. These populations represented four levels of skill difficulty and four levels of measurement error for pairs of skills which are hierarchically related, independent, and positively related but not hierarchical. Next the rationale for selecting appropriate sample sizes is described.

Samples of size 25, 40, 100 and 400 were used in the data generation. Samples of size 25 were employed to determine the accuracy of the selected validation techniques on samples the size of a typical school classroom -- the smallest sample upon which a learning hierarchy would likely be validated. The remaining three sample sizes of 40, 100 and 400 were chosen to represent, in order, samples of large school classes, several classes combined, and large groups of pupils such as all those at a particular grade level in a school or schools.

To this point, the method of generating simulated data and the specification of parameters and sample sizes used in the data generation have been described. Now the manner in which the simulated data were employed for testing the research hypothesis is presented.

3. Procedure for Testing the Hypothesis.

The research hypothesis proposed at the end of Chapter I has to do with the relative accuracy of four selected validation techniques over different sample sizes from various populations. The following procedure was used to test the hypothesis.

First, for each of the 48 simulated populations described in the last section, 1000 random samples of each of the sample sizes 25, 40, 100 and 400 were generated. Then after the random samples were generated, the validation techniques were applied to each sample to test the validity of the connection. Since the random samples were generated yielding simulated responses to two test questions per skill, a decision had to be made conerning the number of questions a simulated subject must answer correctly in order to be deemed to possess a skill for the purposes of WR and DR. Fortunately, some precedent was found in the literature. In a discussion of WR, Eisenberg and Walbesser[6] state that a subject possesses a skill if in the opinion of the instructor the task has been mastered. Gagné et al.[7] require

---

6 T. A. Eisenberg and H. H. Walbesser, "Learning Hierarchies - Numerical Considerations", Journal for Learning in Mathematics, 1971, 2, 244-256.

7 R. M. Gagné et al., op cit.

subjects to answer both of two questions before concluding that they mastered a particular task. Thus for this study the Gagné et al. criterion was adopted: a subject was given a "1" denoting possession of a skill if both skill-testing questions were answered correctly, and a "0" denoting non-possession if one or none was answered correctly.

When the validation techniques were applied to each sample, two decision criteria were used for each technique to test the validity of the connection. This was done to investigate the relationship between the decision criterion of a technique and its accuracy. WR-1 was defined as WR with the original criterion of 0.85 on all three of its ratios, and WR-2 was defined as WR with 0.85, 0.70 and 0.50 for the consistency ratio, the adequacy ratio and the completeness ratio respectively. The use of the criterion values for WR-2 is recommended by Kane, McDaniel and Phillips[8], since they suggest that the technique is too stringent. This is equivalent to stating that the probability of committing type I-V error is too great. Thus the probability of committing type I-V error should be less for WR-2 than for WR-1. A consequence of reducing the

---

8 R. B. Kane, E. D. McDaniel, and E. R. Phillips, "Analyzing Learning Relative to Transfer Relationships with Arithmetic", U.S.O.E. Project No. 9-E-128, Purdue University, 1971, p. 28.

probability of committing type I-V error, however, is to increase the probability of committing type V-I error - the more serious of the two decision errors. The use of WR-2 appears justified however because too stringent a technique would be of almost no value for, with such a technique, valid connections would rarely, if ever, be detected.

DR-1 was defined as DR using the original criterion of 0.50 as Durell suggests. In order to reduce the probability of committing type V-I error the second criterion used with DR was arbitarily set at 0.60. This technique with the second criterion was defined as DR-2.

WC was used with α set at 0.05 and 0.10. The White and Clark technique with these respective criteria was defined as WC-1 and WC-2. Theoretically the probability of committing type V-I error with WC-2 should be less than that with WC-1.

The technique MAX was used with two confidence intervals. MAX-1 was the technique with a 95% confidence interval and MAX-2 was with a 90% confidence interval. As with DR and WC use of the technique with the second criterion should reduce the probability of committing type V-I error.

Next, for every set of 1000 samples generated, the number of correct decisions yielded by each technique about the validity of the connection was tallied. An estimate of the probability of committing an incorrect decision about the validity of a connection was then obtained for each technique by subtracting the proportion of correct decisions from unity. The estimated probability of making an incorrect decision about a valid connection was defined as Prob(I-V), and of an invalid connection as Prob(V-I). Theoretically, the maximum possible standard error of a proportion[9] estimated from 1000 samples is 0.0158. This amount of error occurs when a proportion of 0.50 is estimated. Hence the accuracy of estimates of Prob (V-I) and Prob(I-V) for any given technique was considered satisfactory for the direct comparison of techniques without having to use an inferential statistical test. It is not uncommon in Monte Carlo studies to consider a proportion estimated from 1000 random samples to be close enough to the true population value for most practical purposes. Glass et al.[10] cited a number of studies where this was done for the purpose of estimating the probability of committing type II error in ANOVA and ANCOVA.

The above procedure yielded for each technique 64 estimates of each of Prob(I-V), Prob(V-I) for independent skills and Prob(V-I) for positively related skills.Each set of 64 estimates was comprised of estimates at all possible

---

9 Leslie Kish, Survey Sampling, London, Wiley, 1965, pp. 49 - 53.

10 G. V. Glass et al., "Consequences of Failure to Meet Assumptions Underlying the Analysis of Variance and Covariance", Review of Educational Research, 1972, 42, 3, 228 - 237.

combinations of the four levels of skill difficulty, four levels of measurement error and four sample sizes mentioned in the previous section. Therefore, estimates of the probability of committing both types of decision error were available for each technique at given levels of skill difficulty and measurement error for various sample sizes. These estimates were then used to test the hypothesis.

For all possible combinations of skill difficulty, measurement error and sample size, the estimates of Prob(V-I) and Prob(I-V) for related skills for each technique were examined. Next, the number of times Prob(I-V) was less than or equal to 0.20 and Prob(V-I) was less than or equal to 0.05 simultaneously was determined. The decision error criteria of requiring Prob(I-V) and Prob(V-I) to be no greater than 0.20 and 0.05 respectively were established arbitrarily. Implicit with the criteria is the conjecture that an incorrect decision about an invalid connection should not be made more than 5% of the time, and that to make an incorrect decision about an invalid connection is four times as serious as to make an incorrect decision about a valid connection. Then for the purpose of testing the research hypothesis, the most accurate technique was defined as the one with which both of the decision error criteria were satisfied most often. Therefore if the decision error criteria were satisfied most often with MAX, the research hypothesis would be supported;

otherwise it would not be supported. It is important to note that the estimates of Prob(V-I) for independent skills were not included in the test of the hypothesis. These estimates were not included because when an incorrect decision is made about an independent skill an incorrect decision must necessarily be made about the corresponding related skill.

After the hypothesis was tested the data were inspected to determine on which populations and sample sizes the validation techniques satisfied the decision error criteria simultaneously. This was considered to be an important follow-up to the hypothesis test because, ideally, if the hypothesis was supported, the technique MAX should satisfy the criteria on the same populations and sample sizes as the other three validation techniques, as well as on other populations and sample sizes. If this did not occur the support for the hypothesis would be weakened.

Since one aim of this study is to provide the researcher with a delineation of the conditions under which each validation technique is most useful, some exploratory analyses were carried out. For each technique the relationship between decision errors and skill difficulty, probability of measurement error and sample size respectively were investigated. Wherever appropriate, graphs were drawn to illustrate the result of this investigation.

FORTRAN IV computer programs were written to perform all of the computations necessary for each technique, and to tally the number of correct decisions yielded by each validation technique for each set of 1000 random samples. Listings of these programs are given in Appendix A.

In summary this chapter contained the experimental design of the study under the headings: data generation, selection of population parameters, and procedure for testing the hypothesis. The results are presented and discussed in the next chapter.

CHAPTER III

PRESENTATION AND DISCUSSION OF RESULTS

The problem investigated in this study pertains to the accuracy of four techniques used for testing the validity of connections in learning hierarchies. A research hypothesis was formulated in Chapter I about the relative accuracy of the techniques and the experimental procedure used to test the hypothesis was described in Chapter II. In this chapter, the results of testing the hypothesis are presented and discussed. Following this, each of the four validation techniques is discussed in detail and specific recommendations about their utility are made.

1. Test of the Hypothesis.

It was hypothesized that MAX is a more accurate validation technique than either WC, WR or DR over various sample sizes, probabilities of measurement error and skill difficulties likely to be encountered by the researcher. In order to test the hypothesis, data were generated to simulate subjects' responses to two questions testing possession of each of two skills. For the data generation, parameters were selected to represent four levels of skill difficulty and four probabilities of measurement error for valid connections, invalid connections where the skills are independent and invalid connections where the skills are positively related. Estimates of the probability of committing type V-I error(Prob(V-I)) and type I-V error(Prob(I-V)) with each technique were obtained for all possible combinations of sample size, measurement error and skill difficulty. The sample sizes, probabilities of measurement error and skill difficulties used represent those most likely to be encountered by the researcher. For each of the resulting 48 simulated populations, 1000 samples of size 25, 40, 100 and 400 were generated. Prob(V-I) and Prob(I-V) were then used to determine which technique is the most accurate. The most accurate technique was defined as the one with which both of the decision error criteria were simultaneously satisfied the most often.

The obtained values of Prob(V-I) and Prob(I-V) for all techniques, populations and sample sizes are given in Appendix B. Shown in Table VII is the number of times each technique satisfied the decision error criterion of Prob(I-V) $\leq$ 0.20 for valid connections, Prob(V-I) $\leq$ 0.05 for invalid connections where the skills are positively related and the number of times both these criteria were satisfied simultaneously. The maximum possible number of times the decision error criteria could be satisfied was 64 in each case. As mentioned in the last chapter, the estimates of Prob(V-I) for invalid connections where the skills are independent were not required for the hypothesis test because when an incorrect decision is made about an independent skill, an incorrect decision must necessarily be made about the corresponding related skill. An inspection of the data in Appendix B confirmed that this indeed was the case.

From Table VII one may see that with neither of the deterministic techniques - DR and WR - could the decision error criteria be reached simultaneously. However, with both WC and MAX the decision error criteria could be reached simultaneously. With WC-1 and WC-2 the criteria could be reached 9 and 6 times respectively, with MAX-1 and MAX-2 they could be reached simultaneously 12 and 14 times respectively. Since with MAX-1 and MAX-2 the criteria were reached simultaneously more often than with WC, DR and WR it was concluded that MAX is the most accurate technique.

In the previous chapter it was mentioned that support of the hypothesis is contingent upon the cases where the less accurate techniques satisfy the decision error criteria simultaneously being a subset of those of MAX; otherwise, there would be situations whereby the most accurate technique would not be the best one to use. Since neither WR nor DR satisfied the decision error criteria at all, it was necessary to examine only the cases where WC and MAX did this to see if the cases

TABLE VII

Number of Times Validation Techniques Satisfied the Decision Error Criteria on Valid Connections, Invalid Connections, and on Both Valid and Invalid Connections Simultaneously[a].

Technique	Number of times criteria satisfied[b]		
	Valid Connections	Invalid Connections (ϕ = .25)	Valid and Invalid Connections Simultaneously
WC-1	60	11	9
WC-2	51	13	6
MAX-1	64	12	12
MAX-2	64	14	14
DR-1	12	47	0
DR-2	0	51	0
WR-1	0	48	0
WR-2	0	48	0

[a] The decision error criteria were Prob(I-V) $\leq$ 0.20 for valid connections, Prob(V-I) $\leq$ 0.05 for invalid connections and both Prob(I-V) $\leq$ 0.20 and Prob(V-I) $\leq$ 0.05 for valid and invalid connections simultaneously.

[b] The maximum possible number of times the criteria could be satisfied was 64 in all three cases.

where WC satisfied the criteria are a subset of those of MAX. The results of this examination are presented in Table VIII.

Shown in Table VIII are the conditions under which the decision error criteria were attainable simultaneously for various combinations of measurement error (θ_i), sample size, and skill difficulty for MAX-2 and WC-1. MAX-1 and WC-2 are not indicated as the conditions for these techniques to satisfy the criteria are a subset of those for MAX-2 and WC-1 respectively. Samples of size 25 and 40 are not shown in Table VIII as the criteria could not be attained with either technique on samples of these sizes.

An examination of Table VIII shows that for the various combinations of measurement error, skill difficulty and sample size, the decision error criteria were not attainable for MAX-2 when they were not attainable for WC-1. Conversely, when the criteria were attainable for WC-1, they were also attainable for MAX-2. Thus it is evident that the cases when WC-1 satisfied the decision error criteria are a subset of those when MAX-2 satisfied the criteria.

Therefore the results of this study support the hypothesis that MAX is a more accurate validation technique than either WC, WR or DR over various sample sizes, probabilities of measurement error and skill difficulties likely to be encountered by the researcher. Furthermore both MAX-1 and MAX-2 are more accurate than either of the other three techniques - WC, DR and WR.

TABLE VIII

Sample Sizes, Probabilities of Measurement Error (θ_i) and Skill Difficulty Levels When Decision Error Criteria Were Satisfied Simultaneously with MAX-2 and WC-1.

Difficulty Level	N = 100				N = 400			
	θ_i = .05	.10	.20	.25	θ_i = .05	.10	.20	.25
1					x	x		
2	⊗	⊗			⊗	⊗	⊗	x
3	⊗	⊗			⊗	⊗	x	
4					x			

⊗ indicates decision error criteria satisfied by both WC-1 and MAX-2

x indicates decision error criteria satisfied by only MAX-2

Next, each of the four validation techniques studied, MAX, WC, DR and WR, are examined and discussed in detail. Three main topics are explored: (1) the most appropriate decision criterion for each technique, (2) the nature of the relationships between decision errors of a technique and measurement error, skill difficulty, and sample size, and (3) the utility of each technique. In the latter discussion specific recommendations are made about each technique. The first technique to be examined is MAX.

2. The Technique MAX.

a) Decision criterion. - The technique MAX involves the process of estimating the proportion of the population possessing only the superordinate skill in a hierarchical connection ($\hat{P}_{II}$) and constructing a confidence interval about $\hat{P}_{II}$ to determine if the interval is likely to include the value zero. In this investigation two different confidence intervals were used for MAX - 95% and 90%. The techniques using these intervals were labelled MAX-1 and MAX-2 respectively.

Theoretically, since 95% and 90% confidence intervals were used, correct decisions about valid connections should be made in the long run 95% and 90% of the time with MAX-1 and MAX-2 respectively. An inspection of the data in Appendix B shows that, in practice, one may expect to make correct decisions about valid connections with MAX-1 between 92.8%

and 100% of the time, and between 91.8% and 100% of the time with MAX-2 depending upon the population parameters and sample size. Furthermore, it was found that the number of correct decisions about valid connections yielded by MAX-1 always exceeded the number of correct decisions yielded by MAX-2 or, in terms of the probability of committing an incorrect decision about a valid connection, Prob(I-V) for MAX-1 was always less than Prob(I-V) for MAX-2.

Since type I-V error and type V-I error are related, one would expect that with MAX-2 there would be a lower risk committing type V-I error than there would be with MAX-1 in any given situation. This was confirmed by the data. Indeed, Prob(V-I) was less for MAX-2 than for MAX-1 on all samples from populations where the connection was invalid. Therefore a 90% confidence interval appears to be the more useful of the two intervals studied as type V-I error is the most serious decision error to commit.

The results also suggest the possibility of decreasing the confidence interval to 80% or 85%. This suggestion stems from the fact that on all four sample sizes and sixteen sets of population parameters, the accuracy criterion was met for type I-V error. Thus if the confidence interval is reduced, the probability of committing type V-I error would be reduced without likely exceeding the type I-V error accuracy criterion of 0.20. The result would likely be an overall increase in accuracy.

Further research into the optimum confidence interval to use with MAX at this point appears desirable. Also, as an alternative way to improve the accuracy of MAX, it may be of value to investigate the procedure recently suggested by Dayton and Macready[1] which uses χ^2 to compare vectors of observed and expected frequencies in order to test the hypothesis of skill dependency.

b) Decision Error Relationships. - As mentioned above, Pob(V-I) with MAX remained relatively stable for all population-sample size combinations studied. However, Prob(V-I) was found to be dependent upon skill difficulty, measurement error, and sample size, and hence, these relationships are discussed next. To facilitate the discussion, the relationship between the probability of making a correct decision about an invalid connection and the square root of the sample size, $\sqrt{N}$, for MAX-2 is shown in Figure 10. This relationship is shown for independent skills (ϕ=0.00) and for positively related skills (ϕ=0.25) for four probabilities of measurement error (θ_i) and four levels of skill difficulty. It is recalled from Chapter II that skill difficulty levels 1 to 4 were defined as (skill I, skill II) equal to (.95, .90), (.70, .50), (.40, .10), and (.10, .05) respectively. The relationship was plotted for only MAX-2 because MAX-2 was found to be more accurate than MAX-1, and because the relationship was the same for MAX-1 except that the probabilities of a correct decision

---

1 C. M. Dayton and G. B. Macready, "A Probabilistic Model for Validation of Behavioral Hierarchies", Psychometrika, 1976, 41, 189-204.

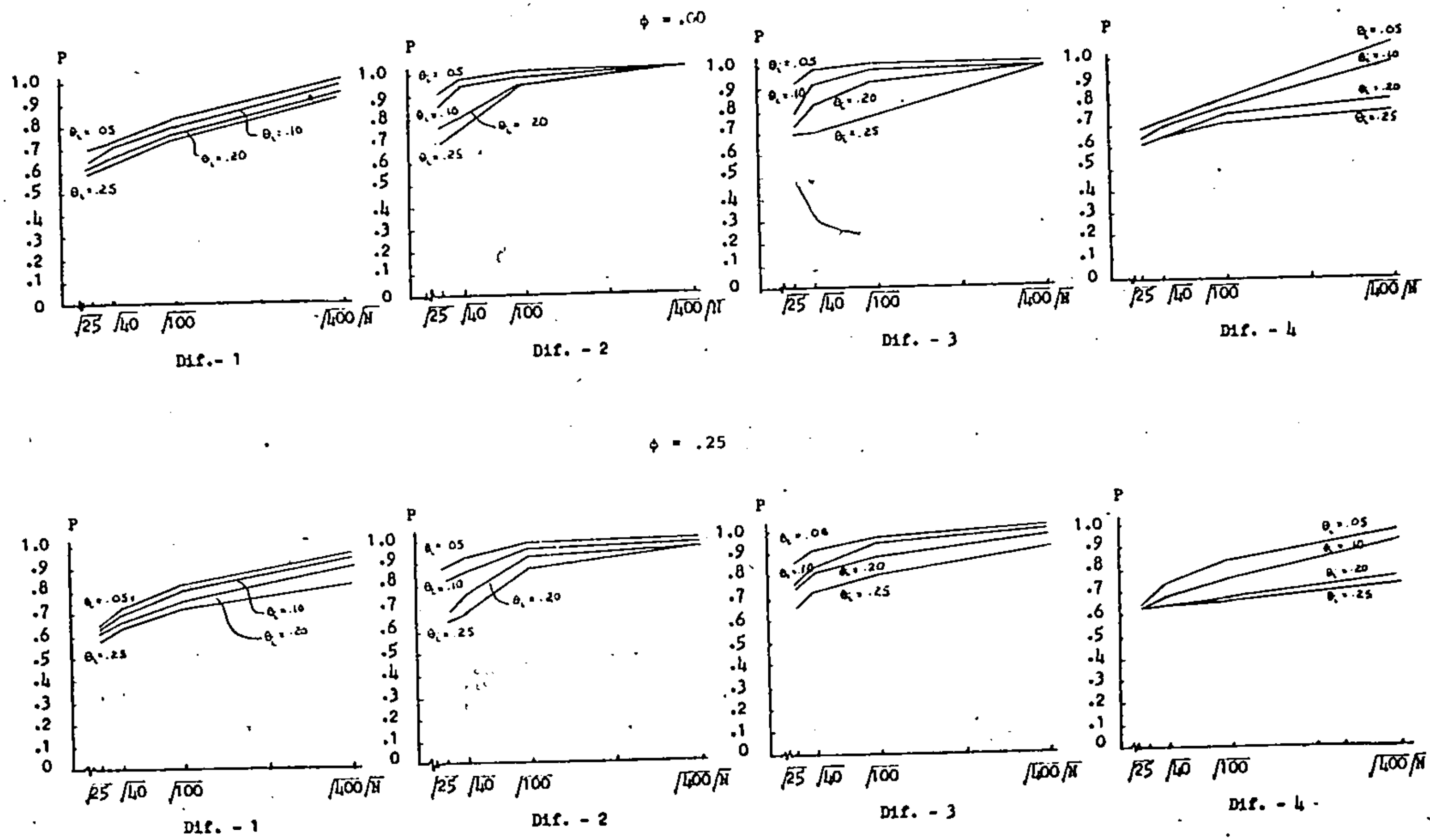

Figure 10: Relation Between Probability (P) of Making a Correct Decision About an Invalid Connection with MAX-2 and Square Root of Sample Size ($\sqrt{N}$) for Four Probabilities of Measurement Error (θ_i) at Four Levels of Skill Difficulty (Dif) When $\phi = 0$ and $\phi = .25$.

were slightly lower in all cases. In this and subsequent discussions it is important to note that Prob(V-I) and Prob (I-V) are the additive inverses of the probability of making a correct decision about an invalid and valid connection respectively.

From Figure 10, it can be seen that a relatively consistent relationship holds for all probabilities of measurement error and skill difficulty levels for $\phi = 0.00$ and $\phi = 0.25$. The relationship between the probability of a correct decision and $\sqrt{N}$ is non-linear. Prob(V-I) is relatively large for samples of size 25. This was not unexpected as the maximum likelihood method of estimation is based upon the theory of large samples. Generally speaking, when the sample size was quadrupled to 100, Prob(V-I) was reduced by approximately half. Quadrupling the sample size again to 400, resulted in a smaller proportional decrease in Prob(V-I). However, when the sample size was 400, Prob(V-I) was quite small except when $\theta'_i = 0.25$.

The relationship between probability of a correct decision and probability of measurement error can be seen from Figure 10 by comparing the vertical distance between adjacent points on the curves for θ'_i at given skill difficulty levels and sample sizes. When θ'_i is 0.05, Prob(V-I) is the least. Generally little change is noted in Prob(V-I) when θ_i is doubled to 0.10. However, when θ'_i is doubled again to 0.20 a substantial increase in Prob(V-I) occurs. When θ_i is increased to 0.25 from 0.20, Prob(V-I) increases approximately the same amount as it does when θ_i is increased from 0.05 to 0.10. Thus even though the technique MAX uses parameter estimates obtained from all frequencies the technique is definitely affected by measurement error. In order to compensate for this, larger samples should be used particularly when θ_i is expected to be 0.20 or greater. Such a situation may be expected to occur when skill possession is tested with multiple choice items.

The relationship between the probability of a correct decision and skill difficulty is generally of an inverted U-shape with the peak (the least Prob(V-I)) above the middle difficulty point. A relationship of this nature was expected with MAX because P_{II} is closer to zero for very easy or very difficult skills than it is for skills of middle difficulty. The population values of P_{II} for difficulty levels 1 to 4 are 0.03, 0.09, 0.10 and 0.03 respectively when $\phi = 0.25$ due to differences between the difficulty of skills I and II at each of the four levels. As a result estimates of P_{II} would tend to be better at the middle difficulty range than at the extremes for a given sample size and probability of measurement error.

c) Specific Recommendations. - To this point, the results of the present investigation have suggested that MAX is the most accurate of the four selected validation techniques over various levels of skill difficulty, probability of measurement error, sample sizes and skill intercorrelations. In no instance were the accuracy criteria attainable with WC, WR and DR and not with MAX. Also the results suggested that MAX-2 is slightly more accurate than MAX-1. This observation led to the suggestion that the use of MAX with 80% or 85% confidence interval be examined in future research as a means of improving the accuracy of the technique. Thus when the researcher is most concerned with the accuracy of the validation technique he is using, it is recommended that MAX be employed.

Even though MAX is the most accurate technique, caution must still be exercised when it is used to validate hierarchies. The accuracy of MAX is generally unacceptable with sample sizes smaller than 40. This is true even when the probability of measurement error is 0.05.

If the experimenter uses multiple choice test items he must be prepared to sacrifice accuracy because of the probability of subjects who do not possess the skills correctly guessing the answers. In fact the data suggest that the decision error criteria can be satisfied only at difficulty levels 2 and 3, with samples of size 400, when items having four or five possible responses are used. At difficulty level 1 only items having five possible responses can be used with samples of size 400, while at difficulty level 4, neither type of multiple choice items should be used.

The above statements are based on Prob(V-I) estimated from independent skills. When the skills are positively related (ϕ=0.25), the decision error criteria are less likely to be satisfied. For example, with samples of size 400, the criteria were satisfied with four and five alternative items only at difficulty level 2, while at difficulty level 3 they were satisfied with only five alternative items. In addition, the data suggested that neither type of multiple choice item should be used at difficulty levels 1 and 4.

A final point of caution is that attention must be given to the level of difficulty of the skills in the hierarchy. In the validation methodology suggested by Gagné, subjects are successively given instruction and immediately tested on each skill. Thus one would expect, certainly for the lower level skills in a hierarchy, that a large proportion of the subjects would possess the skill. Unfortunately, Prob(V-I) was found to be higher when testing the validity of connections with easy questions than with questions of middle difficulty. This could have been due to the fact that, in this study, skills I and II were close in difficulty at level 1, and further apart at levels 2 and 3. Nevertheless, the results suggest that at difficulty level 1 samples of size 400 must be used in order to satisfy the decision error criterion of Prob(V-I) $\leq$ 0.05. Even with samples of this size though, the probability of measurement error should not exceed 0.20 for independent skills and 0.10 for positively related skills if the criterion is to be satisfied. Therefore when the experimenter uses MAX to validate hierarchies where a large proportion of the subjects are expected to answer the skill-testing questions correctly, it is recommended that he examines Figure 10 carefully to determine the probability of committing type V-I error. If Prob(V-I) appears excessive, the experimenter may wish to delay testing so that fewer subjects would be expected to answer the test questions correctly, thereby reducing the probability of committing this kind of error.

In the next section the second most accurate technique, WC, is discussed.

3. The Technique WC.

a) Decision Criterion. - The technique WC involves obtaining estimates of four of the parameters in the White and Clark model, and then from these, p_{O2} is estimated. The estimate of p_{O2} is the probability that a member of the sample will be classified in the O2 cell of the White and Clark model. According to White and Clark, f_{O2}, the expected frequency in the O2 cell, is a binomial random variable of parameters (N, p_{O2}). If f_{O2} is observed to be significantly greater than that which would be expected by chance, the null hypothesis, H_O: $p_{II} = 0$, is rejected. Two significance levels, $\alpha = 0.05$ and $\alpha = 0.10$, were used in this study. The technique WC using these levels was defined as WC-1 and WC-2 respectively.

In the first section of this chapter it was seen that WC-1 is a slightly more accurate technique than WC-2. With WC-1 the decision error criteria were met or exceeded 10 times, while with WC-2 they were met or exceeded 6 times. This perhaps is contrary to what would be expected. It would be expected that WC-2 would be the more accurate technique as the probability of type V-I error, the most serious decision error, was found to be less overall than that of WC-1. However even though Prob (I-V) was theoretically set at 0.05 for WC-1 and 0.10 for WC-2, Prob(I-V) was found in a number of cases to be greater by more than 0.05 with WC-2 which resulted in the type I-V error criterion of 0.20 not being met as often with WC-2 as with WC-1. In fact, a closer examination of the data given in Appendix B revealed that with WC-2, Prob(I-V) increased to as much as 0.30 with samples of size 400. Although Prob(I-V) also increased with samples of size 400 with WC-1, it exceeded 0.20 only on four occasions during the experiment. The reason for this unexpected characteristic of WC is discussed in the next section, however, the results of this study suggest that using $\alpha = 0.05$ leads to greater overall accuracy with WC than $\alpha = 0.10$.

b) Decision Error Relationships. - Skill difficulty and measurement error affected type I-V error to some extent, however the effect was not serious enough to warrant further discussion. On the contrary, sample size seriously affected this kind of error, and therefore, their relationship is now examined.

In Figure 11 the relationship between the probability of making a correct decision about a valid connection with WC-1 and square root of sample size is illustrated. This relationship is shown at four levels of skill difficulty and four levels of measurement error. From the figure it can be seen that Prob (I-V) decreases as sample size is increased from 25 to 100. However as sample size is quadrupled again to 400, Prob(I-V) increases to a magnitude approximately the same or greater than that with samples of size 25. The reason for this unexpected increase in Prob(I-V) for samples of size 400 is not readily apparent but it is likely due to the measurement error assumptions made with the technique. WC requires the assumption that θ_b, the probability of correctly guessing the answer to a skill I question, is zero: and that θ_c, the probability of not blundering on a skill II question, is unity. As mentioned in the discussion of WC in Chapter I, these assumptions tend to over-estimate p_{O2}, the probability that a member of the sample will be classified in the O2 cell of the White and Clark model. A consequence of over estimating p_{O2} is that the null hypothesis, $p_{II} = 0$, is more difficult to reject, and hence that estimates of Prob(I-V) would tend to be lower than they would be if $\theta_b > 0$ and $\theta_c < 1$. During the experiment θ_b was never less than 0.05 and θ_c was set at 0.95. As sample size increased from 100 to 400, the

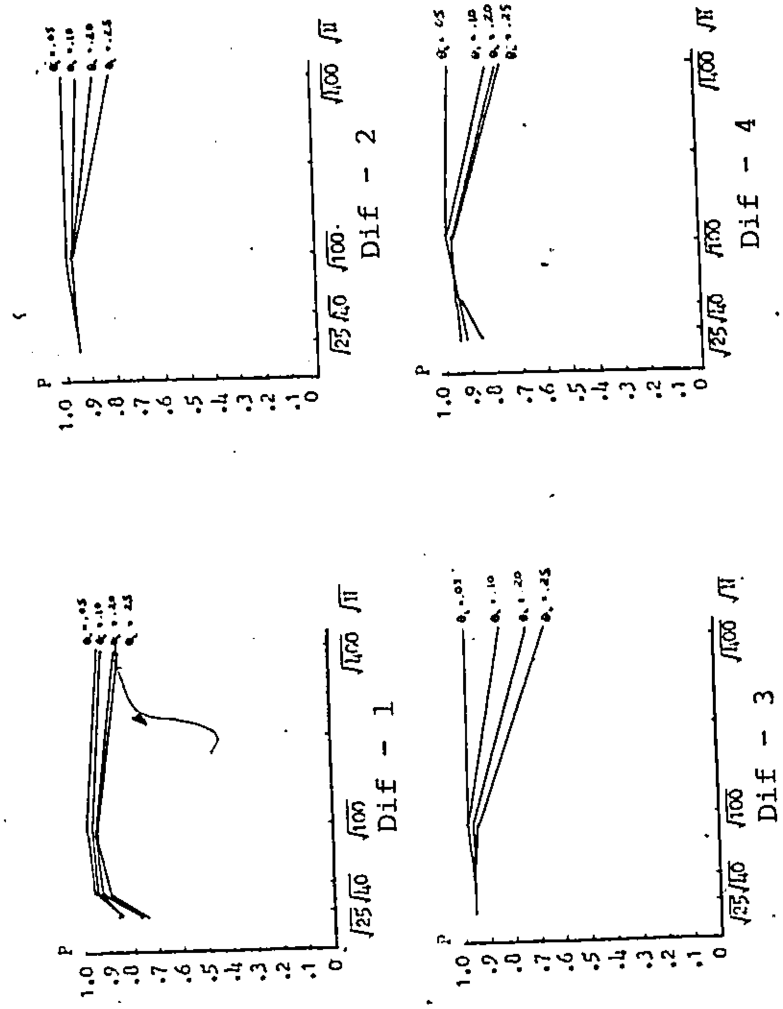

Figure 11: Relation Between Probability (P) of Making a Correct Decision About a Valid Connection with WC-1 and Square Root of Sample Size for Four Probabilities of Measurement Error(θ_i) at Four Levels of Skill Difficulty (Dif).

estimates of p_{O2} would have tended to increase in accuracy for the model with $\theta_b = 0$ and $\theta_c = 1$. This then would have resulted in a decrease in accuracy of the estimate of p_{O2} for the model with $\theta_b \geqq 0.05$ and $\theta_c = 0.95$, and hence, an overall increase in Prob (I-V) for larger samples. It is likely that the decrease in Prob(I-V) noted when sample size was increased from 25 to 100 was due to insufficient sample sizes to adequately represent the difference between the population with $\theta_b = 0$ and $\theta_c \doteq 1$ and the population with $\theta_b \geqq 0.05$ and $\theta_c = 0.95$. Further research should be directed towards this type I-V error problem with WC. In particular, sample sizes between 100 and 400 should be studied to locate the point where Prob(I-V) is least, and sample sizes greater than 400 should be studied to determine if there is a general trend for Prob(I-V) to decrease as sample size increases.

Next the type V-I error relationships are discussed.

Illustrated in Figure 12 is the relationship between probability of a correct decision about invalid connections with WC-1 and square root of sample size for four probabilities of measurement error at four levels of skill difficulty when $\phi = 0.00$ and $\phi = 0.25$. Generally the probability of a correct decision increases (Prob(V-I) decreases) rapidly when sample size is quadrupled from 25 to 100, however, when the sample size is quadrupled again the increase in probability

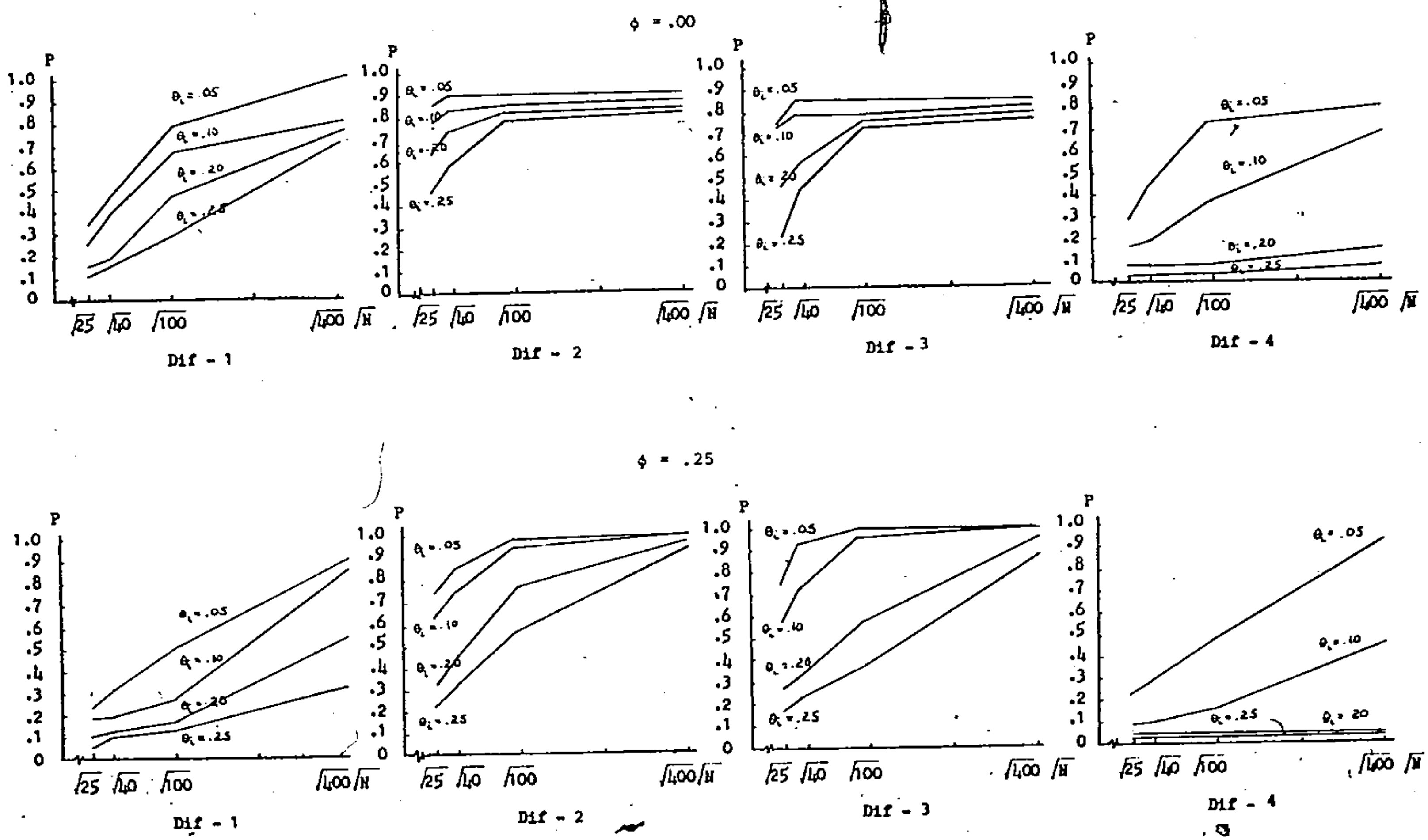

Figure 12: Relation Between Probability (P) of Making a Correct Decision About an Invalid Connection with WC-1 and Square Root of Sample Size ($\sqrt{N}$) for Four Probabilities of Measurement Error (θ_i) at Four Levels of Skill Difficulty (Dif) When $\phi = 0$ and $\phi = .25$.

of a correct decision is not as marked. In no instance could the type V-I error criterion of 0.05 be attained with samples smaller than 100.

The relationship between probability of a correct decision and square root of sample size for WC-1 is non-linear, similar to that for MAX. The chief difference between the family of curves for WC-1 and MAX-2 is that the probability of a correct decision is on the whole less for WC-1 and hence the curves are lower on the vertical axis. Another difference is that at difficulty level 4 when θ_i is equal to 0.20 and 0.25, the probability of making a correct decision is considerably lower with WC-1 than with MAX-2 at all sample sizes. As with MAX, the relationship between probability of a correct decision and square root of the sample size was not unexpected as WC is a probabilistic technique of which the accuracy of parameter estimates is related to sample size.

Also from Figure 12, the relationship between probability of a correct decision about an invalid connection and measurement error with WC-1 can be seen. The decision error criterion of 0.05 for Prob(V-I) is not always attainable even when the sample size is 400 and the probability of measurement error is 0.05. Furthermore, the decision error criterion is never attainable when the probability of measurement error is 0.20 or 0.25 on any of the sample sizes studied.

The relationship between probability of a correct decision and skill difficulty is, as with MAX-2, of an inverted U-shape. The peak of the U-shape curve is located approximately half way between difficulty levels 2 and 3. From Figure 12 it can be seen that the probability of a correct decision drops off rapidly for difficulty levels 1 and 4 except when N = 400 and $\theta_i = 0.05$ at difficulty level 1. Again this relationship could have been due to the difference between the difficulty of skills I and II at levels 1 and 4, compared to the difference between the difficulty of the skills at levels 2 and 3. However, given the present findings, the accuracy of WC is insufficient to test the validity of connections where the skills are very easy or very difficult.

c) Specific Recommendations. - The technique WC was found to be less accurate overall than MAX. WC-1 is slightly more accurate than WC-2 due to a lower probability of committing type I-V error. It was recommended that further research be aimed at identifying the reason why Prob(I-V) increased substantially with samples larger than 100.

An examination of the data suggested that WC is useful for validating hierarchies only when the skills are of middle difficulty, skill possession is tested by free response questions and samples of size 400 are used.

4. The Technique DR.

a) Decision Criterion. - DR is a deterministic technique with which ratios of observed frequencies are compared to a criterion in order to test the validity of hierarchical connections. A criterion of 0.50 was used with DR-1 and a criterion of 0.60 for DR-2. Since the decision error criteria could not be reached with either DR-1 or DR-2, another procedure was followed to determine which criterion is most appropriate. This procedure involved comparing the average number of correct decisions that were made with DR-1 and DR-2 for valid connections and for invalid connections at difficulty level 1. Comparisons could not be made at the remaining three difficulty levels because, when averaged, Prob(V-I) was zero and Prob(I-V) was unity for both DR-1 and DR-2.

From the raw data in Appendix B it was found that an average of 886 correct decisions about valid connections were made with DR-1, while an average of 561 correct decisions were made with DR-2. When the connection was invalid averages of 463 and 308 correct decisions were made with DR-1 when the skills were partially correlated and independent respectively. These same figures for DR-2 were 855 and 770. Therefore Prob(V-I) for DR-1 is 0.392 and 0.462 greater than Prob(V-I) for DR-2 for partially correlated and independent skills respectively, while Prob(I-V) for DR-1 was 0.325 less than Prob(I-V) for DR-2. Since type V-I error is considered to be approximately four times as serious as type I-V error, clearly DR-2 is more accurate overall than DR-1.

b) Decision Error Relationships. - Relationships between decision errors and sample size and probability of measurement error for DR were evident only at difficulty level 1. No relationships were evident at difficulty levels 2, 3, and 4 because, as mentioned above, Prob(V-I) tended towards zero and Prob (V-I) tended towards unity at these levels.

In Figure 13 the relationship between the probability of a correct decision with DR-2 and square root of sample size is shown for valid and invalid connections at four probabilities of measurement error. Only the relationship for DR-2 is given as it is more accurate than DR-1 and the nature of the relationship is similar for both. From the figure it is clear that both Prob(I-V) and Prob(V-I) tend to decrease slightly as the square root of sample size increases and the relationships are generally non-linear for both kinds of errors.

The relationship between probability of measurement error (θ_i) and Prob(V-I) and Prob(I-V) can be seen from Figure 13 by comparing the vertical distance between lines of θ_i at given sample sizes. Both Prob(V-I) and Prob(I-V)

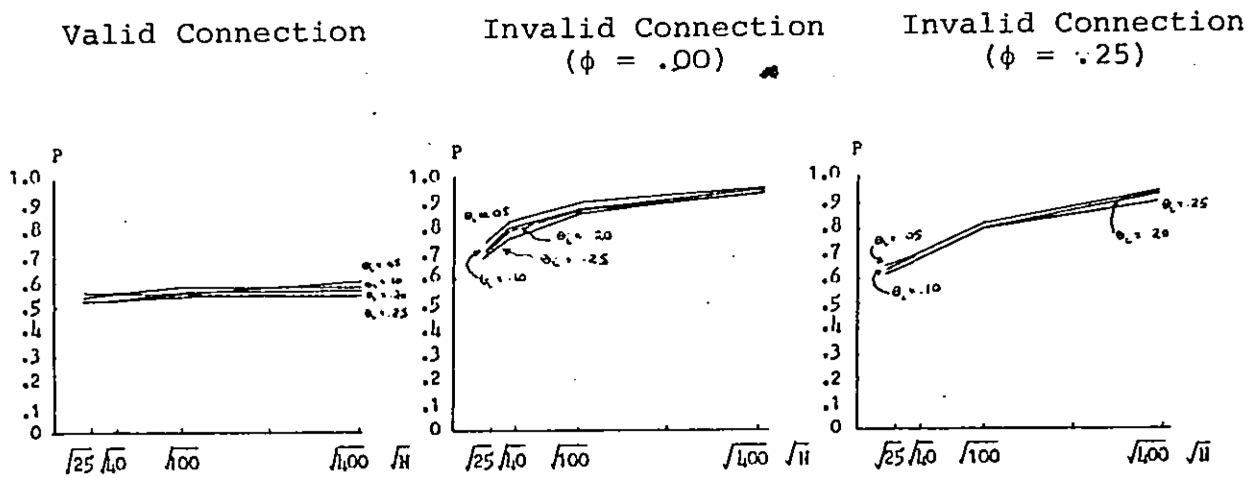

Figure 13: Relation Between Probability (P) of Making a Correct Decision about Valid and Invalid Connections with DR-2 and Square Root of Sample Size for Four Probabilities of Measurement Error (θ_i) at Skill Difficulty Level 1.

tend to increase as the probability of measurement error increases and the relationships also appear to be non-linear.

The data suggest that Prob(V-I) and Prob(I-V) are sensitive to changes in skill difficulty, however, insufficient data were gathered to show the relationship between the probability of making correct decisions and skill difficulty. In order to show these relationships a number of skill difficulties between levels 1 and 2 would have to have been used. However, it is clear that one is virtually certain never to validate a connection where the skills are as difficult as, or more difficult than the skills at level 2 ($d_I = 0.70$; $d_{II} = 0.50$).

c) Specific Recommendations. - Since the decision error criteria were never met with DR, the use of this technique cannot be recommended.

An exception to this may be when the experimenter wants to use a technique where calculations are very simple and straightforward and where he is willing to sacrifice accuracy for ease of calculation. If this is done, the decision error will be least when skills are easy(i.e. when approximately 90% of the sample can answer the skill-testing questions correctly) and when DR-2 is used. The data suggest however that Prob(I-V) may be as great as 0.45 and Prob(V-I) as great as 0.25 with certain populations and samples.

5. The Technique WR.

a) Decision Criterion. - Like DR, WR is a deterministic technique used to test the validity of connections by comparing ratios of observed frequencies. The technique consists of three ratios - the consistency ratio, the adequacy ratio, and the completeness ratio. WR-1 used the criterion of 0.85 for all three ratios; WR-2 used the criteria of 0.85, 0.70 and 0.50 respectively for the three ratios.

The decision error criteria could not be met with either WR-1 or WR-2. Thus it is useful to examine the average number of correct decisions made with WR-1 and WR-2 to determine which of the two techniques is most accurate. As was done with DR, the average numbers of correct decisions were calculated only at difficulty level 1 since Prob(V-I) and Prob(I-V) approach zero and unity respectively at difficulty levels 2, 3, and 4. With WR-1, an average of 514 correct decisions were made about valid connections, and with WR-2 an average of 882 correct decisions were made. However in the case of invalid connections, averages of 833 and 763 correct decisions were made with WR-1 when the skills were independent and partially correlated respectively; the same averages for WR-2 are 267 and 396. When the relative seriousness of type V-I and type I-V errors is taken into consideration, the data collected in this study leads one to conclude that WR-1 is more accurate than WR-2.

The above conclusion is consistent with what one would expect since lowering the 0.85 criteria for two of the three ratios of WR makes it easier to validate a connection and hence results in an increase in Prob(V-I). However the lower criteria of WR-2 were used because Kane, McDaniel and Phillips[2] suggest that the criteria of WR-1 are too stringent. It is evident from the present findings that the original 0.85 criteria are the preferred ones to use with WR.

b) Decision Error Relationships. - The relationships between decision error and sample size, probability of measurement error and skill difficulty for WR were similar to those of DR.

The relationship between the probability of making a correct decision with WR-1 and square root of sample size at four probabilities of measurement error for valid and invalid connections is shown in Figure 14. This relationship is illustrated only for WR-1, as it is more accurate than WR-2, and only at difficulty level 1 because Prob (V-I) and Prob(I-V)

---

2 R. B. Kane, E. D. McDaniel, and E. R. Phillips, "Analyzing Learning Relative to Transfer Relationships within Arithmetic", U.S.O.E. Project 9-E-128, Purdue University, 1971, p.28.

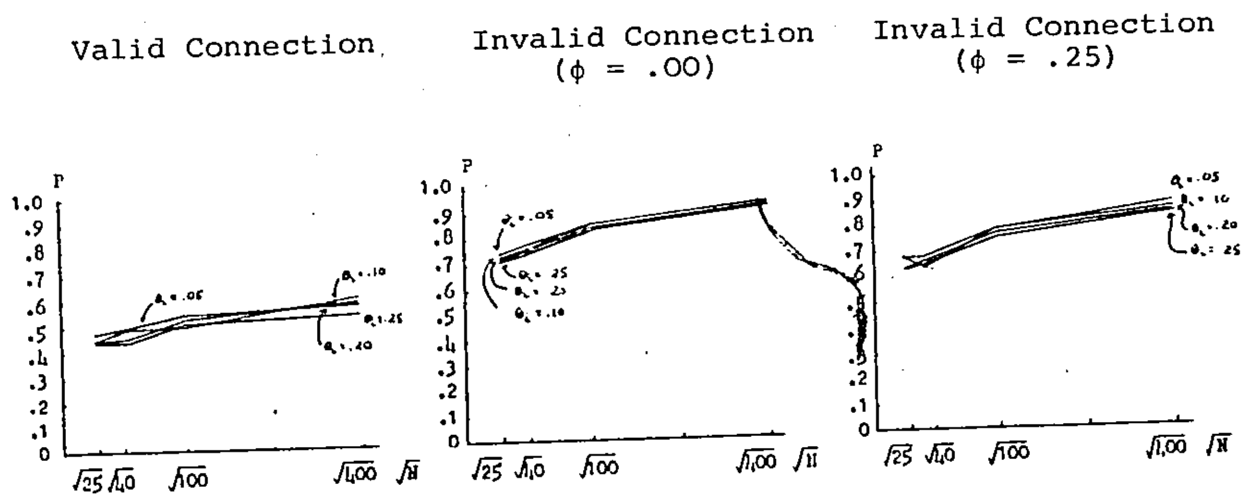

Figure 14: Relation Between Probability of Making a Correct Decision about Valid and Invalid Connections with WR-1 and Square Root of Sample Size for Four Probabilities of Measurement Error (θ_i) at Skill Difficulty Level 1.

approach zero and unity respectively at the other three difficulty levels. It is clear from Figure 14 that both Prob(V-I) and Prob(I-V) decrease non-linearly as the square root of sample size increases. A similar relationship is also evident between Prob(V-I) and Prob(I-V) and probability of measurement error when the vertical distances between lines of measurement error are compared at a given sample size.

Again like DR, WR was found to be very dependent upon skill difficulty since a connection could not be validated on skills at difficulty levels 2, 3, and 4 but for a few exceptions. The exceptions occurred at difficulty level 2 where with WR-2 a connection could be validated; however, Prob(I-V) was greater than 0.73 in these instances. In order to have illustrated the relationships between Prob(V-I) and Prob(I-V) it would have been necessary to study the technique on a number of populations where the skills were very easy. However the general relationship was that the easier the skills, the greater probability of committing type V-I error and the lesser the probability of committing type I-V error.

c) Specific Recommendations. - The findings of this study indicate that WR-1 is more accurate than WR-2. WR-1 could be used when testing the validity of connections where approximately 90% of the sample are able to answer skill-testing questions correctly. When the skills are more difficult, Prob(V-I) approaches unity and Prob(I-V) approaches zero. Even when the skills are easy, however, Prob(V-I) could well be as high as 0.30 and Prob(I-V) as high as 0.50 - unacceptable probabilities of committing these kinds of errors.

Also, the findings suggest that WR-1 is not as accurate as DR-2 overall since with DR-2 an average of 561, 855 and 770 correct decisions were made in this study for valid connections, invalid connections with independent skills and invalid connections with positively correlated skills. For WR-1 these averages were 514, 833 and 763 and hence the technique is less accurate. This conclusion supports that of Durell[3] in a computer simulation study which involved a comparison of several deterministic techniques.

---

3 A. B. Durell, "A Comparison of Measures for Validating Learning Hierarchies", Unpublished Doctoral Dissertation, University of Toronto, 1973.

SUMMARY AND CONCLUSIONS

Learning hierarchies represent groups of tasks arranged in a manner such that a substantial amount of positive transfer may be expected to occur from tasks at lower levels in the hierarchy to connected tasks at higher levels. Since Gagné's initial studies in the early 1960's a substantial body of literature on learning hierarchies has emerged. Their usefulness has been suggested in areas which include the design of instructional sequences and curricula, mastery learning and tailored testing. Despite these claims, little conclusive research evidence is available to support their utility. A major reason for this appears to be because of the lack of adequate techniques to validate postulated hierarchies.

Two general methods are available to validate learning hierarchies. These methods have been referred to as the positive and negative methods. With the positive method an attempt is made to demonstrate positive transfer by teaching subjects the skills in a postulated hierarchy in all possible combinations in order to see which combination results in the most efficient learning. On the other hand, with the negative method of validation, all subjects are taught the skills in the hierarchy and then tested on their possession of the skills. Following this, the test data are analyzed with any one of a number of available validation techniques to establish which skills in the hierarchy are necessary for possession of the terminal skill or skills. This latter method of validation has been found to be the most practical and efficient one to use and hence was selected as the focus for this study.

The validation techniques used with the negative method may be classified as either deterministic or probabilistic. With deterministic techniques, the validity of a hierarchical relationship between a pair of skills, or an entire sequence of skills, is tested by comparing ratios of observed frequencies to arbitrary criteria. However with probabilistic techniques, the parameters in an underlying model are estimated for the purpose of establishing how well the model fits the observed data.

In the first chapter of this study, available validation techniques were discussed and critically analyzed. The deterministic techniques were generally criticized for not taking into account measurement error, for being dependent upon skill difficulty and for not having known sampling distributions. Consequently, it was suggested that they would be of dubious value for testing the validity of learning hierarchies. WC, the probabilistic technique developed by White and Clark, does take into account measurement errors to some extent and has a known sampling distribution; however, it was criticized on two main points. First, it is necessary to make assumptions about the value of two measurement error parameters. Second, the significance test for the technique is based only upon partial information. It was suggested that both of these difficulties would ultimately affect the accuracy of WC in validating hierarchies and that therefore there was a need for an alternative technique.

The technique MAX was developed to overcome the above mentioned deficiencies of available validation techniques. With MAX, estimates of the parameters in the White and Clark model are obtained by the method of maximum likelihood from cell frequencies. A confidence interval is then constructed using the estimate of the probability that an individual possesses only the superordinate skill in a pair of skills hypothesized to be hierarchical and its variance estimate. If the interval includes the value zero, the conclusion is that the skills are hierarchical; otherwise the skills are not hierarchical.

Also in the first chapter, it was pointed out that no matter what validation technique is used, two kinds of error may be committed when testing the validity of a hierarchy.

One may conclude that a valid hierarchy is invalid, or that an invalid hierarchy is valid. These errors were defined as type I-V and type V-I errors respectively. Throughout the discussion of the validation techniques it was noted that little evidence is available to guide the researcher on the probability of committing either of these kinds of errors with the techniques. Thus the problem investigated in this study arose from the need for more information on the accuracy of existing validation techniques.

Theoretically, MAX appeared to be a superior technique to the other techniques reviewed, hence it was hypothesized that MAX is more accurate than those techniques over a wide variety of testing situations likely to be encountered by the researcher. In order to delimit the study, the hypothesis was tested by comparing MAX to three other techniques - WC,WR and DR. WC was included in the study as it is a probabilistic technique, while WR and DR were included despite their apparent weaknesses because they appeared to be superior to other deterministic techniques.

The hypothesis was tested using simulated data. A total of 48 experimental populations were studied. From each of these populations, 1000 random samples of four different sizes were generated by computer. Estimates of the probability of making an incorrect decision about the validity of a connection in a hierarchy (Prob(V-I) and Prob(I-V)) were obtained for each technique on the basis of sets of

1000 random samples. The most accurate technique was defined as the one with which a set of type V-I and type I-V error criteria were simultaneously reached or exceeded the most often over all population - sample size combinations studied.

The results of the study supported the hypothesis that MAX is more accurate than WC, WR and DR over various probabilities of measurement error, skill difficulties and sample sizes likely to be encountered by the researcher. Indeed, there was not a population - sample size combination, where the accuracy criteria were reached with another technique and not with MAX.

After the hypothesis was tested each of the four validation techniques - MAX, WC, WR and DR - were discussed in detail, and when appropriate, graphs were drawn to illustrate the relationship between the probability of making a correct decision about a connection and the variables of measurement error, skill difficulty and sample size. In these sections a number of observations and conclusions were made about the techniques.

First, conclusions about the most appropriate decision error criteria to use with the techniques were made. It was found that the most appropriate ones were a 90% confidence interval for MAX, $\alpha = 0.05$ for WC, a criterion of 0.60 for DR and the original criteria suggested by Walbesser for WR.

Next, the effects of measurement error, skill difficulty and sample size on the accuracy of the techniques was observed. While the accuracy of both MAX and WC was affected by these factors to some extent, MAX was the least affected. With MAX and WC, the probability of a correct decision about an invalid connection increased non-linearly as measurement error increased, decreased non-linearly as sample size increased and was much greater with skills of middle difficulty than with easy or difficult skills. The probability of making a correct decision about a valid connection with MAX was quite constant over all population - sample size combinations used, however, this probability increased with sample sizes greater than 100 with WC. Whereas, the deterministic techniques, WC and WR, were much more seriously affected by measurement error, skill difficulty and sample size. Of these three factors, the effect of skill difficulty predominated. In fact, so serious was the effect of skill difficulty on WR and DR that a connection could not be validated when much fewer than 90% of the sample answered the skill-testing questions correctly.

Following this, specific recommendations were made about each of the techniques for the purpose of providing guidelines about when the techniques may best be used. It was recommended that when MAX is used and skill possession is tested by multiple choice type items, the sample size should be at least 400 and the skills be of middle difficulty. However when the skills are tested with items where the probability of measurement error is lower (i.e. free response type items), smaller samples can be used but under no circumstances should samples smaller than 40 be employed. WC should be used only when the skills are of middle difficulty, skill possession is tested by free response type items and sample sizes of at least 400 are used. The use of either of the deterministic techniques was not recommended except perhaps that DR could be used when the skills are very easy and the experimenter wants to obtain a rough idea of whether or not a connection is valid without having to do the more complex calculations required for the probabilistic techniques.

In conclusion, several implications for future research have resulted from this investigation. Specifically, future research should be carried out to determine the optimum size of confidence interval to use with MAX since the results suggested that a 90% confidence interval is superior to a 95% interval and hence one less than 90% may be more desirable. Also research is needed to ascertain why with WC, Prob(I-V) increases as sample size increases - a difficulty which appears to be contrary to what would be expected. Furthermore, since WR and DR appear to be most effective with very easy skills, it would be useful to examine in detail the effects of skill difficulty on type V-I and type I-V errors with a number of different difficulty levels between 80% and 100%.

Finally, a much broader implication has arisen from this study for future research. Since learning hierarchy research began, most investigators have used deterministic techniques to validate hierarchies. The findings of this study suggest, however, that deterministic techniques are so innaccurate that they are virtually of no use for hierarchy validation. Consequently, every effort should be made to use probabilistic techniques, preferably MAX, whenever feasible. These efforts should result in more accurately validated hierarchies. Nevertheless caution must still be exercised when employing probabilistic techniques since with certain populations and sample sizes the decision error may be excessive.

BIBLIOGRAPHY

Cox, R. C. and Graham, G. T., "The Development of a Sequentially Scaled Achievement Test", Journal of Educational Measurement, 1966, 3, 147-150.

A frequently quoted report of research on the development of sequential achievement tests.

Durell, A. B., "A Comparison of Measures for Validating Learning Hierarchies", Unpublished Doctoral Dissertation, University of Toronto, 1973.

The technique referred to as the Durell ratio is developed in this report and its accuracy is compared to several other deterministic techniques.

Gagné, R. M., "Learning Hierarchies", Educational Psychologist, 1968, 6, 1-9.

An important paper in the historical development of learning hierarchy theory. In this paper Gagné makes a distinction between intellectual skills and verbalized knowledge.

Gagné, R. M., The Conditions of Learning, 2nd. ed., New York, Holt, Rinehart and Winston, 1970.

Gagné's most thorough treatment of the theoretical underpinnings of learning hierarchies. A must for a detailed understanding of the topic.

Gagné, R. M. and Paradise, N. E., "Abilities and Learning Sets in Knowledge Acquisition", Psychological Monographs, 1961, 75, Whole Number 518.

The first major investigation of learning hierarchies by Gagné and his co-workers. Worthwhile as background in learning hierarchy research.

Gagné, R. M. and White, R. T., "Past and Future Research on Learning Hierarchies", Educational Psychologist, 1974, 11, 19-28.

A most useful paper in that it identifies areas in which future research on learning hierarchies should be conducted and suggests methods of carrying out the research.

Glaser, R. and Resnick, L., "Instructional Psychology", Annual Review of Psychology, 1972, 23, 207-276.

The application of learning hierarchies to tailored testing is suggested in this article.

Heimer, R. T., "Conditions of Learning in Mathematics", Review of Educational Research, 1969, 39, 493-508.

An informative article on the application of learning hierarchies to mathematics learning and instruction.

Kane, R. B., Mc Daniel, E. D., and Phillips, E. R., "Analyzing Learning Hierarchies Relative to Transfer Relationships within Arithmetic", U.S.O.E. Project No. 9-E-128, Purdue University, 1971.

A report on the effectiveness of the Walbesser ratios for validating learning hierarchies.

Rao, C. R., Linear Statistical Inference and its Applications, New York, Wiley, 1968.

A most authoritative work on maximum likelihood theory and the method of scoring for parameter estimation.

Walbesser, H. H., An Evaluation Model and its Application, American Association for the Advancement of Science, Washington, D.C., 1968.

The paper in which the Walbesser validation ratios were originally presented.

Walbesser, H. H. and Eisenberg, T. A., A Review of Research on Behavioral Objectives and Learning Hierarchies, S.M.E.A.C. Report, Columbus, Ohio, ERIC Information Analysis Center, 1972.

A very comprehensive review of research on all aspects of learning hierarchies. Particularly recommended for an overview of the subject.

White, R. T., "Research into Learning Hierarchies", Review of Educational Research, 1973, 43, 361-374.

An important paper in learning hierarchy validation methodology. In this paper the author calls for a much more rigorous approach to hierarchy validation by pointing out numerous deficiencies in previous studies.

White, R. T. and Clark, R. M., "A Test of Inclusion Which Allows for Errors of Measurement", Psychometrika, 1973,38, 77-86.

The article in which the White and Clark validation technique is presented. Particularly important as it is the first probabilistic technique to be published.

APPENDIX A

Listings of Computer Programs

```
C     MAIN PROGRAM - - GENERATES RANDOM SAMPLES AND TESTS FOR VALIDITY
C                      OF CONNECTION.  REQUIRES SUBROUTINES  GEN , MAX,
C                      WR, DR, AND WC.
C
C
      DIMENSION M(4)
      DIMENSION NX(4)
      COMMON AF,AE,AD,BF,BE,BD,CF,CE,CD,N
      INTEGER AF,AE,AD,BF,BE,BD,CF,CE,CD
C     **** IS = 9 DIGIT STARTING INTERGER FOR SUBROUTINE RANDU
      IS=360500398
      M(1)=25
      M(2)=40
      M(3)=100
      M(4)=400
      DO 50 L=1,4
      N=M(L)
      NRS=0
      NI =0
      NI2=0
      NJ1=0
      NJ2=0
      NK1=0
      NK2=0
      NL1=0
      NL2=0
C     **** GENERATE 1000 RANDOM SAMPLES OF SIZE M
      DO 20 I=1,1000
C     POPULATION PARAMETERS FOLLOW
      PB=.15
      P2=.1
      P1=.25
      P0=.5
      TA=.95
      TC=.95
      TD=.25
      TB=.25
      CALL GEN(P0,P1,P2,PB,TA,TB,TC,TD,IS)
      CALL WC(I1,I2)
      CALL MAX(J1,J2)
      CALL DR(K1,K2)
      CALL WR(L1,L2)
      IR=I1+I2+J1+J2+K1+K2+L1+L2
   15 IRS=IR**2
      NRS=IRS+NRS
      NI1=NI1+I1
      NI2=NI2+I2
      NJ1=NJ1+J1
      NJ2=NJ2+J2
      NK1=NK1+K1
      NK2=NK2+K2
      NL1=NL1+L1
      NL2=NL2+L2
      JJ=I
   20 CONTINUE
      WRITE (3,60)PO,P1,P2,PB,TA,T3,TC,TD
   60 FORMAT('1',' POPULATION PARAMETERS  P0=',F7.3,'P1=',F7.3,'P2=',F7.3,'PB=',
     13,'PB=',F7.3,'TA=',F7.3,'TB=',F7.3,'TC=',F7.3,'TD=',F7.3)
      WRITE(3,59)N
   59 FORMAT('0','SAMPLE SIZE',I6)
      WRITE(3,61)
   61 FORMAT('-',40X,'NO. OF CORRECT (INCORRECT) DECISIONS')
      WRITE(3,62)
   62 FORMAT('0','TECHNIQUE  1',13X,'2',13X,'3',13X,'4',13X,'5',13X,
     1'6',13X,'7',13X,'8')
      WRITE(3,63)
   63 FORMAT('0',9X,'WC(.05)',5X,'WC (.10)',5X,'MAX (.05)',5X,
     1'MAX (.10)',5X,'DR (.5)',5X,'DR (.6)',5X,'WR (.85)',5X,'WR (.85.7
     2.5)')
      WRITE(3,30)  NI1,NI2,NJ1,NJ2,NK1,NK2,NL1,NL2
   30 FORMAT(8I13)
      STOP
      END
```

```
C.      SUBROUTINE GEN - - GENERATES A RANDOM SAMPLE ACCORDING TO
C                          PARAMETERS SPECIFIED IN MAIN PROGRAM.
C                          USES I.B.M. SSP SUBROUTINE RANDU.
C
      SUBROUTINE GEN(P0,P1,P2,PB,TA,TB,TC,TD,IS)
      COMMON AF,AE,AD,BF,BE,BD,CF,CE,CD,N
      DIMENSION M(3,3),A(4),J(4)
      DO 5 JJ=1,3
      DO 5 K=1,3
    5 M(JJ,K)=0
      RA=P0+P1
      RB=RA+P2
      DO 27 II=1,N
C     ****SELECT TRUE STATE SKILL POSSESSION FOR A SUBJECT
      CALL RANDU(IS,IS,R)
      IF (R.LE.P0) GO TO 10
      IF (R.LE.RA) GO TO 13
      IF(R.LE.RB) GO TO 16
      A(1)=TA
      A(2)=TA
      A(3)=TC
      A(4)=TC
      GO TO 18
   10 A(1)=TB
      A(2)=TB
      A(3)=TD
      A(4)=TD
      GO TO 18
   13 A(1)=TA
      A(2)=TA
      A(3)=TD
      A(4)=TD
      GO TO 18
   16 A(1)=TB
      A(2)=TB
      A(3)=TC
      A(4)=TC
C     ****SELECT OBSERVRED RESPONSES
   18 DO 24 I=1,4
      CALL RANDU(IS,IS,R)
      IF  (R-A(I))21,21,23
   21 J(I)=2
      GO TO 24
   23 J(I)=1
   24 CONTINUE
      K=5-(J(1)+J(2))
      L=J(3)+J(4)-1
   27 M(K,L)=M(K,L)+1
      AF=M(1,1)
      AE=M(1,2)
      AD=M(1,3)
      BF=M(2,1)
      BE=M(2,2)
      BD=M(2,3)
      CF=M(3,1)
      CE=M(3,2)
      CD=M(3,3)
      RETURN
      END
```

```
C     SUBROUTINE WC - - WHITE & CLARK TEST OF INCLUSION
C
      SUBROUTINE WC(IV1,IV2)
      COMMON AF,AE,AD,BF,BE,BD,CF,CE,CD,N
      COMMON /AREA1/P0,P1,P2,PB,TA,TB,TC,TD
      P2=0
      TB=0
      TC=1
      A=AF+AE+AD
      B=BF+BE+BD
      C=CF+CE+CD
      D=AD+BD+CD
      E=AE+BE+CE
      F=AF+BF+CF
   15 TA=(2*A)/(2*A+B)
      TD=E/(E+2*F)
      Q=((2*A+B)**2)/(4*A*N)
      PB=1.-(((E+2*F)**2)/(4*F*N))
      P1=Q-PB
      P0=1.-(P1+PB)
      P02=P0*TD**2+P1*((1.-TA)**2)*(TD**2)+PB*(1.-TA)**2
      Z=(CD-N*P02)/SQRT(N*P02*(1-P02))
      Z=ABS(Z)
      IF(Z.GT.1.96) GO TO 33
      IV1=1
      GO TO 44
   33 IV1=0
   44 IF(Z.GT.1.65) GO TO 66
      IV2=1
      GO TO 77
   66 IV2=0
   77 RETURN
      END
```

```
C       SUBROUTINE WR -- COMPUTES WALBESSER RATIOS
C
        SUBROUTINE WR(LV1,LV2)
        DIMENSION LV(2),W(3),CR(6)
        COMMON AF,AE,AD,BF,BE,BD,CF,CE,CD,N
      1 CR(1)=.95
        CR(2)=.95
        CR(3)=.85
        CR(4)=.85
        CR(5)=.7
        CR(6)=.5
        W(1)=AD/(AD+BD+CD)
        W(2)=AD/(AD+AF+AE)
        W (3)=AD/(AD+BF+BE+CF+CE)
        DO 12 I=1,2
        DO 6 J=1,3
        IF(W(J)-CR(J))7,6,6
      6 CONTINUE
        LV(I)=1
        GO TO 10
      7 LV(I)=0
     10 CR(1)=CR(4)
        CR(2)=CR(5)
     12 CR(3)=CR(6)
        LV1=LV(1)
        LV2=LV(2)
      4 RETURN
        END
```

```
C       SUBROUTINE DR - - COMPUTES OVERALL DIFFERENCE RATIO
C
        SUBROUTINE DR(KV1,KV2)
        DIMENSION CR(2),KV(2)
        COMMON AF,AE,AD,BF,BE,BD,CF,CE,CD,N
        CR(1)=.5
        CR(2)=.6
        DR=(AD/(N+ ABS((AF+AE)-(BF+BE+CF+CE))))-((BD+CD)/N)
        DO 10 I=1,2
        IF (DR-CR(I))7,9,9
      7 KV(I)=0
        GO TO 10
      9 KV(I)=1
     10 CONTINUE
        KV1=KV(1)
        KV2=KV(2)
        RETURN
        END
```

```
C       SUBROUTINE MAX -- MAXIMUM LIKELIHOOD TECHNIQUE
C
        SUBROUTINE MAX(JV1,JV2)
        DIMENSION S(7),ST(7,1),P(7,1),C(7,1)
        DIMENSION H(7,7),L(7),M(7)
        EQUIVALENCE(S(1),ST(1,1))
        COMMON AF,AE,AD,BF,BE,BD,CF,CE,CD,N
C       ROUGH STARTING PARAMETER ESTIMATES OBTAINED FROM SUBROUTINE MC
        COMMON /AREA1/P0,P1,P2,PB,TA,TB,TC,TD
C       **** JJ= NO. OF ITERATIONS
        DO 199 JJ=1,5
        TAA=1.-TA
        TBB=1.-TB
        TCC=1.-TC
        TDD=1.-TD
        TAS=TA**2
        TBS=TB**2
        TCS=TC**2
        TDS=TD**2
        TAAS=(1.-TA)**2
        TBBS=(1.-TB)**2
        TCCS=(1.-TC)**2
        TDDS=(1.-TD)**2
C       ********* ML **********
C       PROBABILITIES OF  A RANDOMLY SELECTED SUBJECT BEING CLASSIFIED
C       INTO EACH CELL OF 3 X 3 TABLE COMPUTED NEXT
        P20=P0*TBS*TDDS+P1*TAS*TDDS+P2*TBS*TCCS+PB*TAS*TCCS
        P21=2*P0*TBS*TD*TDD+2*P1*TAS*TD*TDD+2*P2*TBS*TC*TCC+2*PB*TAS*TC*TCC
       1C
        P22=P0*TBS*TDS+P1*TAS*TDS+P2*TBS*TCS+PB*TAS*TCS
        P10=2*(P0*TB*TBB*TDDS+P1*TA*TAA*TDDS+P2*TB*TBB*TCCS+PB*TA*TAA*TCCS
       1)
        P11=4.*(P0*TB*TBB*TD*TDD+P1*TA*TAA*TD*TDD+P2*TB*TBB*TC*TCC
       1+PB*TA*TAA*TC*TCC)
        P12=2.*(P0*TB*TBB*TDS+TA*P1*TAA*TDS+P2*TB*TBB*TCS+PB*TA*TAA*TCS)
        P00=P0*TBBS*TDDS+P1*TAAS*TDDS+P2*TBBS*TCCS+PB*TAAS*TCCS
        P01=2.*(P0*TBBS*TD*TDD+P1*TAAS*TD*TDD+P2*TBBS*TC*TCC+PB*TAAS*TC*T
       1CC)
        P02=P0*TBBS*TDS+P1*TAAS*TDS+P2*TBBS*TCS+PB*TAAS*TCS
C       ********** ML1 ***********
C       PARTIAL DERRIVATIVES COMPUTED NEXT
        P20P0=TBS*TDDS-TAS*TCCS
        P20P1=TAS*TDDS-TAS*TCCS
        P20P2=TBS*TCCS-TAS*TCCS
        P20TA=2.*(P1*TA*TDDS+PB*TA*TCCS)
        P20TB=2.*(P0*TB*TDDS+P2*TB*TCCS)
        P20TC=2.*(-P2*TBS+TC*P2*TBS-PB*TAS+TC*PB*TAS)
        P20TD=2.*(-P0*TBS+TD*P0*TBS-P1*TAS+TD*P1*TAS)
        P21P0=2.*(TBS*TD*TDD-TAS*TC*TCC)
        P21P1=2.*(TAS*TD*TDD-TAS*TC*TCC)
        P21P2=2.*(TBS*TC*TCC-TAS*TC*TCC)
        P21TA=4.*(P1*TA*TD*TDD+PB*TA*TC*TCC)
        P21TB=4.*(P0*TB*TD*TDD+P2*TB*TC*TCC)
        P21TC=2.*P2*TBS-4.*P2*TBS*TC+2.*PB*TAS-4.*PB*TAS*TC
        P21TD= 2.*P0*TBS-4*P0*TBS*TD+2.*P1*TAS-4*P1*TD*TAS
        P22P0=TBS*TDS-TAS*TCS
        P22P1=TAS*TDS-TAS*TCS
        P22P2=TBS*TCS-TAS*TCS
        P22TA=2.*(P1*TDS+PB*TCS)*TA
        P22TB=2.*(P0*TDS+P2*TCS)*TB
        P22TC=2.*(P2*TBS+PB*TAS)*TC
        P22TD=2.*(P0*TBS+P1*TAS)*TD
        P10P0=2.*(TB*TBB*TDDS-TA*TAA*TCCS)
        P10P1=2.*TA*TAA*(TDDS-TCCS)
        P10P2=2.*(TB*TBB-TA*TAA)*TCCS
        P10TA=(2.*P1-4.*P1*TA)*TDDS+(2.*PB-4.*PB*TA)*TCCS
        P10TB=(2*P0-4.*P0*TB)*TDDS+(2.*P2-4.*P2*TB)*TCCS
        P10TC=2.*(TBB*P2*TB+TAA*TA*PB)*(-2.+2.*TC)
        P10TD=2.*(P0*TB*THB+P1*TA*TAA)*(-2.+2.*TD)
        P11P0=4.*(TB*THB*TD*TDD-TA*TAA*TC*TCC)
        P11P1=4.*TA*TAA*(TD*TDD-TC*TCC)
        P11P2=4.*TC*TCC*(TB*TBB-TA*TAA)
        P11TA=(4.*P1-8.*P1*TA)*TD*TDD+(4.*PB-8.*PB*TA)*TC*TCC
        P11TB=(4.*P0-8.*P0*TB)*TD*TDD+(4.*P2-8.*P2*TB)*TC*TCC
        P11TC=4.*(1.-2.*TC)*(P2*TBB*TB+PB*TA*TAA)
        P11TD=4.*(1.-2.*TD)*(P0*TB*TBB+P1*TA*TAA)
```

```
      P00P1=TAAS*TDDS-TAAS*TCCS
      P00P0=TBBS*TDDS-TAAS*TCCS
      P00P2=TCCS*(TBBS-TAAS)
      P00TA=(-2.+2.*TA)*(P1*TDDS+PB*TCCS)
      P00TB=(-2.+2.*TB)*(P0*TDDS+P2*TCCS)
      P00TC=2.*(-1.+TC)*(P2*TBBS+PB*TAAS)
      P00TD=P0*(-2.+2.*TD)*TBBS+P1*(-2.+2.*TD)*TAAS
      P01P0=2.*(TD*TDD*TBBS-TC*TCC*TAAS)
      P01P1=2.*TAAS*(TD*TDD-TC*TCC)
      P01P2=2.*TC*TCC*(TBBS-TAAS)
      P01TA=(-4.*P1+4.*P1*TA)*TD*TDD+(-4.*PB+4.*PB*TA)*TC*TCC
      P01TB=(-4.*P0+4.*P0*TB)*TD*TDD+(-4.*P2+4.*P2*TB)*TC*TCC
      P01TC=2.*(1.-2.*TC)*(P2*TBBS+PB*TAAS)
      P01TD=2.*(1.-2.*TD)*(P0*TBBS+P1*TAAS)
      P12P0=2.*(TB*TBB*TDS-TA*TAA*TCS)
      P12P1=2.*TA*TAA*(TDS-TCS)
      P12P2=2.*(TB*TBB*TCS-TA*TAA*TCS)
      P12TA=(2.*P1-4.*P1*TA)*TDS+(2.*PB-4.*PB*TA)*TCS
      P12TB=(2.*P0-4.*P0*TB)*TDS+(2.*P2-4.*P2*TB)*TCS
      P12TC=4.*(P2*TB*TBB+PB*TA*TAA)*TC
      P12TD=4.*TD*(P0*TB*TBB+P1*TA*TAA)
      P02P0=TBBS*TDS-TAAS*TCS
      P02P1=TAAS*TDS-TAAS*TCS
      P02P2=TBBS*TCS-TAAS*TCS
      P02TA=(-2.*P1+2.*P1*TA)*TDS+(-2.*PB+2.*PB*TA)*TCS
      P02TB=(-2.*P0+2.*P0*TB)*TDS+(-2.*P2+2.*P2*TB)*TCS
      P02TC=2.*TC*(P2*TBBS+PB*TAAS)
      P02TD=2.*TD*(P0*TBBS+P1*TAAS)
C     ********* ML2 **********
C     EFFICIENCY SCORES COMPUTED NEXT
      S(1)=AF*P20P0/P20+AE*P21P0/P21+AD*P22P0/P22+BF*P10P0/P10+BE*P11P0
     1/P11+BD*P12P0/P12+CF*P00P0/P00+CE*P01P0/P01+CD*P02P0/P02
      S(2)=AF*P20P1/P20+AE*P21P1/P21+AD*P22P1/P22+BF*P10P1/P10+BE*P11P1/
     1P11+BD*P12P1/P12+CF*P00P1/P00+CE*P01P1/P01+CD*P02P1/P02
      S(3)=AF*P20P2/P20+AE*P21P2/P21+AD*P22P2/P22+BF*P10P2/P10+BE*P11P2
     1/P11+BD*P12P2/P12+CF*P00P2/P00+CE*P01P2/P01+CD*P02P2/P02
      S(4)=AF*P20TA/P20+AE*P21TA/P21+AD*P22TA/P22+BF*P10TA/P10+BE*P11TA/
     1P11+BD*P12TA/P12+CF*P00TA/P00+CE*P01TA/P01+CD*P02TA/P02
      S(5)=AF*P20TB/P20+AE*P21TB/P21+AD*P22TB/P22+BF*P10TB/P10+BE*P11TB
     1/P11+BD*P12TB/P12+CF*P00TB/P00+CE*P01TB/P01+CD*P02TB/P02
      S(5)=AF*P20TC/P20+AE*P21TC/P21+AD*P22TC/P22+BF*P10TC/P10+BE*P11TC
     1/P11+BD*P12TC/P12+CF*P00TC/P00+CE*P01TC/P01+CD*P02TC/P02
      S(7)=AF*P20TD/P20+AE*P21TD/P21+AD*P22TD/P22+BF*P10TD/P10+BE*P11TD
     1/P11+BD*P12TD/P12+CF*P00TD/P00+CE*P01TD/P01+CD*P02TD/P02
C     ******* ML3 ************
C     INFORMATION MATRIX ELEMENTS COMPUTED NEXT
      H(1,1)=P20P0**2/P20+P21P0**2/P21+P22P0**2/P22+P10P0**2/P10+P11P0**
     12/P11+P12P0**2/P12+P00P0**2/P00+P01P0**2/P01+P02P0**2/P02
      H(1,2)=P20P0*P20P1/P20+P21P0*P21P1/P21+P22P0*P22P1/P22+P10P0*P10P1/P10
     1/P10+P12P0*P12P1/P12+P00P0*P00P1/P00+P01P0*P01P1/P01+P02P0*P02P1/
     2P02+P11P0*P11P1/P11
      H(1,3)=P20P0*P20P2/P20+P21P0*P21P2/P21+P22P0*P22P2/P22+P10P0*P10P2
     1/P10+P11P0*P11P2/P11+P12P0*P12P2/P12+P00P0*P00P2/P00+P01P0*P01P2/P
     201+P02P0*P02P2/P02
      H(1,4)=P20P0*P20TA/P20+P21P0*P21TA/P21+P22P0*P22TA/P22+P10P0*P10TA
     1/P10+P11P0*P11TA/P11+P12P0*P12TA/P12+P00P0*P00TA/P00+P01P0*P01TA/
     2P01+P02P0*P02TA/P02
      H(1,5)=P20P0*P20TB/P20+P21P0*P21TB/P21+P10P0*P10TB/P10+P11P0*P11TB/P
     1/P11+P12P0*P12TB/P12+P00P0*P00TB/P00+P01P0*P01TB/P01+P02P0*P02TB/
     2P02+P22P0*P22TB/P22
      H(1,6)=P20P0*P20TC/P20+P21P0*P21TC/P21+P10P0*P10TC/P10+P11P0*P11TC
     1/P11+P12P0*P12TC/P12+P00P0*P00TC/P00+P01P0*P01TC/P01+P02P0*P02TC/
     2P02+P22P0*P22TC/P22
      H(1,7)=P20P0*P20TD/P20+P21P0*P21TD/P21+P22P0*P22TD/P22+P10P0*P10TD/
     1/P10+P12P0*P12TD/P12+P00P0*P00TD/P00+P01P0*P01TD/P01+P02P0*P02TD/
     2P02+P11P0*P11TD/P11
      H(2,2)=P20P1**2/P20+P21P1**2/P21+P22P1**2/P22+P10P1**2/P10+P11P1**2/
     12/P11+P12P1**2/P12+P00P1**2/P00+P01P1**2/P01+P02P1**2/P02
      H(2,3)=P20P1*P20P2/P20+P21P1*P21P2/P21+P22P1*P22P2/P22+P10P1*P10P2/P1
     1/P10+P11P1*P11P2/P11+P12P1*P12P2/P12+P00P1*P00P2/P00+P01P1*P01P2/P01
     201+P02P1*P02P2/P02
      H(2,4)=P20P1*P20TA/P20+P21P1*P21TA/P21+P22P1*P22TA/P22+P10P1*P10TA
     1/P10+P11P1*P11TA/P11+P12P1*P12TA/P12+P00P1*P00TA/P00+P01P1*P01TA/
     2P01+P02P1*P02TA/P02
      H(2,5)=P20P1*P20TB/P20+P21P1*P21TB/P21+P22P1*P22TB/P22+P10P1*P10TB
     1/P10+P11P1*P11TB/P11+P12P1*P12TB/P12+P00P1*P00TB/P00+P01P1*P01TB/P01+P02P1
     201+P02P1*P02TB/P02
```

```
      H(2,6)=P20P1*P20TC/P20+P21P1*P21TC/P21+P22P1*P22TC/P22+P10P1*P10TC
     1/P10+P11P1*P11TC/P11+P12P1*P12TC/P12+P00P1*P00TC/P00+P01P1*P01TC/
     2P01+P02P1*P02TC/P02
      H(2,7)=P20P1*P20TD/P20+P21P1*P21TD/P21+P22P1*P22TD/P22+P10P1*P10TD/
     1/P10+P11P1*P11TD/P11+P12P1*P12TD/P12+P00P1*P00TD/P00+P01P1*P01TD/
     2P01+P02P1*P02TD/P02
      H(3,3)=P20P2**2/P20+P21P2**2/P21+P22P2**2/P22+P10P2**2/P10+P11P2**
     12/P11+P12P2**2/P12+P00P2**2/P00+P01P2**2/P01+P02P2**2/P02
      H(3,4)=P20P2*P20TA/P20+P21P2*P21TA/P21+P22P2*P22TA/P22+P10P2*P10TA/P1
     1/P10+P11P2*P11TA/P11+P12P2*P12TA/P12+P00P2*P00TA/P00+P01P2*P01TA/
     2P01+P02P2*P02TA/P02
      H(3,5)=P20TB*P20P2/P20+P21P2*P21TB/P21+P22P2*P22TB/P22+P10P2*P10TB
     1/P10+P11P2*P11TB/P11+P12P2*P12TB/P12+P00P2*P00TB/P00+P01P2*P01TB
     2/P01+P02P2*P02TB/P02
      H(3,6)=P20P2*P20TC/P20+P21P2*P21TC/P21+P22P2*P22TC/P22+P10P2*P10TC/P1
     1/P10+P11P2*P11TC/P11+P12P2*P12TC/P12+P00P2*P00TC/P00+P01P2*P01TC
     2/P01+P02P2*P02TC/P02
      H(3,7)=P20P2*P20TD/P20+P21P2*P21TD/P21+P22P2*P22TD/P22+P10P2*P10TD
     1/P10+P11P2*P11TD/P11+P12P2*P12TD/P12+P00P2*P00TD/P00+P01P2*P01TD
     2/P01+P02P2*P02TD/P02
      H(4,4)=P20TA**2/P20+P21TA**2/P21+P22TA**2/P22+P10TA**2/P10+P11TA**
     12/P11+P12TA**2/P12+P00TA**2/P00+P01TA**2/P01+P02TA**2/P02
      H(4,5)=P20TA*P20TB/P20+P21TA*P21TB/P21+P22TA*P22TB/P22+P10TA*P10TB
     1/P10+P11TA*P11TB/P11+P12TA*P12TB/P12+P00TA*P00TB/P00+P01TA*P01TB
     2/P01+P02TA*P02TB/P02
      H(4,6)=P20TA*P20TC/P20+P21TA*P21TC/P21+P22TA*P22TC/P22+P10TA*P10TC
     1/P10+P11TA*P11TC/P11+P12TA*P12TC/P12+P00TA*P00TC/P00+P01TA*P01TC
     2/P01+P02TA*P02TC/P02
      H(4,7)=P20TA*P20TD/P20+P21TA*P21TD/P21+P22TA*P22TD/P22+P10TA*P10TD
     1/P10+P11TA*P11TD/P11+P12TA*P12TD/P12+P00TA*P00TD/P00+P01TA*P01TD
     2/P01+P02TA*P02TD/P02
      H(5,5)=P20TB**2/P20+P21TB**2/P21+P22TB**2/P22+P10TB**2/P10+P11TB**
     12/P11+P12TB**2/P12+P00TB**2/P00+P01TB**2/P01+P02TB**2/P02
      H(5,6)=P20TB*P20TC/P20+P21TB*P21TC/P21+P22TB*P22TC/P22+P10TB*P10TC
     1/P10+P11TB*P11TC/P11+P12TB*P12TC/P12+P00TB*P00TC/P00+P01TB*P01TC/
     2P01+P02TB*P02TC/P02
      H(5,7)=P20TB*P20TD/P20+P21TB*P21TD/P21+P22TB*P22TD/P22+P10TB*P10TD
     1/P10+P11TB*P11TD/P11+P12TB*P12TD/P12+P00TB*P00TD/P00+P01TB*P01TD
     2/P01+P02TB*P02TD/P02
      H(6,6)=P20TC**2/P20+P21TC**2/P21+P22TC**2/P22+P10TC**2/P10+P11TC**
     12/P11+P12TC**2/P12+P00TC**2/P00+P01TC**2/P01+P02TC**2/P02
      H(6,7)=P20TC*P20TD/P20+P21TC*P21TD/P21+P22TC*P22TD/P22+P10TC*P10TD/P1
     1/P10+P11TC*P11TD/P11+P12TC*P12TD/P12+P00TC*P00TD/P00+P01TC*P01TD/
     2P01+P02TC*P02TD/P02
      H(7,7)=P20TD**2/P20+P21TD**2/P21+P22TD**2/P22+P10TD**2/P10+P11TD**
     12/P11+P12TD**2/P12+P00TD**2/P00+P01TD**2/P01+P02TD**2/P02
      DO 80 I=1,7
      DO 80 J=I,7
   80 H(I,J)=H(I,J)*N
      DO 99 J=1,5
      K=J+1
      DO 99 I=K,7
   99 H(I,J)=H(J,I)
      H(7,6)=H(6,7)
      CALL MINV(H,7,D,L,M)
      CALL GMPRD(H,ST,C,7,7,1)
      P(1,1)=P0
      P(2,1)=P1
      P(3,1)=P2
      P(4,1)=TA
      P(5,1)=TB
      P(6,1)=TC
      P(7,1)=TD
      CALL MADD(C,P,P,7,1,0,0)
      P0=P(1,1)
      P1=P(2,1)
      P2=P(3,1)
      PB=1-(P(1,1)+P(2,1)+P(3,1))
      TA=P(4,1)
      TB=P(5,1)
      TC=P(6,1)
      TD=P(7,1)
  199 CONTINUE
```

APPENDIX A

```
      DO 198 L=1,7
198   IF(C(L,1).GT..001) GO TO 196
      GO TO 200
196   WRITE(3,197)
197   FORMAT('**** WARNING NO CONVERGENCE***')
      GO TO 207
200   SO=SQRT(H(3,3))
      Z1=1.96*SO
      Z2=1.65*SO
      IF(P2-Z1)211,211,202
211   JV1=1
      GO TO 203
202   JV1=0
203   IF (P2-Z2)204,204,206
204   JV2=1
      GO TO 207
206   JV2=0
207   CONTINUE
      RETURN
      END
```

APPENDIX B

Raw Data

Population	N	Number of Correct Decisions about Validity of Correction (maximum 1000)							
	Technique	WC	WC-1	MAX	MAX-1	DR	DR-1	WR	WR-1
1-1	25	852	846	981	973	782	563	483	792
	40	965	961	970	965	846	561	477	823
	100	995	994	982	974	944	560	499	929
	400	964	753	999	998	1000	651	607	999
1-2	25	867	862	967	964	797	540	464	784
	40	951	946	966	959	852	566	493	852
	100	993	988	968	961	952	571	551	943
	400	933	697	993	991	998	592	578	996
1-3	25	789	779	958	949	757	636	462	767
	40	930	923	945	938	831	532	456	805
	100	994	984	944	934	940	574	539	929
	400	884	708	963	959	1000	589	592	997
1-4	25	752	748	959	950	752	535	458	777
	40	895	889	951	941	802	504	444	801
	100	993	984	928	918	932	546	545	922
	400	874	694	934	923	999	563	581	998

Population	N	Number of Correct Decisions about Validity of Correction (maximum 1000)								
		Technique	WC	WC-1	MAX	MAX-1	DR	DR-1	WR	WR-1
2-1	25		962	957	998	998	55	10	1	266
	40		974	963	1000	999	27	1	0	220
	100		994	988	1000	1000	3	0	0	174
	400		996	977	1000	997	0	0	0	37
2-2	25		972	961	992	990	61	9	0	257
	40		978	963	999	997	27	1	0	226
	100		987	984	1000	1000	5	0	0	157
	400		939	755	998	993	0	0	0	33
2-3	25		965	950	966	961	50	14	0	228
	40		981	976	979	974	24	1	0	226
	100		992	975	981	978	1	0	0	147
	400		890	668	995	985	0	0	0	18
2-4	25		976	965	937	931	52	14	2	218
	40		980	971	957	954	31	1	0	226
	100		989	916	946	936	2	0	0	141
	400		833	697	979	969	0	0	0	32

Population	N	Number of Correct Decisions about Validity of Correction (maximum 1000)								
		Technique	WC	WC-1	MAX	MAX-1	DR	DR-1	WR	WR-1
3-1	25		970	961	1000	1000	0	0	0	5
	40		965	943	1000	1000	0	0	0	0
	100		992	979	1000	999	0	0	0	0
	400		998	944	1000	998	0	0	0	0
3-2	25		972	962	1000	999	0	0	0	4
	40		981	972	999	999	0	0	0	0
	100		993	988	1000	999	0	0	0	0
	400		839	684	995	988	0	0	0	0
3-3	25		980	965	995	992	0	0	0	6
	40		991	976	998	996	0	0	0	3
	100		971	844	1000	997	0	0	0	0
	400		701	533	993	979	0	0	0	0
3-4	25		989	970	986	979	0	0	0	12
	40		992	975	996	991	0	0	0	2
	100		953	859	1000	991	0	0	0	0
	400		654	509	995	977	0	0	0	0

Population	N	Technique	WC	WC-1	MAX	MAX-1	DR	DR-1	WR	WR-1
		Number of Correct Decisions about Validity of Correction (maximum 1000)								
4-1	25		851	845	989	988	0	0	0	0
	40		957	949	992	991	0	0	0	0
	100		999	996	997	997	0	0	0	0
	400		969	757	998	997	0	0	0	0
4-2	25		921	914	979	972	0	0	0	0
	40		984	981	984	978	0	0	0	0
	100		998	978	995	992	0	0	0	0
	400		782	582	999	991	0	0	0	0
4-3	25		947	939	966	956	0	0	0	0
	40		983	972	958	951	0	0	0	0
	100		978	928	982	974	0	0	0	0
	400		801	666	998	991	0	0	0	0
4-4	25		956	947	940	932	0	0	0	0
	40		986	973	941	932	0	0	0	0
	100		970	920	970	961	0	0	0	0
	400		842	707	997	989	0	0	0	0

Population	N	Technique	Number of Correct Decisions about Validity of Correction (maximum 1000)							
			WC	WC-1	MAX	MAX-1	DR	DR-1	WR	WR-1
5-1	25		336	396	696	700	466	741	747	425
	40		464	535	717	730	480	819	795	460
	100		773	814	796	812	495	913	866	381
	400		1000	1000	997	999	490	999	969	350
5-2	25		255	312	652	659	451	708	730	433
	40		380	443	709	720	483	801	783	451
	100		643	712	763	785	492	925	874	405
	400		787	801	995	996	514	998	972	353
5-3	25		168	218	641	644	441	703	748	430
	40		195	248	643	659	473	795	782	462
	100		421	526	780	795	427	904	847	372
	400		797	809	931	956	454	997	953	314
5-4	25		108	152	605	612	446	698	730	424
	40		148	191	612	632	457	774	766	437
	100		292	388	715	733	437	882	816	359
	400		747	764	845	912	402	994	945	277

Population	N	Number of Correct Decisions about Validity of Correction (maximum 1000)								
		Technique	WC	WC-1	MAX	MAX-1	DR	DR-1	WR	WR-1
6-1	25		859	875	905	921	1000	1000	1000	1000
	40		894	906	984	988	1000	1000	1000	1000
	100		884	896	1000	1000	1000	1000	1000	1000
	400		877	886	1000	1000	1000	1000	1000	1000
6-2	25		798	815	826	859	1000	1000	1000	995
	40		832	846	939	956	1000	1000	1000	1000
	100		849	867	1000	1000	1000	1000	1000	1000
	400		874	885	1000	1000	1000	1000	1000	1000
6-3	25		610	684	737	749	1000	1000	1000	998
	40		740	781	761	777	1000	1000	1000	1000
	100		818	841	971	982	1000	1000	1000	1000
	400		839	856	1000	1000	1000	1000	1000	1000
6-4	25		443	502	677	694	999	1000	1000	996
	40		579	650	709	726	1000	1000	1000	998
	100		798	820	909	938	1000	1000	1000	1000
	400		835	854	1000	1000	1000	1000	1000	1000

Population	N	Technique: WC	WC-1	MAX	MAX-1	DR	DR-1	WR	WR-1
		Number of Correct Decisions about Validity of Correction (maximum 1000)							
7-1	25	747	751	916	933	1000	1000	1000	1000
	40	851	856	983	990	1000	1000	1000	1000
	100	854	864	1000	1000	1000	1000	1000	1000
	400	826	851	1000	1000	1000	1000	1000	1000
7-2	25	750	789	778	796	1000	1000	1000	1000
	40	790	811	890	923	1000	1000	1000	1000
	100	760	768	996	997	1000	1000	1000	1000
	400	861	870	1000	1000	1000	1000	1000	1000
7-3	25	451	530	734	747	1000	1000	1000	1000
	40	567	619	795	814	1000	1000	1000	1000
	100	788	821	873	901	1000	1000	1000	1000
	400	819	852	997	998	1000	1000	1000	1000
7-4	25	229	362	689	700	1000	1000	1000	1000
	40	445	524	684	698	1000	1000	1000	1000
	100	721	779	732	753	1000	1000	1000	1000
	400	749	763	998	999	1000	1000	1000	1000

Population	N	Number of Correct Decisions about Validity of Correction (maximum 1000)								
		Technique	WC	WC-1	MAX	MAX-1	DR	DR-1	WR	WR-1
8-1	25		285	345	717	735	1000	1000	1000	1000
	40		434	498	674	689	1000	1000	1000	1000
	100		728	741	782	833	1000	1000	1000	1000
	400		799	814	999	1000	1000	1000	1000	1000
8-2	25		147	190	711	720	1000	1000	1000	1000
	40		188	249	657	681	1000	1000	1000	1000
	100		353	449	746	760	1000	1000	1000	1000
	400		781	800	870	930	1000	1000	1000	1000
8-3	25		50	74	647	657	1000	1000	1000	1000
	40		51	85	691	706	1000	1000	1000	1000
	100		62	112	668	687	1000	1000	1000	1000
	400		192	299	770	780	1000	1000	1000	1000
8-4	25		21	40	605	624	1000	1000	1000	1000
	40		36	57	673	691	1000	1000	1000	1000
	100		24	57	707	717	1000	1000	1000	1000
	400		75	157	729	748	1000	1000	1000	1000

Population	N	Number of Correct Decisions about Validity of Correction (maximum 1000)							
Technique		WC	WC-1	MAX	MAX-1	DR	DR-1	WR	WR-1
9-1	25	220	266	655	660	478	668	683	345
	40	313	362	692	705	371	693	612	372
	100	521	591	811	831	306	807	776	250
	400	929	962	957	959	176	965	894	98
9-2	25	184	222	639	644	389	639	684	356
	40	197	248	683	693	364	693	696	351
	100	389	467	775	788	316	814	778	277
	400	849	890	944	953	167	967	887	117
9-3	25	102	134	623	631	359	606	632	343
	40	134	179	627	646	337	668	680	358
	100	182	251	706	732	282	793	751	261
9-4	25	64	85	585	592	335	606	640	323
	40	101	144	621	633	333	699	689	338
	100	139	203	693	708	301	803	748	272
	400	353	440	809	825	140	939	823	95

Number of Correct Decisions
about Validity of Correction (maximum 1000)

Population	N	Technique WC	WC-1	MAX	MAX-1	DR	DR-1	WR	WR-1
10-1	25	765	799	880	885	994	999	999	970
	40	864	889	911	914	999	1000	1000	988
	100	994	998	990	991	1000	1000	1000	1000
	400	1000	1000	1000	1000	1000	1000	1000	1000
10-2	25	615	663	787	818	893	999	1000	964
	40	731	770	840	863	1000	1000	1000	990
	100	959	970	973	976	1000	1000	1000	999
	400	1000	1000	999	999	1000	1000	1000	1000
10-3	25	313	378	670	689	992	999	999	966
	40	415	491	712	746	998	1000	1000	974
	100	723	789	885	903	1000	1000	1000	1000
	400	993	965	998	999	1000	1000	1000	1000
10-4	25	265	325	627	657	995	1000	1000	968
	40	308	376	649	679	999	1000	1000	988
	100	557	641	828	853	1000	1000	1000	1000
	400	974	986	993	994	1000	1000	1000	1000

Population	N	Number of Correct Decisions about Validity of Correction (maximum 1000)								
		Technique	WC	WC-1	MAX	MAX-1	DR	DR-1	WR	WR-1
11-1	25		745	782	847	856	1000	1000	1000	1000
	40		912	931	909	923	1000	1000	1000	1000
	100		996	997	970	972	1000	1000	1000	1000
	400		1000	1000	999	999	1000	1000	1000	1000
11-2	25		589	646	748	764	1000	1000	1000	1000
	40		735	778	817	836	1000	1000	1000	1000
	100		952	968	956	961	1000	1000	1000	1000
	400		1000	1000	996	996	1000	1000	1000	1000
11-3	25		288	334	626	645	1000	1000	1000	1000
	40		323	399	686	722	1000	1000	1000	1000
	100		598	688	848	876	1000	1000	1000	1000
	400		981	989	992	992	1000	1000	1000	1000
11-4	25		180	236	654	663	1000	1000	1000	1000
	40		237	307	728	731	1000	1000	1000	1000
	100		391	483	772	776	1000	1000	1000	1000
	400		842	854	878	924	1000	1000	1000	1000

Population	N	Technique								
		Number of Correct Decisions about Validity of Correction (maximum 1000)								
			WC	WC-1	MAX	MAX-1	DR	DR-1	WR	WR-1
12-1	25		222	271	658	663	1000	1000	1000	1000
	40		291	351	721	728	1000	1000	1000	1000
	100		497	570	829	845	1000	1000	1000	1000
	400		931	956	949	952	1000	1000	1000	1000
12-2	25		74	107	623	643	1000	1000	1000	1000
	40		100	149	671	688	1000	1000	1000	1000
	100		145	211	732	760	1000	1000	1000	1000
	400		418	540	924	931	1000	1000	1000	1000
12-3	25		33	53	633	647	1000	1000	1000	1000
	40		33	43	649	658	1000	1000	1000	1000
	100		22	49	645	665	1000	1000	1000	1000
	400		31	69	708	721	1000	1000	1000	1000
12-4	25		13	24	631	651	1000	1000	1000	1000
	40		20	40	654	663	1000	1000	1000	1000
	100		15	40	644	654	1000	1000	1000	1000
	400		20	50	687	704	1000	1000	1000	1000

APPENDIX C

ABSTRACT OF

A Comparison of Learning Hierarchy Validation Techniques Using Simulated Data[1]

Learning hierarchy theory appears to provide a useful model to aid in instructional design. However, to date, progress in the implementation of learning hierarchy theory to practice has been hindered by the lack of adequate techniques to validate hypothesized hierarchies.

In this investigation available validation techniques were critically reviewed for the purpose of providing a rationale upon which an alternative validation technique could be developed. An alternative technique was then developed using the method of maximum likelihood for parameter estimation. The alternative validation technique was compared to three other validation techniques currently in use - the Walbesser Ratios, the Durell Difference Ratio, and the White and Clark Test of Inclusion. This comparison followed a Monte Carlo procedure and was based upon four different sample sizes from each of 48 populations simulating valid and invalid hierarchies.

---

1 Ronald D. Owston, Doctoral thesis presented to the Faculty of Education of the University of Ottawa, Ontario, March, 1978.

The results supported the hypothesis that the alternative validation technique developed in this study was more accurate, over a wide variety of sample sizes and populations, than the other three techniques studied. A detailed examination of each of the validation techniques was then made. From this examination several problems associated with various techniques were identified and specific recommendations concerning the utility of each technique were made.

A major implication of this study is that learning hierarchies should be validated with probabilistic techniques whenever possible, preferably with the maximum likelihood technique.