Portland State University
PDXScholar

Dissertations and Theses   Dissertations and Theses

1983

Derivation and test of predictions of a discrete latent state model for signed number addition test performance

Kentaro Yamamoto
Portland State University

Let us know how access to this document benefits you.

Follow this and additional works at: http://pdxscholar.library.pdx.edu/open_access_etds

🌀 Part of the Discrete Mathematics and Combinatorics Commons, and the Psychology Commons

AN ABSTRACT OF THE THESIS OF Kentaro Yamamoto for the Master of Science in Psychology presented July 27, 1983.

Title: Derivation and Test of Predictions of a Discrete Latent State Model for Signed Number Addition Test Performance.
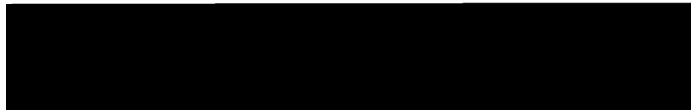
APPROVED BY MEMBERS OF THE THESIS COMMITTEE:

James A. Paulson, Chairman

Chadwick Karr

Craig A. Magwire

This study is an investigation of the performance of a discrete latent state model devised by Paulson (1982) to account for signed-number arithmetic test data gathered by Birenbaum and Tatsuoka (1980). One hundred twenty nine students took a test which consists of sixteen item types

with four parallel arithmetic items of each type. The present study utilizes the five addition item types of four items each; hence, there are four parallel subtests. Responses to the addition items can be analyzed in terms of two components: the sign component (is the sign correct?), and the absolute value component (is the size of the answer correct?). Paulson's model describes how students perform on the two components separately and how the component responses are related. This study examines the parallelism of the four subtests, in terms of equality of means, standard deviations, and correlations between all pairs of subtests. Decision consistency between subtests is another useful indicator of measurement reliability, particularly for tests of concept mastery. The model implies that the consistency between any two pairs of subtests should be equal; this implication is tested. The specific numerical values predicted by the model for the means, standard deviations, correlations, and decision consistency indices are tested against the corresponding observed statistics. All the analyses described so far are done separately for both the sign and the absolute value components of the responses. A method to synthesize overall correct response from estimated parameter values of two components is derived and tested against observed values. The results are that "parallel" items within item types are not all parallel and finer characterization would be needed to describe the items

completely. However, the deviations from strict parallelism are slight. Paulson's model demonstrates good predictive ability; on both components and on the overall responses. Most of the deviations from the prediction can be attributed to not strictly parallel subtests and estimated parameter values not being the best possible estimates.

DERIVATION AND TEST OF PREDICTIONS

OF A DISCRETE LATENT STATE MODEL

FOR SIGNED NUMBER ADDITION TEST PERFORMANCE

by

KENTARO YAMAMOTO

A thesis submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE
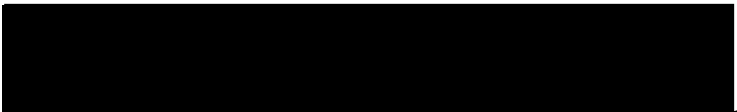
in

PSYCHOLOGY

Portland State University

1983

TO THE OFFICE OF GRADUATE STUDIES AND RESEARCH

The members of the Committee approve the thesis of Kentaro Yamamoto presented July 27, 1983.
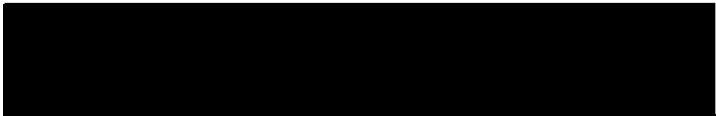
James A. Paulson, Chairman

Chadwick Karr

Craig A. Magwire

APPROVED:

David Wrench, Head, Department of Psychology

Stanley E. Rauch, Dean of Graduate Studies and Research

ACKNOWLEDGMENTS

It is with sincere appreciation that the author acknowledges the meticulous readings and insightful comments of members of the thesis committee, Dr. Chadwick Karr and Dr. Craig W. Magwire. Both were always available and willing to help the author in various stages of this research.

The author is particularly grateful to Dr. James A. Paulson who introduced the author to the field of Psychometrics, and inspired the author with the idea of this present study. Without Dr. Paulson's instructions, encouragements and friendship, this study could not have been undertaken.

The author would also like to express his thanks to Katy Peterson who skillfully and diligently typed the final copy of this thesis.

Special gratitude is due to my wife, Lisa who has supported the author with tolerance and understanding to complete this project.

TABLE OF CONTENTS

        PARAMETER FREE PREDICTION

            QUARTER TEST ABSOLUTE VALUE COMPONENT

            HALF TEST ABSOLUTE VALUE COMPONENT

            QUARTER TEST SIGN COMPONENT

            HALF TEST SIGN COMPONENT

        FORMULAS FOR PARAMETER SPECIFIC PREDICTIONS

            QUARTER TEST

            HALF TEST

        PARAMETER SPECIFIC PREDICTIONS

            QUARTER TEST ABSOLUTE COMPONENT

            HALF TEST ABSOLUTE COMPONENT

            QUARTER TEST SIGN COMPONENT

            HALF TEST SIGN COMPONENT

SYNTHESIS OF OVERALL RESPONSES FROM

COMPONENT RESPONSES

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Lord and Novick (1968) describe unidemensional latent trait theories which provide models to explain the relationships between unobservable real abilities (latent traits) and observed test scores. The relationships are expressed in the form of mathematical functions. The theory begins with the assumption that the unobservable hypothetical traits explain the most important aspect of the test performance. If a set of test items is fitted to a latent trait model and the item parameters are known, the estimation of each examinee's ability can be made on the single scale from any subset of items that have been fitted to the model. This type of estimation could never have been dreamed of in classical test theory. The theoretical advantages of latent trait theory have potentially far reaching practical implications for testing, particularly in the area of adaptive testing (Weiss, 1980). Despite these potentially beneficial characteristics, there are serious problems of application to real test items because of violation of unidimensionality. Most models, such as Rasch's logistic model and the two and three parameter

models, assume that one's probability of answering an item correctly increases as the latent trait value increases.

Latent trait models start by assuming items to be unidimensional. When this assumption is violated, then the model has to be adjusted. Even though it has been known for a long time that test items are very likely to be multidimensinal, a sound adjustment procedure has not been known. Attempts have been made to understand the interaction of dimensionality and latent trait models. Hambleton and Traub (1973) dealt with this problem by clustering items using results from factor analysis. Hence each cluster contained homogeneous items. This would work fine if all items in a subset are unidimensional and multidimensionality emerges only when they are combined together. But in many cases the responses to the items are the results of multidimensional latent traits. Also by clustering items into several subsets, not enough items are contained in each set to calibrate the parameter values successfully (Bejar, 1980). The use of common factor analysis for multidimensional items was discussed by McDonald (1981, 1982). Adjustment of unidimensional models to multidimensional data was also discussed by Reckase and McKinley (1982). Both of these approaches are still in a very early stage of development, and the field as a whole has not yet found a general method to deal with the problem of multidimensionality.

In recent developments in the assessment of mastery Brown and Burton (1978), and also by Birenbaum and Tatsuoka (1980), express concern with identification of misconceptions which would produce patterns of systematic errors and right answers. The main objectives of both studies are oriented toward computer algorithms, as opposed to more traditional measurement concepts. Brown and Burton call misconceptions procedural bugs. The process of deriving answers is divided into several steps based on special skills.

Misconceptions produce the incorrect implementation of these skills. On many mathematical problems, it is possible to get many right answers with an erroneous procedure. Also procedural bugs produce characteristic patterns of errors and correct responses. For example, the student with the bug that subtracts the small digit from the large digit in each column, regardless of which is on top, would answer incorrectly on the item, "62-34." But the same student can answer correctly on the item "64-32." The same rule was applied to both items but the results were not compatible.

Brown and Burton's model performed successfully with data obtained from fourth, fifth, and sixth grade Nicaraguan students. This study also found that about 40% of students had procedural bugs, but many bugs were not consistent. One advantage of this model is that a few bugs can account for entire patterns of errors for many

students.   Students with procedural bugs are not making random errors but are making responses according to a rule that happens to be incorrect.

One of the problems this model has is that it is a deterministic model; because of this it cannot incorporate any random errors. It is likely that some patterns created by the random errors will be classified into a particular bug. Theoretically there can be as many bugs to account for error patterns as the number of errors. But the majority of these so called bugs would not be plausible as rules students are following. One quickly reaches a point of diminishing returns where one has to hypothesize a particular bug for each student.

The discrete latent-state approach by Paulson (1982) is similar to Tatsuoka and Birenbaum's model (1979) with respect to erroneous rules in arithmetic achievement tests. Unlike Tatsuoka and Birenbaum's model, Paulson's model provides room for excecution errors for anyone in any state and the use of a few discrete states rather than an indefinite number of possible states.

Paulson devised his model to account for signed-number arithmetic data collected by Tatsuoka and Birenbaum. The data of concern in the proposed study is the collection of responses made on the problems of addition of two different one or two digit signed numbers. The model analyzes responses from two perspectives, as was done by Tatsuoka and

Birenbaum: the sign component response and the absolute value component response. States are described in terms of the rule the student appears to follow in making a component response. With respect to the sign component, the possible states which any student can be in are 1) correct rule, 2) always positive, 3) always negative, 4) sign of first, 5) sign of second, and 6) random guessing. With respect to the absolute value component the possible states are 1) correct rule, 2) always adds, 3) always subtracts, and 4) random guessing.

The sources of errors are three: guessing errors, systematic errors, and execution errors. Being in the correct rule state should produce a perfect score, except for occasional execution errors. In the random guessing state the pattern of errors should be random, with possibly different probabilities of correct guessing for different types of items. In states corresponding to systematic misconceptions, errors can be due to the misconception, or to an error of execution causing a mistake on an item which would have otherwise been correct. Paulson's model was very successful in reproducing the means, variances, and covariances for the different item types in the data collected by Tatsuoka and Birenbaum. The good performance of the model with real data shows that the systematic states in the model are accountable for a good part of the variability we see in the data.

The performance of the model in terms of descriptive statistics does not ensure that other detailed predictions of the model are accurate. This study tests specific predictions of Paulson's model with real data. The model predicts subtests to be strictly parallel in classical sense; that is, means and variances of all parallel subtests ought to be equal, and correlations between any new pairs of subtests also ought to be equal. This prediction of strong parallelism holds for any model which implies that items within types are equivalent, provided the characterization of items suggested by Tatsuoka is used. If parallelism among the four subtests exists on all statistics mentioned before, then the characterization of items used is supported. If Paulson's model is to be held true without further modification, parallelism of subtests must be found. The theoretical predictions of means, standard deviations of subtests, and correlation between subtests will be derived from Paulson's model and compared with the data collected by Tatsuoka.

Two ways to assess reliability of measurement will be considered. Paulson's model predicts how the reliability indices ought to come out. One index is the correlation coefficient between parallel subtests already mentioned. Another is the decision consistency. In classical test theory, the reliability of measurement can be viewed as the correlation between parallel tests, or between two

measurements obtained by using the same test on two different occasions. In criterion-referenced tests, we frequently find the range of test scores to be small. This restricted range causes the correlation coefficients between parallel tests to be small. Thus, the correlational estimates of reliability and validity will be low under these circumstances (Lord & Novick, 1968). A more appropriate way to assess reliability of measurement is the decision consistency. Decision consistency refers to the extent to which decisions on classification agree across two measurements. The comparison of observed extent of consistent decision with the theoretical prediction of consistent decision on mastery will be made. At the appropriate cut-off point, the decison consistency is expected to provide more useful information on the reliability of a test than the correlation coefficients between the tests. Both outcomes should be closely related.

CHAPTER II

METHOD

The data used in this study were collected by Tatsuoka
and Birenbaum (1979) in January 1979 at Urbana Junior High
School. The data include 64 responses made on a 64
open-ended item test, consisting of 16 tasks of four
parallel items each in addition and subtraction of one or
two digit signed integers. One hundred twenty nine junior
high school students who had just started studying signed
numbers took this test. This study utilizes a subtest of
five item types in addition; each type consists of four
parallel items, so there are 20 items in all. All 20 items
are presented in Table I. The five item types are; 1) large
positive plus small negative, 2) small negative plus large
positive, 3) large negative plus small negative, 4) small
positive plus large negative, 5) large negative plus small
positive. Therefore the four subtests, each consisting of
one item from each of five item types, are parallel.

When the entire test is divided into 4 subtests each
containing 5 item types of one item each, then the subtest
is called quarter test, and identified by subscript q1 to
q4. The term half test is used for two quarter tests which

TABLE I

20 ITEMS OF SIGNED NUMBER ARITHMETIC TEST

Item type	Subtests			
	I	II	III	IV
L+-S	12+-3	7+-5	15+-6	4+-2
-S+L	-3+12	-1+10	-4+13	-2+11
-L+-S	-14+-5	-10+-1	-7+-5	-10+-8
S+-L	3+-5	2+-11	6+-8	1+-10
-L+S	-6+4	-5+3	-4+2	-8+6

have been combined into one. There can be 6 pairs of half tests made up out of 4 quarter tests. Therefore, for these 6 half tests the identities of quarter tests are expressed by the subscript. For example, a half test consists of q1 and q3 would be h13. These 6 half tests also should be parallel to each other since each contains 5 item types of 2 items each.

This study utilizes already estimated parameter values which came from previous efforts by Paulson (1982). In Table II estimated parameters of subject state probability, π_j, and conditional probability of correct response on item type i given state j, P_{ij}, of both absolute and sign components are presented.

The results of this study will be presented in the next section as follows. First, the tests of predictions of the model which do not depend on the parameter values will be given. Second, formulas will be devised for making more specific predictions which do depend on the parameters. Then the comparison of the data to these predictions will be made. For each kind of prediction, results for quarter tests and half test on the absolute value component will be given first, followed by corresponding results for quarter tests and half tests on the sign component. Finally, results concerning the synthesis of overall responses from component responses will be given.

TABLE II

ESTIMATED PARAMETER VALUES FOR BOTH COMPONENTS

ABSOLUTE VALUE COMPONENT

State	State probability	Conditional probability of correct response				
		1	2	3	4	5
1,correct rule	.202	.945	.945	.945	.945	.945
2,always add	.091	.143	.143	.857	.143	.143
3,always subtract	.253	.896	.896	.104	.896	.896
4,random errors	.454	.656	.434	.344	.502	.464

SIGN COMPONENT

State	State probability	Conditional probability of correct response				
		1	2	3	4	5
1,correct rule	.442	.920	.920	.920	.920	.920
2,always positive	.113	.904	.096	.096	.904	.096
3,always negative	.186	.128	.872	.872	.128	.872
4,sign of first	.039	.928	.928	.072	.072	.072
5,sign of second	.024	.069	.069	.931	.931	.069
6,random errors	.196	.805	.436	.895	.676	.583

CHAPTER III

RESULTS

PARAMETER FREE PREDICTIONS

Since each quarter test consists of one item each of five item types, each quarter test should be parallel to each other quarter test. This parallelism implies that all quarter tests should have equal means and variances. Also this equality should extend to the correlations of any pairs and decision consistencies of any pairs of quarter tests at a given cut-off point for mastery.

Quarter tests for absolute value component

Descriptive statistics for the absolute value component responses on quarter tests are presented in Table III. The t-tests for the difference of means of quarter tests found marginal significance at two pairs of comparisons, quarter test 1 against 3 and quarter test 2 against 3 at $p=.04$. The test of the significance of the difference between correlated variances (Ferguson, 1981) found no significant difference between any pairs of variances.

TABLE III

OBSERVED MEANS AND STANDARD DEVIATIONS
OF QUARTER TESTS

Quarter test	Mean	Standard deviation
1	3.209	1.434
2	3.163	1.520
3	2.961	1.465
4	3.109	1.470

Friedman two-way analysis of variance found that the variance due to the quarter tests was not significant, $\chi^2=2.256$, $\underline{p}=.521$. This test can not say anything about within item types. The quarter test scores may not differ each other, because quarter test scores are the combined scores of responses to five item types. Two-way repeated measures analysis of variance of item type by quarter test was performed, even though the assumption of normality of response value distribution is clearly violated for the data at this level of analysis since responses are either 0 or 1 for each item. The strength of association (Meyers, 1979) was calculated also. The two-way analysis of variance was done even though the data obviously violate the assumption of normality, in order to be critical about accepting prediction of parallelism to be true. The results are presented in Table IV. Even though the variance due to the quarter tests was significant, the proportion of variance due to the quarter tests was very small. These results imply the existence of unequal items in some item types. In order to locate the unequal items, item analysis within the type was conducted.

Cochran's Q test within item types showed the existence of inequality among "parallel" items in item type 1 and item type 2. Within each of these two item types, one specific item each which differs from the other three items in the same item type was found by McNemar's test. These 2

TABLE IV

TWO WAY REPEATED MEASURES ANALYSIS OF VARIANCE;
ABSOULUTE SIGN COMPONENT

Source	SS	df	MS	F	p
Subject	190.541	128			
Tests	.897	3	.299	3.636	<.025
Item type	21.442	4	5.361	26.498	<.001
Interaction	4.605	12	.384	2.723	<.005
Test x Type					
Error-test	31.578	384	.0822		
Error-type	140.317	512	.2023		
Error-interaction	217.161	1536	.1414		
Total	606.541	2579			

Strength of association

$$\omega^2 \text{ score} | \text{tests} = .0020$$
$$\omega^2 \text{ score} | \text{subjects} = .6549$$

items were item number 3 of item type 1 and item number 5 of item type 2. The results of Cochran's Q tests and McNemar's tests are presented in Table V, with the number of subjects passing each item in the two item types.

The results of the test of the equality of observed correlations and observed decision consistencies at three cut-off points are presented in Table VI and Table VII respectively. It should be noted that all correlations which involve quarter test 1 were consistantly lower than any other correlations. The lowest value of correlation was of quarter test 1 and 3; both have a discrepant item in different item types.

Half tests for absolute value component

There are six half tests which are all possible combinations of pairing 4 quarter tests. Observed means, standard deviations, and correlations of half tests are presented in Table VIII.

The t-tests on the difference of means between any two half tests found no significant difference. Also no significant differences were found between any two variances by the test for the significance of difference between two correlated variances (Ferguson, 1981).

The study on the absolute component found that presumably parallel tests were not exactly parallel. Within item type there were significant differences between items,

TABLE V

RESULTS OF COCHRAN'S Q TEST AND McNEMAR'S TEST:
ABSOLUTE VALUE COMPONENT

ITEM TYPE I

Item number	1	2	3	4
Item	12+-3	7+-5	15+-6	4+-2
Correct response	99	94	81	102
Incorrect response	30	35	48	27

$$Q = 15.636 \qquad (p<.001)$$

McNemar's test results

Item number	Item number		
	2	3	4
1	.456	7.225 (p<.007)	.114
2		4.645 (p<.031)	1.750
3			13.793 (p<.000)

ITEM TYPE II

Item number	5	6	7	8
Item	-3+12	-1+10	-4+13	-2+11
Correct response	96	90	75	79
Incorrect response	33	39	54	50

$$Q = 17.091 \qquad (p<.001)$$

McNemar's test results

Item number	Item number		
	6	7	8
5	.893	12.121 (p<.000)	6.919 (p<.009)
6		6.759 (p<.009)	2.564
7			.281

TABLE VI

OBSERVED CORRELATIONS OF QUARTER TESTS; ABSOLUTE VALUE COMPONENT

Correlation pair	Observed correlation	Correlation pair					
		1,2	1,3	1,4	2,3	2,4	3,4
1,2	.611				*	*	
1,3	.543				*	*	**
1,4	.634						
2,3	.719						
2,4	.723						
3,4	.713						

$$*; \underline{p} < .05$$
$$**; \underline{p} < .01$$

TABLE VII

OBSERVED DECISION CONSISTENCIES OF QUARTER TESTS; ABSOLUTE VALUE COMPONENT

Cut-off point	Decision consistency pair					
	1,2	1,3	1,4	2,3	2,4	3,4
5	.380	.440	.405	.727	.659	.765
4	.535	.415	.490	.540	.552	.579
3	.315	.440	.472	.582	.703	.598

Consistency pair	Significantly different pair					
	1,2	1,3	1,4	2,3	2,4	3,4
1,2			*3	**5 **3	**5 **3	**5 **3
1,3				*5	*4 **3	*5 *4
1,4				**5	**5 **3	**5
2,3						
2,4						
3,4						

*; $p < .05$
**; $p < .01$

Note: The number following the asterisks indicates the cut-off points at which the decision consistencies differ.

TABLE VIII

OBSERVED DESCRIPTIVE STATISTICS OF HALF TESTS;
ABSOLUTE VALUE COMPONENT

Half test	Mean	Standard deviation
h12	6.372	2.651
h34	6.070	2.717
h13	6.171	2.548
h24	6.271	2.775
h14	6.318	2.625
h23	6.124	2.768

Half test I	Half test II	Correlation
h12	h34	.790
h13	h24	.821
h14	h23	.774

that is to say that the item type does not provide complete characterization of items, therefore we may require finer characterizations of item qualities. But the strength of association between subtests and the dependent variable shows that the proportion of variance due to the subtests is very small compared to the between subjects variance.

Quarter tests for sign component

Observed means and variances of the sign component responses on the four quarter tests are presented in Table IX. The t-tests for the difference of means of quarter tests found significance at two pairs of comparisons, quarter test 1 against 4 at the P =.001 level and quarter test 2 against 4 at the P =.05 level. No significant difference was found between any pairs of observed variances using the test on the significance of the difference between correlated variances.

A Friedman two-way analysis of variance found that the variance due to the subjects was not significant, χ^2=5.128, p=.163.

A two-way repeated measures analysis of variance test was performed and the strength of association of subtests with the dependent variable was calculated. The variance due to the quarter test was significant, but the strength of association index was very small. The results are in Table X.

TABLE IX

OBSERVED DESCRIPTIVE STATISTICS OF QUARTER TESTS;
SIGN COMPONENT

Quarter test	Mean	Standard deiviation
q1	3.806	1.132
q2	3.674	1.091
q3	3.628	1.146
q4	3.512	1.200

CORRELATION

	q2	q3	q4
q1	.594	.558	.631
q2		.590	.695
q3			.696

TABLE X

TWO WAY REPEATED MEASURES ANALYSIS OF VARIANCE;
SIGN COMPONENT

Source	SS	df	MS	F	p
Subject	96.422	128			
Tests	1.148	3	.383	3.984	<.010
Item type	4.889	4	1.222	3.126	<.025
Interaction	1.732	12	.144	1.333	<.200
Test x Type					
Error-test	37.357	384	.097		
Error-type	200.007	512	.391		
Errot-interaction	165.764	1536	.108		
Total	507.319	2579			

Strength of association

ω^2 score|tests   = .0064
ω^2 score|subjects = .6219

Cochran's Q test on the item types showed the significant inequalities in item types 1 and 4. Unlike absolute value component, the inequality emerged in the combinations, not because a single item differs from the other three, which was the case in absolute value component. In item type 1, McNemar's tests between pairs of items found significant differences between item 1 and item 2 and item 1 and item 4 at $p=.05$. In item type 4, a significantly different pair was item 13 and item 15 at $p=.05$. No other pairs were found significantly different. The results of Cochran's Q tests and McNemar's tests are presented in Table XI.

It should be noted that even though item type 1 was found to have non parallel items in both components, the unequal item in the absolute value component was not significantly different from any of the other three items in the sign component. A discrepant item in one component is not necessarily discrepant item in the other component.

The results of the rest of the equality of observed correlation and decision consistencies at three cut-off points are presented in Table XII and Table XIII consecutively.

Half tests for sign component

The means, variances and correlations of six half tests are presented in the Table XIV.

TABLE XI

RESULTS OF COCHRAN'S Q TEST AND McNEMAR'S TEST;
SIGN COMPONENT

ITEM TYPE I

Item number	1	2	3	4
Item	12+-3	7+-5	15+-6	4+-2
Correct response	103	89	94	90
Incorrect response	26	40	35	39

$$Q = 8.512 \qquad (\underline{p}<.037)$$

McNemar's test results

Item number	Item number		
	2	3	4
1	5.633 ($\underline{p}$<.018)	2.065	4.114 ($\underline{p}$<.043)
2		.640	0.
3			.450

ITEM TYPE IV

Item number	13	14	15	16
Item	3+-5	2+-11	6+-8	1+-10
Correct response	100	94	85	92
Incorrect response	29	35	44	37

$$Q = 8.345 \qquad (\underline{p}<.039)$$

McNemar's test results

Item number	Item number		
	14	15	16
13	.893	5.939 ($\underline{p}$<.015)	1.531
14		2.370	1.042
15			1.714

TABLE XII

OBSERVED CORRELATION OF QUARTER TESTS;
SIGN COMPONENT

Correlation pair

Correlation pair	Observed correlation	1,2	1,3	1,4	2,3	2,4	3,4
1,2	.594						
1,3	.558					*	*
1,4	.631						
2,3	.590					*	*
2,4	.695						
3,4	.696						

*; significantly different at $p < .05$

TABLE XIII

OBSERVED DECISION CONSISTENCIES OF QUARTER TESTS;
SIGN COMPONENT

Cut-off point	Decision consistency pair					
	1,2	1,3	1,4	2,3	2,4	3,4
5	.472	.547	.628	.428	.538	.584
4	.477	.494	.514	.598	.646	.674
3	.424	.439	.473	.360	.488	.525

Consistency pair	Significantly different pair					
	1,2	1,3	1,4	2,3	2,4	3,4
1,2				*4	**4	**4
1,3					**4	**4
1,4				**5		*4
2,3						*5
					*3	*3
2,4						
3,4						

*; $p < .05$
**; $p < .01$

Note: The number following the asterisks indicates the cut-off points at which the decision consistencies differ.

TABLE XIV

OBSERVED DESCRIPTIVE STATISTICS OF HALF TEST; SIGN COMPONENT

Half test	Mean	Standard deviation
h12	7.481	1.985
h34	7.140	2.161
h13	7.434	2.011
h24	7.186	2.109
h14	7.318	2.106
h23	7.302	1.995

Half test I	Half test II	Correlation
h12	h34	.753
h13	h24	.775
h14	h23	.791

The $\underline{t}$-tests on the differences of means between two half tests found significant differences between half test 12 and half test 34, and also between half test 13 and half test 24. There were no significant differences between any two variances by the test for the significance of difference between two correlated variances.

FORMULAS FOR PARAMETER SPECIFIC PREDICTIONS

It is necessary to generate specific values of unconditional means, variances, correlations, and decison consistencies to examine the predictive ability of Paulson's model against observed statistics. The following formulas utilize the already estimated values for the parameters; proportions of subjects in states, πj, and conditional probabilities of correct response on item type i given state j, P_{ij}. The model assumes that the response of a student in state j to item type i is independent of the response to item type i'. This is referred to as conditional independence in this study. Under this assumption of conditional independence the responses may be thought of as independent Bernouli trials having parameters P_{ij}.

Quarter test

Let j represent state, $E(q|j)$ the conditional expectation of quarter test score given state j, and $var(q|j)$ and $cov(q,q'|j)$ the conditional variance of a quarter test q and covariance of quarter tests q and q' respectively. Let $E_j(\cdot)$ denote the expectation operator, averaging over states, and $var_j(\cdot)$ and $cov_j(\cdot)$ the corresponding variance and covariance operators with respect to the distribution over states. Then the unconditinal means, variances, and covaiances are given by

$$E[q] = E_j[E(q|j)] \tag{1}$$

$$Var(q) = E_j[var(q|j)] + Var_j[E(q|j)] \tag{2}$$

$$Cov(q,q') = E_j[cov(q,q'|j)] + cov_j[E(q|j),E(q'|j)] \tag{3}$$

The conditional expectation of a quarter test score, given the state, is the sum of conditional probabilities of correct response to item types one through five, given the state. Therefore, the unconditional mean of quarter test score is the sum of conditional expectations of quarter test scores weighted with respect to state probability. Let π_j

be the proportion of subjects in state j, P_{ij} the conditional probability of correct response on item type i given state j, and M the number of states on the component under study. Recall that responses are being analyzed in two perspectives, absolute value component responses and sign component responses. In the absolute value component, M is 4, and in the sign component M is 6. The equation (1) for the unconditional mean of the quarter test score can be rewritten as

$$\text{Mean}(q) = E[q] = \sum_{j=1}^{M} \pi_j \cdot \sum_{i=1}^{5} P_{ij}$$

The formula (2) for unconditional variance can be understood as the sum of the average of the variances within subject states weighted with respect to the state probabilities and the variance of the conditional expectations of score, given subject states. By using elementary statistical knowledge, $\text{var}_j E([q|j])$ can be decomposed into two parts, then the formula (2) can be written as

$$\text{Var}(q) = E_j[\text{var}(q|j)] + E_j[(E[q|j])^2] - \{E_j[E(q|j)]\}^2$$

Notice that the quarter test score of a subject consists of five conditionally independent Bernouli trials with parameters P_{ij}, then $\text{var}(q|j)$ is the sum of variances

of each five Bernoulli distributions. Then the unconditional variance of quarter test is given by

$$
Var(q) = \sum_{j=1}^{M} \pi_j \cdot \sum_{i=1}^{5} P_{ij} \cdot (1-P_{ij}) + \sum_{j=i}^{M} \pi_j \cdot (\sum_{i=1}^{5} P_{ij})^2
$$
$$
- (\sum_{j=1}^{M} \sum_{i=1}^{5} \pi_j \cdot P_{ij})^2
$$

On the formula of covariance (3), because q and q' are conditionally independent scores, given the state of the subject, the average of the conditional covariance is O. For the same reason, the covariance of conditional expected scores over states is the variance of conditional expected scores over states. Then (3) can be written as

$$
Cov(q,q') = O + Cov_j(E[q|j], E[q'|j])
$$
$$
= Var_j(E[q|j])
$$
$$
= E_j[(E[q|j])^2] - (E_j[E[q|j]])^2
$$
$$
= \sum_{j=1}^{M} \pi_j \cdot (\sum_{i=1}^{5} P_{ij})^2 - (\sum_{j=1}^{M} \sum_{i=1}^{5} \pi_j \cdot P_{ij})^2
$$

The correlation of quarter tests q and q' is given by

$$
\rho_{q,q'} = \frac{Cov(q,q')}{Var(q)}
$$

Decision consistency of mastery between two quarter tests q and q' can be expressed in terms of phi coefficients for a 2x2 table.

Subtest A	Subtest B	
	Nonmaster	Master
Master	a	b
Nonmaster	c	d

The phi coefficient of consistency is given by

$$\phi = \frac{bc - ad}{\sqrt{(a+b)\cdot(b+d)\cdot(c+d)\cdot(a+c)}}$$

This table can be thought of as the weighted sum of the probabilities of 2x2 classification, given states. The conditional independence of q and q' for a given state implies that the probabilities of 2x2 classification can be obtained by multiplying the two conditional probabilities. Then values for a, b, c and d can be calculated using following formula.

$$a = d = \sum_{j=1}^{M} \pi_j \cdot P_j(k) \cdot [1-P_j(k)]$$

$$b = \sum_{j=1}^{M} \pi_j \cdot P_j(k)^2$$

$$c = \sum_{j=1}^{M} \pi_j \cdot [1-P_j(k)]^2$$

Where $P_j(k)$ is the probability of scoring equal to or above the cut-off point k given state j. Since P_{ij}'s are the probabilities of conditionally independent trials, $P_j(k)$ is the sum of probabilities of producing scores equal to or above specified cut-off point k given state j. The probability of making correct responses on all five item types given state j, $P_j(5)$, is the product of conditional probabilities. The probability of making four or more correct responses, $P_j(4)$ is the sum of $P_j(5)$ and probabilities of making exactly one error out of five items, and there are five such possibilities. In addition to $P_j(4)$, $P_j(3)$ includes the probability of making exactly two errors out of five items, and there are 10 such combinations of errors. The $P_j(k)$ at various cut-off points can be obtained by using following formulas.

$$P_j(5) = \prod_{i=1}^{5} P_{ij} \qquad \text{for} \quad k=5$$

$$P_j(4) = P_j(5) + \prod_{i=1}^{5} P_{ij} \cdot \sum_{i=1}^{5} \frac{(1-P_{ij})}{P_{ij}} \qquad \text{for} \quad k=4$$

$$P_j(3) = P_j(4) + \prod_{i=1}^{5} P_{ij} \cdot \sum_{i=1}^{4} \sum_{i'=i+1}^{5} \frac{(1-P_{ij}) \cdot (1-P_{i'j})}{P_{ij} \cdot P_{i'j}}$$

$$\text{for} \quad k=3$$

Decision accuracy is the extent of accurate classification given two states, correct rule state and all other states, namely mastery and nonmastery states. In order to calculate decision accuracy, proportions of following 2x2 table need to be determined.

Given state	Classification	
	Nonmaster	Master
Master	a	b
Nonmaster	c	d

Recall that $P_j(k)$ is the probability of scoring equal to or above the cut-off point k given state j. Then the values for a, b, c and d can be obtained through these formulas.

$$a = \pi_1 - b$$

$$b = \pi_1 \cdot P_1(k) = \pi_1 \cdot \sum_{x=k}^{5} \binom{5}{x} \cdot P_{i1}^{x} \cdot (1-P_{i1})^{5-x}$$

$$c = \sum_{j=2}^{M} \pi_j - d$$

$$d = \sum_{j=2}^{M} \pi_j \cdot P_j(k)$$

Then the phi coefficient for accuracy is given by

$$\phi = \frac{bc - ad}{\sqrt{(a+b) \cdot (b+d) \cdot (c+d) \cdot (a+c)}}$$

Recall that the conditional probability of correct response on item i given correct rule state, P_{i1}, is constant for all item types. This constant conditional probability of P_{i1} justifies the binominal form $P_1(k)$.

Half test

The predicted mean of the half test that consists of quarter test 1 and quarter test 2 is given by

$$\text{Mean}(h_{12}) = 2 \cdot \text{Mean}(q)$$

The predicted variance of a half test is given by

$$\text{Var}(h_{12}) = 2 \cdot (1 + \rho_{q,q'}) \cdot \text{Var}(q)$$

The predicted correlation between half test 12 and half test 34 is given by the Spearman-Brown formula.

$$\rho_{h12, \ h34} = \frac{2 \cdot \rho_{q,q'}}{1 + \rho_{q,q'}}$$

The decision consistency between two half tests is analogous to the one for quarter tests, and it is expressed by phi coefficients also. The probability P of scoring equal to or above the cut-off point k for the half test h given state j, $P_{hj}(k)$ can be obtained by summing the multiplications of all combinations that produce specified scores. The values for the $P_{hj}(k)$ are given by

$$P_{hj}(10) = [P_j(5)]^2$$

$$P_{hj}(9) = Ph_j(10) + 2 \cdot P_j(4) \cdot P_j(5)$$

$$P_{hj}(8) = Ph_j(9) + [P_j(4)]^2 + 2 \cdot P_j(3) \cdot P_j(5)$$

$$P_{hj}(7) = Ph_j(8) + 2 \cdot P_j(2) \cdot P_j(5) + 2 \cdot P_j(3) \cdot P_j(4)$$

$$P_{hj}(6) = Ph_j(7) + 2 \cdot P_j(1) \cdot P_j(5) + 2 \cdot P_j(2) \cdot P_j(4) + P_j(3)^2$$

The predicted phi coefficient for decision consistency at cut-off point of k can be obtained by

$$\phi = \frac{bc - ad}{\sqrt{(a+b) \cdot (b+d) \cdot (c+d) \cdot (a+c)}}$$

where

$$a = d = \sum_{j=1}^{M} \pi_j \cdot P_{hj}(k) \cdot [1-P_{hj}(k)]$$

$$b = \sum_{j=1}^{M} \pi_j \cdot [P_{hj}(k)]^2$$

$$c = \sum_{j=1}^{M} \pi_j \cdot [1-P_{hj}(k)]^2$$

The formula for predicted decision accuracy of half tests is also analgous to the one for quarter tests. The Phi coefficient for decision accuracy can be obtained by using the same formula for decision consistency, except the values for a, b, c and d are given by

$$a = \pi_1 \cdot [1-P_{h1}(k)]$$

$$b = \pi_1 \cdot P_{h1}(k)$$

$$c = \sum_{j=2}^{M} \pi_j \cdot [1-P_{hj}(k)]$$

$$d = \sum_{j=2}^{M} \pi_j \cdot P_{hj}(k)$$

PARAMETER SPECIFIC PREDICTIONS

Quarter tests for absolute value component

By using the previously presented formulas the predicted values for the means and standard deviations of quarter test scores are presented in the Table XV.

No significant difference was found by the t-test between any of the quarter test means and the predicted mean. Chi-square tests found no significant difference between any quarter test variances and the predicted variance. The results of the comparison between observed correlations and the predicted correlation are presented in the Table XVI. All the correlations except those involving the quarter test 1 exceeded the predicted value of .588.

The decision consistencies for various cut-offs are presented in the Table XVII. At the cut-off point of 4, agreement between observed and predicted decision consistencies was consistently the best, and none of the deviations was found to be significantly different. When the cut-off point was set at 5, which is the perfect score and often used as the criterion for mastery, the agreement was the lowest.

The decison accuracy was computed only to compare with the predicted decision consistency, since actual decison accuracy is not observable. Predicted values for both

TABLE XV

PREDICTED AND OBSERVED MEANS AND STANDARD DEVIATIONS
OF QUARTER TESTS; ABSOLUTE VALUE COMPONENT

Statistic	Predicted	Observed			
		q1	q2	q3	q4
Mean	3.107	3.209	3.163	2.961	3.109
Standard deviation	1.366	1.433	1.520	1.465	1.470

TABLE XVI

PREDICTED AND OBSERVED CORRELATIONS FOR QUARTER TESTS;
ABSOLUTE VALUE COMPONENT

Predicted correlation   .588

Observed correlation

	q2	q3	q4
q1	6.11	.543	.634
q2		.719**	.723**
q3			.713**

*; significantly different from predicted value at $\underline{p}$<.05

**; significantly different from predicted value at $\underline{p}$<.01

TABLE XVII

PREDICTED AND OBSERVED DECISION CONSISTENCY
OF QUARTER TESTS; ABSOLUTE VALUE COMPONENT

Cut-off point	Predicted consistency	Observed consistency pair					
		1,2	1,3	1,4	2,3	2,4	3,4
5	.570	.380**	.440*	.405*	.727**	.659	.765**
4	.510	.535	.415	.490	.540	.552	.579
3	.478	.315*	.440	.472	.582	.703**	.598

*; $\underline{p} < .05$

**; $\underline{p} < .01$

indices are presented in Table XVIII. Both change monotonically over three cut-off points, but the rate of change was much greater for decision accuracy. Neither index was consistantly larger than the other for all cut-off points. The accuracy index is sometimes larger and sometimessmaller than the consistency index.

Half test for absolute value component

Predicted values and observed means, standard deviations, and correlations of the half tests on the absolute value component are presented in Table XIX.

No significant difference was found between any half test means and predicted mean by t-test. The half test which consist of the quarter tests 2 and 4, and the half test of the quarter tests 2 and 3 were found to have variances which significantly differ from the predicted variance.

The results of the comparisons between predicted and observed values on decision consistency are presented in Table XX and Figure 1. The highest values of the decision consistency, both predicted and observed were obtained at a cut-off point of 9, instead of a perfect score of 10. The second highest value which is nearly equal to the highest value was found at the cut-off point of 7. The lowest observed consistency was found at the cut-off point of 8 between two high consistencies at 7 and 9 (see Figure 1).

TABLE XVIII

PREDICTED DECISION CONSISTENCY AND ACCURACY OF
QUARTER TEST; ABSOLUTE VALUE COMPONENT

Cut-off point	Consistency	Accuarcy
5	.570	.716
4	.507	.524
3	.410	.287

TABLE XIX

PREDICTED AND OBSERVED DESCRIPTIVE STATISTICS
OF HALF TEST; ABSOLUTE VALUE COMPONENT

Statistic	Predicted	Observed					
		h12	h34	h13	h24	h14	h23
Mean	6.214	6.372	6.070	6.171	6.271	6.318	6.124
Standard deviation	2.434	2.651	2.717	2.548	2.775	2.625	2.768
Correlation	.741		.790		.821		.774

TABLE XX

PREDICTED DECISION ACCURACY AND CONSISTENCY AND
OBSERVED CONSISTENCY OF HALF TESTS;
ABSOLUTE VALUE COMPONENT

Cut-off point	Predicted		Observed consistency		
	Accuracy	Consistency	h12,h34	h13,h24	h14,h23
10	.709	.504	.579	.600	.588
9	.860	.749	.742	.739	.695
8	.673	.630	.438**	.536*	.481*
7	.529	.643	.675	.707	.612
6	.415	.546	.642	.656	.609

*;$\underline{p}<.05$
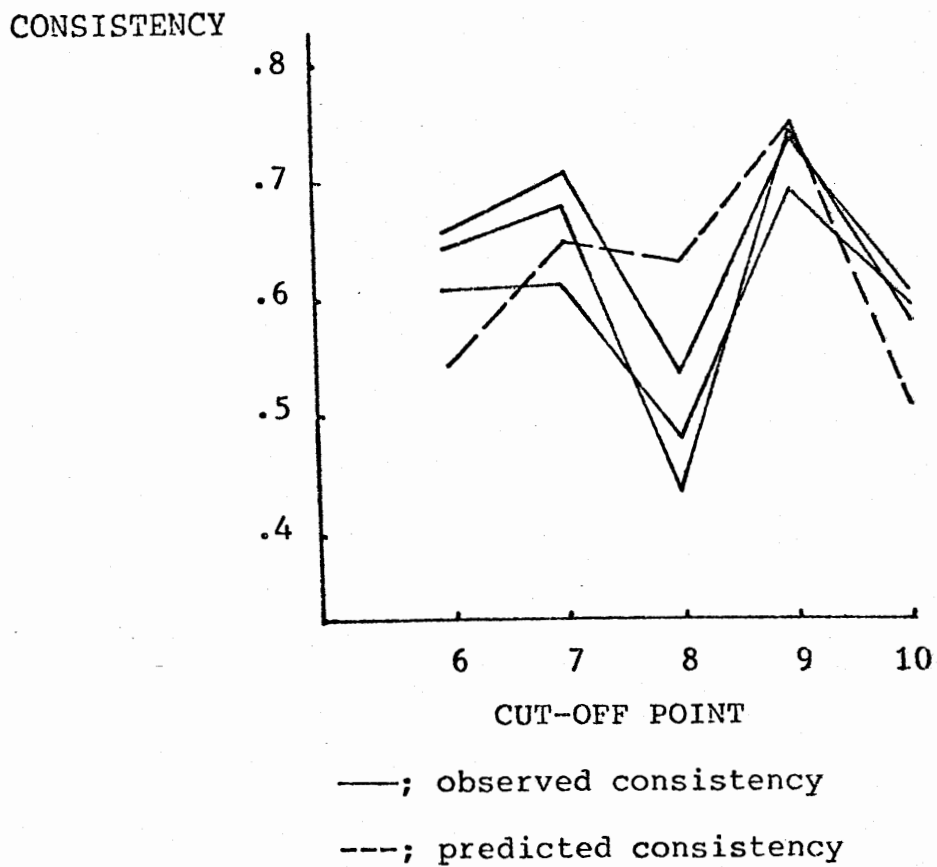**;$\underline{p}<.01$

Figure 1. Predicted and observed decision consistency for half tests: absolute value component.

The value of decision consistency does not change monotonically over various cut-off points.

The comparison between decision accuracy and consistency found that only one peak value is present in the accuracy index but two peak values in the consistency index. It also should be noted that the two indices are not related; there is no evidence of functional relationship between the two (see Figure 2).

Quarter tests for sign component

Predicted values for the means and standard deviations of the sign component quarter test are presented in Table XXI. There were no significant differences between predicted values and observed values of means, variances, and correlations. Predicted and observed correlations of half tests are presented in Table XXII.

Predicted decision accuracy and predicted and observed decision consistencies are presented in Table XXIII. At cut-off point of 4, agreement between predicted and observed decision consistency was best. Unlike the absolute value component, for the sign component the highest values of decision consistency were at the cut-off point of 4. The highest value of decision accuracy was also at the cut-off point of 4.
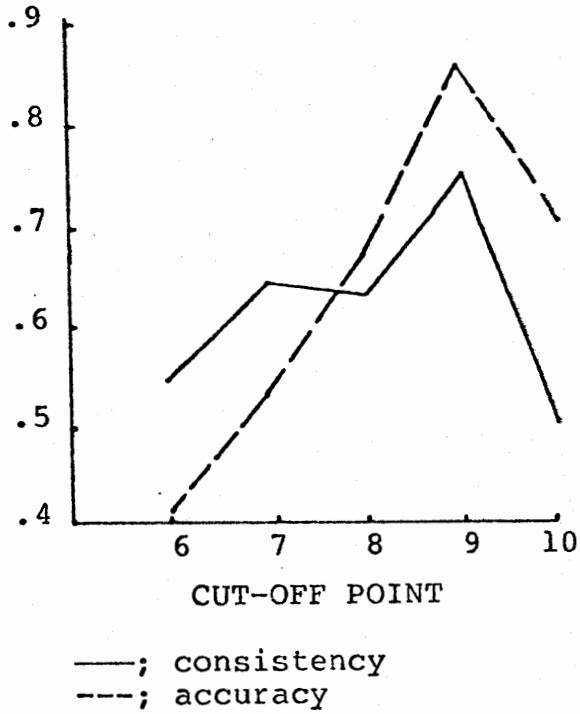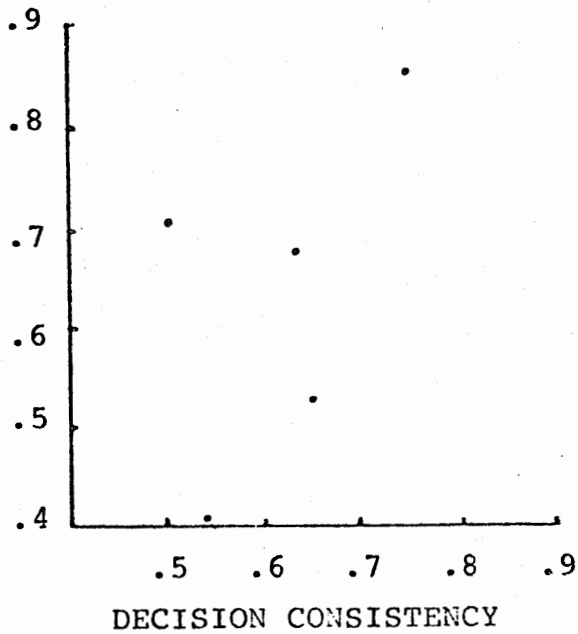
Figure 2. Predicted decision accuracy and consistency: absolute value component.

TABLE XXI

PREDICTED AND OBSERVED MEANS AND STANDARD DEVIATIONS
OF QUARTER TESTS; SIGN COMPONENT

Statistic	Predicted	Observed			
		q1	q2	q3	q4
Mean	3.634	3.806	3.674	3.628	3.512
Standard deviation	1.197	1.132	1.091	1.146	1.200

PREDICTED AND OBSERVED CORRELATIONS OF QUARTER TESTS;
SIGN COMPONENT

Predicted correlation   .612

Observed correlation

	q2	q3	q4
q1	.594	.588	.631
q2		.590	.695
q3			.696

None of correlations was found to be significantly
different from the predicted value.

TABLE XXIII

PREDICTED AND OBSERVED DECISION CONSISTENY AND
PREDICTED ACCURACY OF QUARTER TESTS;
SIGN COMPONENT

Cut-off point	Predicted		Observed consistency by quarter test pair					
	Accuracy	Consistency	1,2	1,3	1,4	2,3	2,4	3,4
5	.652	.434	.472	.547	.625**	.428	.538	.584*
4	.722	.582	.447	.494	.514	.598	.646	.674
3	.443	.409	.424	.439	.473	.360	.488	.525**

*;$\underline{p}<.05$
**;$\underline{p}<.01$

Half tests for sign component

Predicted and observed means, standard deviations and correlations of sign component half tests are presented in Table XXIV.

There were no significant differences between predicted values and observed values of means and variances. There was no significant difference between predicted and observed values of correlations.

Predicted and observed decision consistencies are presented in Table XXV with predicted decision accuracies. In Figure 3, the relationship between predicted and observed decision consistencies over various cut-off points is presented. Other than at the cut-off point of 10, there were no significant differences between predicted and observed values of consistency at all cut-off points. At the cut-off point of 10, predicted value was much lower than observed value.

In Figure 4, the relationship between predicted decision accuracy and consistency over various cut-off points is presented. Both indices have one peak value at the cut-off point of 8. The rate of change is smaller in consistency than accuracy. There is no linear relationship between accuarcy and consistency. In fact, the relationship is not even monotonic.

TABLE XXIV

PREDICTED AND OBSERVED DESCRIPTIVE STATISTICS OF
HALF TESTS; SIGN COMPONENT

Statistics	Predicted	Observed					
		h12	h34	h13	h24	h14	h23
Mean	7.267	7.481	7.140	7.434	7.186	7.318	7.302
Standard deviation	2.149	1.985	2.161	2.011	2.109	2.106	1.995
Correlation	.759		.753		.775		.791

TABLE XXV

PREDCITED DECISION ACCURACY AND CONSISTENCY AND OBSERVED
CONSISTENCY OF HALF TESTS; SIGN COMPONENT

Cut-off point	Predicted		Observed consistency		
	Accuracy	Consistency	h12,h34	h13,h24	h14,h24
10	.538	.289	.561**	.548**	.657**
9	.800	.634	.682	.659	.677
8	.867	.732	.651	.691	.706
7	.709	.631	.561	.571	.619
6	.528	.517	.401	.521	.505

**; $\underline{p} < .01$
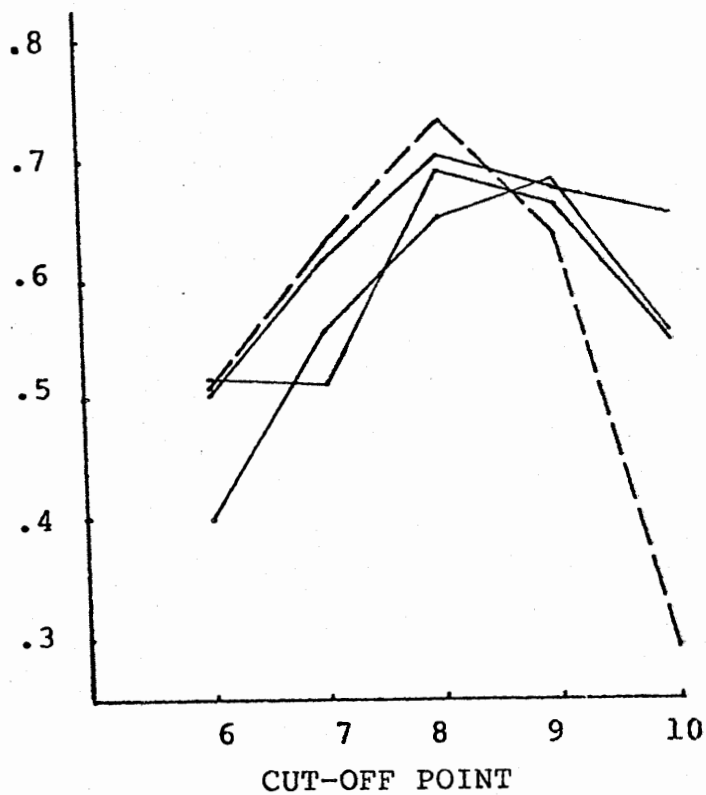
CONSISTENCY

Figure 3. Predicted and observed decision consistency for half tests: sign component.

TWO
INDICES

.9

.8

.7

.6

.5

.4

.3

6   7   8   9   10

———; consistency
---; accuracy

DECISION
ACCURACY

.9

.8

.7

.6

.5

.3   .4   .6   .7   .8   .9
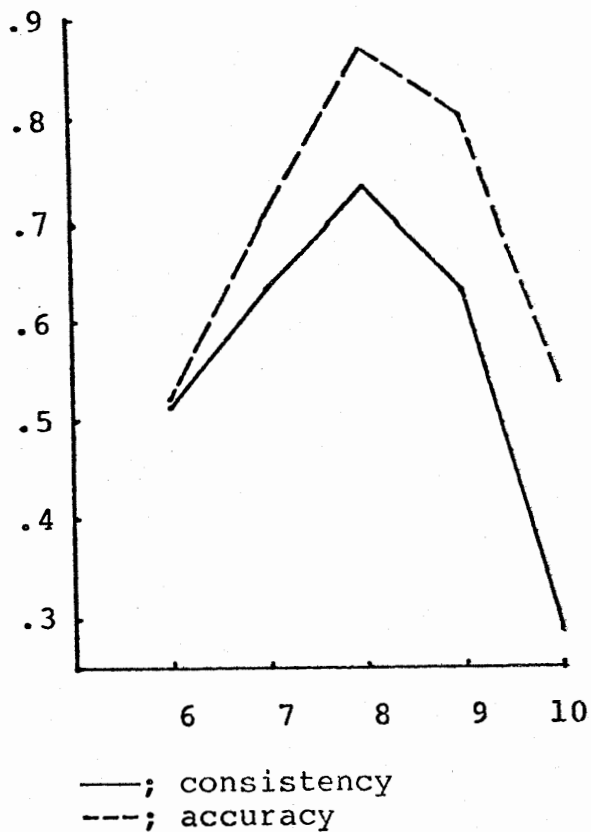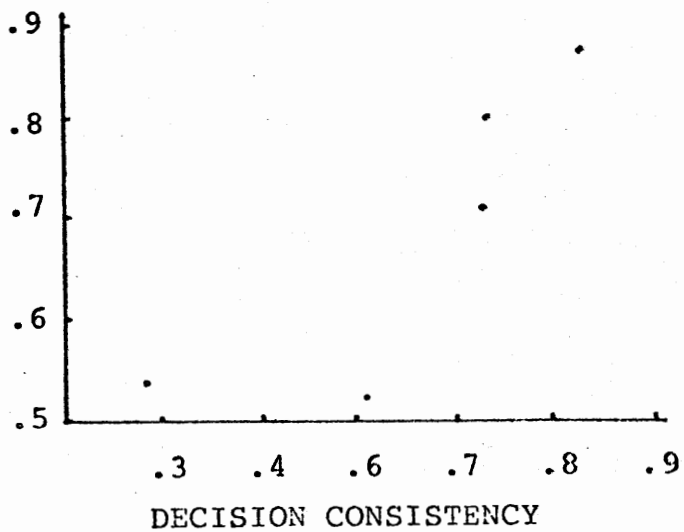
DECISION CONSISTENCY

Figure 4. Predicted decision accuracy and consistency:
sign component.

SYNTHESIS OF OVERALL RESPONSES FROM COMPONENT RESPONSES

For the model to be used in real situations, the model's ability to account for overall correct and incorrect responses is necessary. In practice, test responses are analyzed in terms of correctness and incorrectness without reference to the components.

Although a detailed discussion of the synthesis of component responses will not be made in this study, a preliminary assessment of the ability of the model to account for overall correctness and incorrectness of the responses will be attempted.

This assessment requires a model specifying the relationship between classifications on the two components, There are 24 overall states which represent combinations of four states of absolute value component and six states of sign component. A subject who is in state I of the absolute value component and in state J of the sign component will be classified as being in overall state I,J. Subscripts for the overall state identify the subject's states in two components (see Table XXVI). Having provided such a model which will be described below, assessment of its performance was made in terms of the closeness of predicted overall classification and observed overall classifications between two quarter tests. In mastery versus nonmastery

TABLE XXVI

OVERALL CORRECT RESPONSE STATES
FROM TWO COMPONENTS

Sign component

Absolute value component	1 correct rule	2 always positive	3 always negative	4 sign of first	5 sign of second	6 random error
1,correct rule	1,1	1,2	1,3	1,4	1,5	1,6
2,always add	2,1	2,2	2,3	2,4	2,5	2,6
3,always subtract	3,1	3,2	3,3	3,4	3,5	3,6
4,randon error	4,1	4,2	4,3	4,4	4,5	4,6

classification, the correct rule state is the mastery state, and the other systematic states and the random state are combined together and called the nonmastery state. Overall mastery state means the subject is in the mastery states of both components. Overall nonmastery state means the subject is in the nonmastery state in either one or both components. A finer diagnostic classification into overall mastery state, nonmastery in both components, and two combinations of mastery in one component and nonmastery in the other is also examined.

Predicted decision consistency and observed decision consistencies on the overall mastery versus nonmastery classification between two quarter tests are compared. Chi-square tests were also performed on the joint frequency distribution of the finer diagnostic classifications between two quarter tests, using predicted classifications as expected values.

Observed frequencies of the overall states classifications are presented in Table XXVII with the estimated frequencies. The procedure used to estimate frequencies of overall state classifications is discussed in the following section. After reviewing the observed 4x6 overall correct response state table, the two components were not completely independent. There were no subjects classified in correct rule in the absolute value component

TABLE XXVII

PREDICTED AND OBSERVED FREQUENCIES IN THE
OVERALL CORRECT RESPONSE STATES

Sign state

Absolute state	1	2	3	4	5	6
1	26.1 (26)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
2	3.5 (0)	1.7 (4)	2.7 (3)	.6 (1)	.4 (0)	2.7 (2)
3	9.8 (13)	4.6 (3)	7.6 (4)	1.6 (2)	1.0 (1)	8.0 (8)
4	17.6 (17)	8.3 (7)	13.7 (14)	2.9 (3)	1.8 (2)	14.4 (19)

A chi square test for the independence of systematic
and random error states of two components showed that
the two components are independent.

and classified as being in other than correct rule state in the sign component. However, when neither component has been mastered, classifications on the two components are independent. Therefore classifications on the two components are independent except for subjects who are in one of the correct rule states. The implication of this dependency between correct rule states of two components is that mastery of the absolute component does not occur unless mastery of the sign component is achieved first. The formulas that estimate the proportion of subjects in overall correct response states need to account for this dependency.

Let π_{IJ} be the proportion of subjects in the I state of the absolute component and also in J state of the sign component, π_{aI} the proportion of subjects in I state of the absolute component, π_{sJ} the proportion of subjects in J state of the sign component, and δ the constant accounting for the dependence of correct rule states of two components. The model for π_{IJ} which is tried is given by

$$\pi_{IJ} = \pi_{aI} \cdot \pi_{sI} + \delta \qquad \text{for} \quad I=1, J=1$$

$$= (\pi_{aI} - \frac{\delta}{1-\pi_{sI}}) \cdot \pi_{sJ} \qquad \text{for} \quad I=1, J>1$$

$$= \pi_{aI} \cdot (\pi_{sI} - \frac{\delta}{1-\pi_{aI}}) \qquad \text{for} \quad I>1, J=1$$

$$= \pi_{aI} \cdot \pi_{sJ} \cdot \left[1 + \frac{\delta}{(1-\pi_{aI}) \cdot (1-\pi_{sI})}\right] \qquad \text{for} \quad I>1, J>1$$

Estimation of δ can be made such that a trial value of δ produces a phi coefficient identical to the observed phi coefficient. Because this procedure produces values very close to 0 but negative for π_{IJ}'s (j>1), δ was set to give $\pi_{IJ}=0$ where j>1, that is

$$\hat{\delta} = \pi_{aI} \cdot (1 - \pi_{sI})$$

This yields $\pi_{Ij} = 0$ for j>1. Then estimated frequencies can be obtained by multiplying by the number of subjects, which is 129.

Let $P(a|I)$ be the conditional probability of producing an absolute component response pattern a, given subject state I of the absolute component, and $P(s|J)$ the conditional probability of producing sign component response pattern s, given subject state J of the sign component. Let r_{ia} and r_{is} be the scores of the theoretically most probable response on item type i given subject state a of the absolute component and s of the sign component, respectively. Since a quarter test contains five item types of one item each, the r_{ia}'s and the r_{is}'s are 0's or 1's. Two response pattern matrices, one giving r_{ia}'s and the other giving r_{is}'s are presented in Table XXVIII.

TABLE XXVIII

THEORETICALLY MOST PROBABLE RESPONSE PATTERNS
BY THE SUBJECTS IN A PARTICULAR STATE

ABSOLUTE VALUE COMPONENT

State	Item type				
	1	2	3	4	5
1,correct rule	1	1	1	1	1
2,always add	0	0	1	0	0
3,always subtract	1	1	0	1	1
4,random error					

SIGN COMPONENT

State	Item type				
	1	2	3	4	5
1,correct rule	1	1	1	1	1
2,always positive	1	0	0	1	0
3,always negative	0	1	1	0	1
4,sign of first	1	1	1	0	0
5,sign of second	0	0	1	1	0
6,random error					

The conditional response pattern probabilities $P(a|I)$ and $P(s|J)$ can be obtained by using conditional probabilities of the two separate components and the response pattern matrices (r_{ia}) and (r_{is}). Let P_{iI} be the conditional probability of correct response on item type i given state I of the absolute component and P_{iJ} the conditional probability on item type i given state J of the absolute component. Then the conditional probabilities of response patterns a and s, $P(a|I)$ and $P(s|J)$, can be calculated using the following formulas.

$$P(a|I) = \prod_{i=1}^{5} [1-r_{ia}+P_{iI} \cdot (2 \cdot r_{ia}-1)] \qquad \text{for} \quad I<4$$

$$P(a|I_r) = 1-\sum_{I=1}^{3} P(a|I) \qquad \text{for} \quad I_r=4$$

$$P(s|J) = \prod_{i=1}^{5} [1-r_{is}+P_{iJ} \cdot (2 \cdot r_{is}-1)] \qquad \text{for} \quad J<6$$

$$P(s|J_r) = 1-\sum_{J=1}^{5} P(s|J) \qquad \text{for} \quad J_r=6$$

Then the table of the entire 24 overall correct response state classification probability between two quarter tests can be obtained by the following formula.

$P(a_1, s_1; a_2, s_2)$

= P(overall correct response pattern a_1, s_1 on quarter test 1; overall correct response pattern a_2, s_2, on quarter test 2)

$$= \sum_{J=1}^{6} \sum_{I=1}^{4} \pi_{IJ} \cdot P(a_1 | I) \cdot P(s_1 | J) \cdot P(a_2 | I) \cdot P(s_2 | J)$$

This result can be thought as the weighted sum of probabilities of producing a_1, s_1 pattern and also a_2, s_2 pattern in respect to the overall correct response state. Since the conditional probability of response pattern given state is assumed as conditionally independent, the probability of producing a_1, s_1, and a_2, s_2 patterns is the product of four conditional response pattern probabilities given each component state.

Because the expected values in most cells of the overall correct response state calssification table are very small, instead of the chi-square test, Cohen's Kappa was used to compare the similarity of predicted and observed

decision consistency tables. The formula for Cohen's Kappa of predicted decision consistency table is

$$K = \frac{P_o - P_e}{1 - P_e}$$

Where

P_o = proportion of response state classification agreement

$$= \sum_{j=1}^{6} \sum_{i=1}^{4} P(a_i, s_j; a_i, s_j)$$

P_e = proportion of "expected" response state classification agreement

$$= \sum_{s_1=1}^{6} \sum_{a_1=1}^{4} [\sum_{s_2=1}^{6} \sum_{a_2=1}^{4} P(a_1, s_1; a_2, s_2)]^2$$

(Note: "expected" refers to the null hypothesis of independence which provides the baseline for chance agreement in Cohen's Kappa.)

The formula for Cohen's Kappa for the observed decision consistency table is

$$K = \frac{N_o - N_e}{N - N_e}$$

Where N_O is the observed frequency of classification agreement; N_e is the expected frequency of classification agreement; and N is the total number of subjects.

The predicted and observed Cohen's Kappa are presented in Table XXIX. The average of the observed values of .2845 seems very close to the predicted value of .2648. Evidently, extremely fine diagnostic classifications on the basis of the five item quarter test are not very consistent, but they are about as consistent as one would expect on the basis of the model.

The prediction concerning the consistency of overall mastery state decisions can be expressed by a phi coefficient based on a 2x2 table as was done in the previous single component study.

Test 1	Test 2	
	Overall nonmaster	Overall master
Overall master	a	b
Overall master	c	d

TABLE XXIX

PREDICTED AND OBSERVVED COHEN"S KAPPA FOR THE
AGREEMENT BETWEEN TWO QUARTER TESTS ON
CLASSIFICATIONS INTO THE 24 OVERALL
RESPONSE STATES

	Cohen's Kappa
Predicted	.2648

Observed
quarter tests
pair

q1,q2	.1893
q1,q3	.2078
q1,q4	.2713
q2,q3	.3079
q2,q4	.3050
q3,q4	.4257

The values for a, b, c, and d are the summation of appropriate overall correct response state classifications probabilities which have been derived already. The formulas can be expressed in terms of the $P(a_1,s_1;a_2,s_2)$ notation introduced above in following way.

$$a = d = \sum_{s_2=1}^{6} \sum_{a_2=1}^{4} P(1,1,; a_2,s_2)$$

$$b = P(1,1; 1,1)$$

$$c = 1-(a+d+b)$$

Then the phi coefficient is given by

$$\phi = \frac{bc-ad}{\sqrt{(a+b)\cdot(b+d)\cdot(c+d)\cdot(a+c)}} = \frac{bc-a^2}{(a+b)\cdot(a+c)}$$

The results of comparisons between observed and predicted overall correct response mastery decision consistencies are listed on Table XXX. Only one comparison was found significantly different from the predicted value.

Among the six possible pairs of the finer, four category diagnostic classifications based on quarter test, only one was not found to be significantly different from the predicted joint classification using a chi square test. Two observed diagnostic classifications between two quarter tests which were found to be closest to and the farthest from the predicted classification are presented in Table XXXI with predicted classifications.

TABLE XXX

OVERALL CORRECT RESPONSE MASTERY DECISION CONSISTENCY

Predicted consistency	Observed consistency pair					
	q1,q2	q1,q3	q1,q4	q2,q3	q2,q4	q3,q4
.405	.286	.502	.446	.494	.432	.656**

**;$\underline{p}<.01$

Note: Only one comparison was found significantly different from the predicted value.

TABLE XXXI

JOINT DISTRIBUTIONS OF DIAGNOSTIC CLASSIFICATIONS BASED ON TWO QUARTER TESTS

PREDICTED

Quarter test A	Quarter test B			
	Na,Ns	Na,Ms	Ma,Ns	Ma,Ms
Ma,Ms	1.42	2.55	3.34	6.45
Ma,Ns	5.25	1.42	1.82	3.34
Na,Ms	9.70	13.58	1.42	2.55
Na,Ms	59.79	9.70	5.25	1.42

OBSERVED

1) CLOSEST TO THE PREDICTED VALUE ($\chi^2_{15}=15.85$)

q1	q3			
	Na,Ns	Na,Ms	Ma,Ns	Ma,Ms
Ma,Ms	0	5	3	9
Ma,Ns	3	0	0	0
Na,Ms	7	13	2	3
Na,Ns	65	11	5	3

2) FARTHEST FROM THE PREDICTED VALUE ($\chi^2_{15}=33.74$)

q1	q2			
	Na,Ns	Na,Ms	Ma,Ns	Ma,Ms
Ma,Ms	4	9	2	7
Ma,Ns	2	1	2	3
Na,Ms	4	9	3	2
Na,Ns	64	11	3	3

Ma,Ms; mastery in both components
Ma,Ns; mastery in absolute value component and nonmastery in sign component
Na,Ms; nonmastery in absolute value component and mastery in sign component
Na,Ns; nonmastery in both components

The degrees of freedom for the chi square test is 15, since calculation of the predicted value used marginal frequencies of two components states and total frequencies from the entire length test, not from any single quarter test. Therefore, the chi square test at the quarter test level, the loss of degrees of freedom is only one due to the total number of subjects.

Even though the chi square test found a significant difference between observed and predicted values on five observations, two cases among them were only marginally significant $(.05 > p > .025)$. All three significantly different observed cases always had a significantly larger than predicted value in one particular cell, nonmastery in the absolute component and mastery in the sign component of one test and overall mastery state of the other test. Since the predicted value of that particular cell is fairly small, 2.55, the chi square value can be inflated greatly by the single cell contribution.

The assessment of predictive ability of Paulson's model on diagnostic classification between two quarter tests should not be made solely on these chi square test results, because four quarter tests being used are found to be not exactly parallel. When component classification is based on the quarter test results, the diagnostic classification is not dependable, because the magnitude of influence on the classification by one item is very large.

CHAPTER IV

DISCUSSION

Through item analysis, it is clear that quarter tests
are not strictly parallel. Inequality of items within item
types suggests that finer item characterization would be
necessary to equate items. The violation of the parameter
free prediction of parallelism has great significance for
any model which utilizes the present characterization of
item type as the essential information about items. This
inequality of items within an item type influences the tests
of parameter specific predictions, especially in correlation
and decision consistency predictions. The results of tests
on parameter specific predictions should be examined in
light of item analysis. In the absence of a better model,
Paulson's model should not be discarded even if we find some
minor errors between parameter specific predictions and
observed subtest statistics. Paulson's model was quite
successful in predicting means, variances, correlations, and
decision consistencies for both quarter tests and half tests
involving absolute value and sign components. The errors of
predicted values from observed statistics were relatively
consistent with results of the item analysis and the few

parameter values which are suspected not to be the best estimated values. Better parameter estimation might improve the predictive ability of the model greatly.

The reason why the predicted variances of quarter tests and half tests of absolute component are lower than the observed variances may be that the model does not account for the item differences within item type and the execution error rate differences between subjects within states. The estimated parameters of proportions of subjects in states, π_j's, and conditional probabilities of correct response given state, P_{ij}'s, both of which are used to calculate the predictions might not be the best values.

On the sign component, the large disagreement between observed and predicted decision consistencies of half test at the cut-off point of 10 indicates that the parameters associated with the correct rule state are not very accurate, for almost all subjects who score 10 in either one or both half tests are from the correct rule state. There are two possible reasons for the parameter values to be inaccurate. One is that the estimated conditional probability of correct response, given the "correct rule" state is too low, and the other is that the estimated state proportion is too large, so that there are some subjects from other states being included. Either of these factors, or both of them acting together, would lead to inconsistency for a strict cut-off at 10.

The large difference on the proportions of correct rule state in two components might be due to the difference in the number of possible responses or to the differences between the complexity of judgments of each component. In the absolute value component, for the subject to be classified as being in the correct rule state, the absolute value has to be correct and there are many possible values from which to choose. To obtain the correct absolute value, several judgments have to be made correctly, for example, whether the numbers should be added or subtracted, or borrowed from 10 or brought up to 10, and so on. On the other hand, the choice of responses is limited to only two, positive or negative, for the sign component. There is only one judgment on the sign component, which sign of one of the two numbers is to be chosen.

The changes in decision consistency over several cut-off points can provide a rationale for deciding on the cut-off point for making classification decisions. In practice, the passing score may be taken to maximize or nearly maximize the value of decision consistency. In such a case, if there are two high values and they are not next to each other, there is a problem to choose which one to use. In the absolute value component half test, there are two cut-off points, 7 and 9, where decision consistencies are higher than the rest, with one low value in between at 8. If the decision consistency is the only basis for

deciding the cut-off point, it is conceivable that either one could be chosen as a passing score. The meaning of the existence of two peak values should be closely examined.

In the absolute value component half test, at the cut-off point of 9, 80% of subjects in the correct rule state would score 9 or 10 on both tests. But 83% of subjects in the always subtract state and almost all of the subjects in other states would score 8 or less on both tests. Therefore, inconsistently classified subjects are very few. At the cut-off point of 7, there is another high decision consistency value, and the increase from cut-off point of 8 is substantial in observed consistency. The reason for this increase is, in addition to almost all of the subjects in "correct rule" state scoring above cut-off points in both tests, the majority of subjects in the always subtract state also would score above or equal to the cut-off point of 7, whereas subjects in this state would be almost evenly distributed over the four cells if the cut-off were at 8. Consequently, the decision consistency at the cut-off point of 7 is high but misleading for setting a passing score for mastery. Decision consistency should not be used by itself to set a passing score.

Although it is not possible to test the prediction of decision accuracy against data, there was a clear difference between predicted decision accuracy and consistency, especially in case of half tests. In the absolute value

component quarter tests, the rate of change is much greater in the accuracy index than in the consistency index. This may lead to two implications, if they are studied independently. The loss of consistency is only 11% when the cut-off point is lowered to 4 from 5, but the loss of accuracy is 26%. Therefore, the passing score could be set at 5 or 4 under the consistency index, but 5 may be more likely chosen under the accuracy index. The difference between accuracy and consistency indices is more pronounced in the absolute component half test.   There is only one peak value of accuracy at 9 and the index quickly decreases as the cut-off point is lowered further, in contrast to the two peak values of consistency and its slower rate of change.   The passing score would be decided differently under the two indices.

For the sign component, decision accuracy and consistency had the highest value at identical cut-off points for both quarter tests and half tests. Again, the rate of change was greater in accuracy than in consistency for both tests.

The relationship between decision accuracy and consistency, and their relationships to cut-off scores in this study showed disagreement with the results presented by Huynh (1980). Huynh found a strong linear relationship between the two indices. He also found that the accuracy index was always greater than the consistency index. The

present study found that the two indices are not functionally related at all, and at some cut-off points accuracy is less than consistency (see Figures 2,4). The lack of relationship between the two indices in both components is depicted in Figures 2b and 4b.

It should be pointed out that the notion of a true cut-off score in Huynh's study is to balance costs and benefits, and in this present study a true-score cut-off is expressed in terms of state categorization; mastery state versus other states. In Huynh's study the comparison between accuracy and consistency was made at the test scores where true cut-off score is reflected. The correspondence between the true-score cut-off used in the decision accuracy calculation and the observed score cut-off used in the decision consistency calculation is necessary to establish the relationship between the two indices. This restriction on Huynh's result is much greater than it seems. In practice, one of the common uses of decision consistency is to select a cut-off point where the decision consistency is nearly maximum, or maximum. This would not necessarily correspond with true cut-off scores in general. Therefore, it is vital information that the relationship between the two indices is not linear over various cut-off points and one index is not necessarily greater than the other at all points.

The restrictive conditions which should be examined for generalizing Huynh's results are that equivalent, unidimensional items be used and that cut-off scores be appropriately determined. That test items be equivalent was necessary because he used a beta-binominal model. That model produced the values for the accuracy index, therefore the accuracy index is model dependent. But observed consistency is not model dependent, and it is the information from the items. Therefore, the linear relationship between accuracy and consistency is dependent on the nature of the model being used. Since his model is also unidimensional, when items are multidimensional his results may not hold. This unidimensionality of items should not be assumed casually. The present study found that accuracy is not related to consistency, using items customarily thought of as unidimensional items. Without the presence of the multidimensional model, Huynh's results might have been used on these items, and the results would be misleading. Generalization of Huynh's results should be made carefully.

The parameter free prediction of the quarter tests as well as the half tests to be parallel was found to be not true, and finer categorization of items is probably required to obtain a model that would account for all data exactly. The performance of Paulson's model on parameter specific predictions was generally good in both components and

overall classification studies. Predicted values were close to observed statistics and most of the discrepancies found can be attributed to subtests which are not strictly parallel, and estimated parameter values not being the best. This present study shows the need to continue the study on decision accuracy and consistency of multidimensional items tests.

REFERENCES

Bejar, I. I. "A procedure for investigating the unidimensionality of achievement tests based on item parameter estimate," Journal of Educational Measurement, 283-296, 17, 1980.

Birenbaum, M. & Tatsuoka, K. K. "The use of information from wrong responses in measuring students' achievement," (Research Report 80-1), Urbana, Illinois: University of Illinois, computer-based Education Laboratory, February 1980.

Brown, J. S. & Burton, R. R. "Diagnostic models for procedural bugs in basic mathematical skills," Cognitive Science, 155-192, 2, 1978.

Ferguson, G. A. "Statistical analysis in Psychology Education," N.Y.: McGraw-Hill, 1981.

Hambleton, R. K., Swaminathan, H., Algina, J. & Coulson, D. B. "Criterion-referenced testing and measurement: A review of technical issues and developments," Review of Educational Research, 1-47, 48, 1978.

Hambleton, R. K. & Traub, R. E. "Analysis of empirical data using two logistic latent trait models," British Journal of Mathematical and Statistical Psychology, 105-211, 26, 1973.

Huynh, H. & Saunders, J. "Relationship between decision accuracy and decision consistency in mastery testing," Solution for Some Technical Problems in Domain-Referenced Mastery Testing. Final Report, University of South Carolina College of Education, South Carolina, 1980.

Lord, F. M. & Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

McDonald, R. P. "The dimensionality of tests and items," British Journal of Mathematical and Statistical Psychology, 100-117, 34, 1981.

McDonald, R. P. "Unidimensional and Multidimensional Models in Item Response Theory: A Factor-Analytic Perspective. A paper presented at 1982 IRT/CAT meeting, Minneapolis, 1982.

Meyers, J. L. Fundamentals of Experimental Design. Boston, N.J.: Allyn and Bacon Inc, 1979.

Paulson, J. A. "A discrete latent-state approach to diagnostic testing," Research proposal submitted to the Office of Naval Research Personnel and Training Research Program, Portland, Oregon: Portland State University, 1982.

Reckase, M. & McKinley, M. "Some Latent Trait Theory for the Multidimensional Latent Space," paper presented at 1982 IRT/CAT meeting, Minneapolis, 1982.

Tatsuoka, K. & Birenbaum, M. "The danger of relying solely on diagnostic adaptive testing when prior and subsequent instructional methods are different," Computer-based Education Research Laboratory Report E-5. University of Illinois, March 1979.

Weiss, D. J. (Ed.) Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis, Minnesota: University of Minnesota, 1980.