AUTHOR Powell, J. C.
TITLE Can Developmental Status Information Be Obtained From Wrong Answers?
INSTITUTION Windsor Univ. (Ontario).
SPONS AGENCY International Business Machines Corp., Armonk, N.Y.; Ontario Educational Research Council, Toronto.
PUB DATE Mar 79
NOTE 57p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April, 1979)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Age Differences; *Cognitive Development; Elementary Secondary Education; *Multiple Choice Tests; Postsecondary Education; *Response Style (Tests); Scoring Formulas; Statistical Analysis; *Test Interpretation; *Thought Processes
IDENTIFIERS *Distractors (Tests); Proverbs Tests

ABSTRACT

The educational significance of wrong answers on multiple choice tests was investigated in over 4,000 subjects, aged 7 to 20. Gorham's Proverbs Test--which requires the interpretation of a proverb sentence--was administered and repeated five months later. Four questions were addressed: (1) what can the pattern of answer choice, across age, using frequencies of response as the raw data, indicate about the psychometric properties of learner development; (2) what process/product inferences might be drawn from these outcomes; (3) is the total correct score adequate for evaluating achievement; and (4) is the cumulative learning hypothesis valid--this hypothesis implies that the principal source of meaning is found in the frequency of right answers, or in interactions between right answers. Analyses considered various interactions between two right answers, one right answer and the wrong answer in another item, wrong answers in both items, equivalent ages, and equivalent student groups. The results suggested that learning involves a complex hierarchical sequence of interactive non-linear events; analysis of error patterns is more meaningful than frequency counts; total correct scores are only adequate for evaluation of simple recall; and the cumulative learning hypothesis is not valid for complex cognitive processes. (GDC)

Abstract

Answers on a multiple choice test were treated as categorical events and then plotted across a broad spectrum sample using a 9 month interval on age of subjects. Using the assumption of homogeneity to remove the impact of selection proportions cell Chi-Squares were obtained for alternative by alternative relationships to determine the presence or absence of "meaningful" ($\chi^2 \geq 2.4$) interactions. About 10 percent of the interactions emerged as "meaningful" at this level. For R*R (right by right answer interactions, observed frequencies were always higher than expected and most W*W interactions were in the same direction within the 5 items studied. The opposite being true for W*R interactions. Both R*R and W*W interactions occurred in significantly greater proportion than W*R interactions, and W*W interactions showed a nearly significant greater proportion of occurrence than R*R interactions. Also R*R interactions tended to be concentrated at the low end of the age sequence with W*W interactions were generally higher than the others throughout the rest of the age span. This observation is opposite to the one reported by Bock (1972). The [illegible] of concentration of interactions for particular pairs generally appeared to be localized within one of four apparent age level divisions of the range or adjacent levels. Interactions between alternatives clearly changed with age, implying that a single key scoring system for all answers across all age levels may be inappropriate for interpreting these patterns. Pairs at the same age were significantly more frequent in number than sequences with the same group suggesting age to be the more important

Abstract cont'd

developmental factor. The assumption that the distribution of these events might be random was not supported ($\chi^2=327.33$) by these data. Cognitive development, as revealed by this procedure appears to be complexly interactive and non-linear in nature. These findings appear to be generally consistant with other "wrong answer" research. A plan for continuing the exploration of these relationships was proposed and an invitation to participate was extended.

Can Developmental Status Information be Obtained from Wrong Answers?

INTRODUCTION

This present paper represents the culmination of more than 10 years of exploration into the educational significance of wrong answers on multiple choice tests, (See: Powell 1968, 1970, 1976, 1977, 1978a, 1978b, Powell and Isbister, 1974). A number of tantilizing observations seem to have emerged in repeated studies with different tests and different populations. Most notable have been:

1. Consistency of reasoning behind specific wrong answer selections has been repeatedly found, (Powell 1968, 1977).
2. Distinctive patterns among wrong answers independent of right answers has occurred consistantly, (Powell 1968, 1970, 1976, 1977, Powell and Isbister, 1974).
3. Wrong answers seem to produce "better" predictors of independent achievement measures than right answers with different tests and populations, (Powell 1970, 1976).
4. Curved line patterns among wrong answers across age have been found using quite different analytic procedures, (Powell 1976, 1978a).
5. A strong developmental trend among wrong answers (Rho=1.00) has been found, (Powell 1977).

This combination of results raises some interesting questions about the current practice of using Total Correct scores for achievement assessment as the role criterion of success.

The use of Total Correct scores for achievement assessment has a long and honourable history. It is built upon a perfectly resonable model and has had considerable support from the results of statistical procedures derived from the model. The fact that educational research has generally tended to be non-conclusive, (See: Walker and Schaffarzik, 1974), has only spurred renewed vigor in pursuit of the confounding variables which produced these results.

Figure 1 below gives the behavioral assumptions and the mathematical equivalent of these which together form the basis for classical test theory.

FIGURE 1

PRINCIPAL ASSUMPTIONS IN
CLASSICAL TEST THEORY

Behavioral Assumption	Mathematical Equivalent
1. THE KNOW-GUESS HYPOTHESIS The learner either knows the answer or guesses (blindly).	1. $X_{ij} = \partial$; i.e. the answer given by person i on item j is a 1 for a right answer and a 0 for a wrong answer.
2. THE CONTINUOUS LEARNING HYPOTHESIS Learning is continuous and cumulative. (Graphically learning would appear as a straight line when age and achievement are equated).	2. $X_i = \sum_{j=1}^{n} X_{ij}$; i.e. the score of a person on a test is the sum of all the 1's and 0's from the items on the test.
3. THE LINEAR MEASURABILITY HYPOTHESIS Achievement is linearly measurable. (Departures from the straight line in # 2 are the product of meaningless measurement error and other random events.)	3. $X_i = T_i + E_i$; i.e. the observed score (X_i) is a linear combination (sum) of the person's linear true score (T_i) and an error of measurement (E_i).

The three basic assumptions presented in Figure 1 are the fundamental formulations behind classical test theory.

The first two of these are more important than the third, because alternative measurability models which are non-linear in nature could be formulated if the need to do so were recognized and established.

In order to challenge classical test theory, then, it would be necessary to refute the first two hypothesis. These two hypothesis cannot be challenged upon mathematical or logical grounds. In order to successfully refute them, it must be conclusively demonstrated in behavioral terms, that learners do not always behave as these assumptions indicate.

The evidence reported above is strongly suggestive that both of these assumptions may be false. However, here-to-fore the evidence has been conclusive only for assumption # 1, the "Know-Guess" Hypothesis.

Treating each wrong answer as a zero (0); or by using formula scoring, is mathematically equivalent to assuming that wrong answers are "blind" guesses. That [illegible] these answer selections could be just as easily achieved by flipping coins, rolling dice, etc. as by reading the question and the proposed answer set.

Stated in this way, the know-guess hypothesis is earsily testable. Strictly random wrong answers will be distributed about equally on all wrong alternatives. It has long been known that such distribution rarely, if ever, occurs. This observation is sufficient to refute the "know-guess" hypothesis. More complex approaches which address the several sub-problems connected with this hypothesis can also be formulated and tested. It is probably safe to say that the "Know-Guess" hypothesis as formulated above has now been conclusively refuted (See: Powell and Isbister 1974, and Powell 1978). Refutation of this hypothesis, however, may be of no consequence if the only meaningful information with respect to achievement in a question is to be found in the right answers.

Another approach takes an alternative behavioral view. In this approach developed by Shuford Albert and Massengill (1966) the assumption is made that the respondant is expected to know something about the questions. In so doing, weights can be assigned to all alternatives in order of their likelihood of being right. In this way, partial information can be accommodated. Partial information, in this approach is assumed to increase the likelihood of getting the answer to the question right. This assumption is also of dubious validity. No theory, apparently considers the role of "misinformation" in answering.

Hakstian and Kansup (1975) tested a wide range of approaches to this problem and concluded that no benefits accured from any of several approaches they tested. They defined "benefit" in terms of increased reliability of test scores by adding any of several possible weighted combinations of wrong answers to the Total Correct score.

A third approach has been to determine from self-report the respondant's reasoning behind choosing wrong answers. Powell (1968) showed that this reasoning was consistant within wrong answer factors and cross-validated about two-thirds (p=.64) of the time. In a later study (Powell 1977) with the same test and different age groups the validation levels were between .50 and .60. These findings imply that the selection of wrong answers may be systematic rather than random. Such a conlusion is consistant with the implied refutation of the know-guess hypothesis, but inconsistant with the findings reported by Hakstian and Kansup.

However, Hakstian and Kansup tried to find some method of adding a weighted combination of wrong answers to the Total Correct score. They examined the impact of their attempts upon the reliability of the test scores. In this respect, they used assumption 2 and 3 to help support assumption 1. This approach to scientific procedure is incorrect. Support or refutation of the fundamental assumptions in a theory must be achieved upon a "stand alone" basis. The reason for this problem is that if learning is NOT a linear phenomenon, then the adding of a non-random distribution of wrong answers would produce either null or inconsistant results. The failure Hakstian and Kansup to find an additive approach which improved reliability could be a product of non-linearity.

To give a bit of background, it would, perhaps be helpful to illustrate how learners apparently select answers on multiple choice achievement tests.

CHILDREN'S REASONING AND WRONG ANSWERS

The easiest method of clarifying the problem of the nature of answer selection is to begin with an example. Figure 2 provides this illustration.

INSERT FIGURE 2 ABOUT HERE

In Figure 2 alternative C was most commonly selected by the 8 year olds in the study. There seems to be no logical relationship between "Quickly Come, Quickly Go" and "Always do things on Time". However, when the commonly reported reason "That's what teacher always says" is taken into account the logic is immediately evident to anyone familiar with a typical Grade 3 classroom. These children have interpreted "Quickly Come, Quickly Go" as a description of their personal classroom behavior. With several instructional groups and the complex interlocking and overlapping timetable of activities which this classroom management system necessitates, "Quickly Come, Quickly Go" and "Always do things on time" describes very succinctly what life in such a setting must be like for the young learner. In dealing with this proverb within the framework of their own experience, these children have given a perfectly valid answer which is "wrong" only because it is not included in the scoring key. Better than 50 percent of the children selecting 18c reported semantically equivalent reasons to this one.

Of course, Piaget's work has long demonstrated that children and particularly young children reason differently from adults. If this

FIGURE 2

EXAMPLE INTERPRETATION

PROVERB:

QUICKLY COME, QUICKLY GO.
(Easy Come, Easy Go)*

TRANSLATIONS:

a. ALWAYS COMING AND GOING AND NEVER SATISFIED.
Characteristic of 13 year olds.
"You Should Stick To A Job Til It's Finished."

b. WHAT YOU GET EASILY DOES NOT MEAN MUCH TO YOU.
Characteristic of adults.

c. ALWAYS DO THINGS ON TIME.
Characteristic of 8 year olds.
"That's What Teacher Always Says."

d. MOST PEOPLE DO AS THEY PLEASE AND GO AS THEY PLEASE.
Characteristic of 10 year olds.
"It Talks About Coming and Going."

*Item No. 18 from Gorham, Donald R. Proverbs Test, Psychological Test Specialists, 1956. Reproduced with permission.

were merely an isolated event, then it could easily be resolved by not using this instrument in settings where unexpected or at least alternative interpretations of these items were likely. Another approach, of course, would be to determine how to identify when children were responding in this way to this and other similar items, and to use this information to identify the problem solving approaches used by these children. This suggestion proposes an intriguing alternative approach to testing. Do the "wrong" answers reveal how the respondants are attacking the problems?

If the 10 year olds were looking for a word-for-word (literal) translation of the proverb, then their typical choice (18d) as reported in Figure 1 makes sense as well, and similarily if the 13 year-olds are searching for simple (linear) cause-effect relationships -- so does their typical choice. (18a) Their choices may be the results of the way they attack and interpret the problem. In problem solving terms, these differences in selection seem to be related to a combination of frame of reference and solution of strategy. Each age group framed the problem differently, and the 13 year-olds added extraneous information.

Using the statistical clustering of items with common modes to produce "homogeneous" answer subsets and the reported reasons to classify these subgroup the present author identified 14 subgroup of answers 12 of them among the wrong answers. These, in aggregate, accounted for 151 of the 160 alternatives in this 40 item test. The internal consistancy of the test increased from r=.76 to r=.94 with this approach. When these 14 subsets of the answers were ordered using the Simplex procedure, the resulting scaling perfectly recapitulated the age sequence of the subtests by their modes of selection (Rho=1.00). In addition, the logic of the answering followed Piaget's stages except that Concrete Operations seemed to have two substages.

Thus it would appear from this behavioral description which emerged from the reasoning behind answer selection, that at least for items of the type present in The Proverbs Test, the "Know-Guess" hypothesis may not be an appropriate behavioral description. Instead, answer selection seemed (as also occurred in the earlier study; Powell 1968) to be related to the world views of the respondants; in Piaget's terminology, to the schemata they hold.

The second approach (also used elsewhere) to refute the "Know-Guess" hypothesis was statistical and was also attempted upon a "stand alone" basis. The results are probably conclusive that the "Know-Guess" hypothesis is false (See: in particular, Powell and Isbister, 1974). Given an opportunity to make a reasonable decision, apparently most people will not respond "blindly" to multiple choice achievement test items.

The details of the study concerning children's reasoning just summarized are reported elseshere. (Powell 1977)

A SEQUENCE OF DEVELOPMENT FROM WRONG ANSWERS

In a general way the study just reported above has already demonstrated a strong developmental trend among wrong answers.

But this observation raises more questions than it answers. If such a powerful developmental trend is present among wrong answers, why has it not shown up before?

In most approaches to testing, the right answers are considered to be the main if not the only source of information about the learner and to form a straight line scale. This approach involves assuming learning to be cumulative, (i.e. assumption # 2). These assumptions justify the counting of right answers and/or the addition of subscores to form subtest scores and the Total Correct score. The Total Correct score or some linear transformation of it is often the sole basis for evaluating learner progress. A good example of this approach is the Peabody Picture Vocabulary Test, which is a broadly used standardized test of receptive vocabulary. The norms, on this test, however, are not presented in terms of vocabulary development patterns, but rather as deviation Intelligence Quotients.

If the relationships among answers are not straight lines, then much more complex approaches to tests than classical test theory may be in order. At this point there are two possible approaches. First is the "arm chair" approach in which a series of assumptions are made and a mathematical model built and tested, often with simulated data. This approach requires much more mathematical skill than the present researcher possesses. A second approach is to collect a large sample of data and to explore these data for their structural and relational properties i.e. to develop a "grounded theory". (After Glaser and Strauss, 1967).

Making the smallest possible number of assumptions forces the use of all data as frequencies only. In this case, the behavioral assumption made must be that the respondant assigns categorical values to each alternative in each item. In this case alternatives should be statistically related, either as artifacts of random variations, or on the basis of similarities among assignment procedures.

The most reasonable method to use is the Chi Square (χ^2) procedure, which though robust, is less sensitive than procedures which make more assumptions. In addition, on an alternative-by-alternative basis, many of the events examined will need to be cell Chi Squares. In the absence of critical values for zero degrees of freedom, cell Chi Square values present an important interpretation problem to the use of the Chi Square technique in this study. Furthermore, the four alternatives in any one item can be compared in several ways. Each approach produces a different contingency table from these same data. They can be compared in pairs of items either across the entire sample or by using subgrouping from within the sample.

Another approach is to compare an item with itself between reasonable subgroupings of the sample such as at different age levels, or between sexes or different administration times.

In addition to the possibility of formulating the contingency tables in several ways, different assumptions can be tested with each table. There are three common approaches. Random, homogeneous, and/or external model assumptions can be used.

In the random event approach, each cell is expected to occur with the same probability or with a distribution reflecting the properties of the "normal curve". In this case the expected frequencies used in each cell within one table are identical to each other or related to the area under the curve. Using equal cell frequencies is pointless here, because the "Know-Guess" hypothesis has already been refuted.

Under the homogeneous assumption, only the marginal proportions of these events are assumed to be meaningful. This is the approach used in this study.

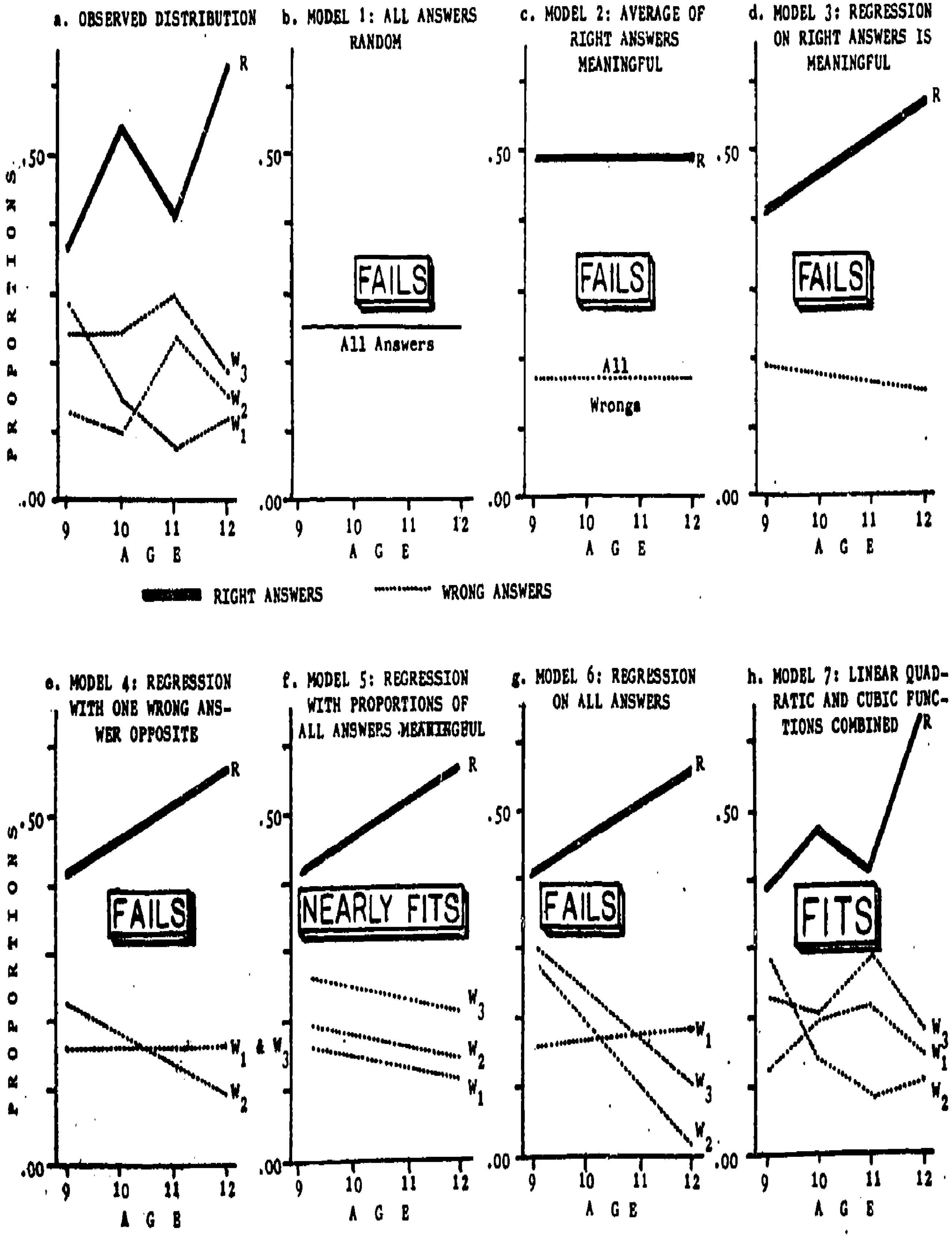
In the third case, a model which is external to the contingency table is tested with the table for goodness of fit. An approach to external model building called MULTIQUAL was developed by Bock (1973) can be used to compare patterns among cell frequencies with some form of external mathematic model. Figure 3 illustrates the outcomes of this last approach.

INSERT FIGURE 3 ABOUT HERE

In summary, Figure 3 shows the testing of all possible straight line formulations of an age sequence of frequencies within one item. Among these straight line models the only one which comes close to a fit includes the regression line for the right answers and the average proportions of selection among the wrong answers. Since the average proportions are apparently a necessary component of the model, these events are not random, further negating the use of the random assumption in this present study. However, even this fails to fit

FIGURE 3

DISTRIBUTION OF RESPONSES TO ITEM ONE FITTED TO SEVERAL POSSIBLE MODEL THEORIES



a. OBSERVED DISTRIBUTION

b. MODEL 1: ALL ANSWERS RANDOM

c. MODEL 2: AVERAGE OF RIGHT ANSWERS MEANINGFUL

d. MODEL 3: REGRESSION ON RIGHT ANSWERS IS MEANINGFUL

e. MODEL 4: REGRESSION WITH ONE WRONG ANSWER OPPOSITE

f. MODEL 5: REGRESSION WITH PROPORTIONS OF ALL ANSWERS MEANINGFUL

g. MODEL 6: REGRESSION ON ALL ANSWERS

h. MODEL 7: LINEAR QUADRATIC AND CUBIC FUNCTIONS COMBINED to the $p > .05$ criterion. The MULTIQUAL formulation (MODEL 7) includes both quadratic and cubic functions along with straight lines before a fit is achieved.

At least for this one item of the Proverbs Test (Gorham 1956) and for these 4 ages, (9 through '2 inclusive), the patterns of answer selection among all alternatives are apparently not straight lines. This same test with a different (larger) sample was used in this present study. Making some assumptions discussed later, the original data validated by replication but the MULTIQUAL model did not. The findings just summarized here are reported in much more detail elsewhere (Powell, 1978a).

The finding of curved line relationships among answers in this item supports findings reported elsewhere, (Powell 1976, Yu 1977), the refutation of the "Know-Guess" hypothesis, and implies possible problems with the "Cumulative Learning" hypothesis. However, this later study (Powell, 1978a) does not deal with the "Cumulative Learning" hypothesis on a "stand alone" basis.

In order to refute the "Cumulative Learning" hypothesis, it would be necessary to approach it in a situation uncontaminated by other influences. The procedure finally decided upon, involves removing the influence of the aggregate frequencies upon events and examining between alternatives among items, interactions for systematic events.

The purpose of this present study then, is to attempt to refute the "Cumulative Learning" hypothesis on a "stand alone" basis.

Refutation of this hypothesis, would be necessary and sufficient to refute the use of classical test theory as an approach to the evaluation of learning progress.

THE SAMPLE USED

Since it had become evident that meaningful information about the development of cognition might be found among wrong answers, a major study has been mounted involving more than 4,000 children in the age range from about 7 years to over 20 (grades 3 to 13 inclusive). The test (Gorham's Proverbs Test) was administered in conjunction with a personality test, and both were repeated after a 5 month interval. Because of the reading level the personality test was not used with Grades 3 and 4. The distribution of ages (condensed into 5 month intervals) for the Proverbs Test is given in Table 1.

INSERT TABLE 1 ABOUT HERE

The purpose of aggregating to 5 month intervals is two-fold. First, it makes a satisfactory minimum sample size in all but 4 of the 60 groups. Second, since, the second administration was 5 months after the first, comparison between two groups of the same age range assures independence of group membership. Also, comparison between October (N) and March (N + 1) age levels gives a sequential comparison of the two administrations with largely the same subjects in each group.

TABLE 1

DISTRIBUTION OF SUBJECTS
IN THIS STUDY BY AGE LEVEL AND THE TIME
OF ADMINISTRATION OF THE TEST.

AGE LEVEL	AGE IN MONTHS	AGE IN YEARS	OCTOBER ADMIN-ISTRATION		MARCH ADMIN-ISTRATION	TOTALS
1	AIM > 96		43		3	46
2	96 - 100	8	68	MOSTLY	32	100
3	101 - 105		70	SAME	55	125
4	106 - 110	9	130	GROUP	53	183
5	111 - 115		101		120	221
6	116 - 120	10	127		78	205
7	121 - 125		137		101	238
8	126 - 130		142		114	256
9	131 - 135	11	145		118	263
10	136 - 140		129		125	262
11	141 - 145	12	165		106	271
12	146 - 150		135		131	266
13	151 - 155		138		100	238
14	156 - 160	13	152		104	256
15	161 - 165		114		132	246
16	166 - 170	14	163		101	264
17	171 - 175		262		150	412
18	176 - 180	15	264		237	503
19	181 - 185		258		247	505
20	186 - 190		251		255	506
21	191 - 195	16	249		228	477
22	196 - 200		219		220	439
23	201 - 205	17	210		219	429
24	206 - 210		171		173	344
25	211 - 215		186		130	316
26	216 - 220	18	125		131	251
27	221 - 225		87		81	168
28	226 - 230	19	47		66	113
29	231 - 240	20	20		43	63
30	240 < AIM		10		14	24
		TOTALS	4319		3676	7995

There s not 100 percent equivalence between the groups in these two sets since they represent administration to everyone in 10 schools in October and only 9 schools in March. One school voluntarily withdrew. It also includes some who were absent in October, yet, were present in March, and so on. Person-by-person comparisons between administrations can be made from these data but were not made in the particular study reported here.

HYPOTHESIS TESTED

The procedure of examining contingency tables for departures from homogeneity effectively removes the impact of aggregate selection proportions.

The "Cumulative Learning" hypothesis is used in a manner which implies that the principal source of meaning is to be found in the aggregate frequency of "right" answers. If interactions are ever considered, these are derived from the right answers only. These sub-test aggregates and whole test aggregates often become the sole basis for achievement evaluation.

Under the null hypothesis condition, then, departures from homogeneity should be at best purely random and at worst should be a phenomenon confined mainly to right by right answer interactions.

If these null hypothesis are supported, then the right answers and/or subgroupings thereof would be <u>necessary</u> <u>and</u> <u>sufficient</u> to describe successful learning. If they are refuted then right answers and/or subgroupings thereof would not be <u>sufficient</u> to describe learning. In this latter case, since the events being examined are interactive, this refutation is by itself sufficient to refute the "Cumulative Learning" hypothesis.

The reason why the finding of meaningful wrong-by-wrong answer interactions refutes the "Cumulative Learning" hypothesis is that such interactions are, by definition non-linear. If right answers are not sufficient to describe achievement, because of meaningful wrong-by-wrong answer interactions being present among these data, then the learning itself must be non-linear and interactive.

If learning is non-linear, then the "Cumulative Learning" hypothesis becomes insufficient to describe learning events and fails as a model.

If the "Cumulative Learning" hypothesis is refuted, then classical test theory becomes an inappropriate model for achievement since the first term in the fundamental equation $(X_i=T_i+E_i)$, namely X_i has been demonstrated to be insufficient to describe learning by means of the refutation of the two hypotheses ("Know-Guess" and "Continuous Learning") upon which it is built. On the basis of this reasoning, present approaches to testing would be expected to either stand, become qualified, or fall on the basis of this present study.

PROCEDURE

Using the breakdown given in Table 1 and the first 5 items on the Proverbs Test, an inter-item comparison for each 5 month aggregate for each of the two adminisrations was obtained. With 10 inter-item comparisons, 30 age levels, and 2 administration times; 600 four-by-four contingency tables were produced. Each alternative of one item being cross-tabulated with each alternative on each other items among all 5 items. This approach examines about 1/80th of the interaction data available from this 40 item test.

These 600 tables were generated including the cell Chi Squares under the assumption of homogeneity. In this assumption, the expected values are generated on the basis that all meaning resides in the marginal totals. For instance, if there are 180 students in the group, 100 get one item correct and 120 get the other correct the the expected joint occurance would be 100 X 120 ÷ 180 = 66.67. If 80 students actually chose the right answer for both items, the cell Chi Square would be $(80 - 66.67)^2 \div 66.67 = 2.67$.

Using the frequency distribution of all 9,600 cell Chi Squares (divided into 10 groups and averaged to get maximum stability) a distribution of the cell Chi Squares was obtained. Comparison of this distribution with extrapolations from a Chi Square table, a critical value $\chi^2 \leq 2.4$ was obtained. Any cell Chi Square in this range was considered to be "meaningful". About 10 percent of the cell Chi Squares fell in this category.

These values cannot be called "significant" since the mathematical distribution for Chi Square with zero degrees of freedom is not available, hence the use of the alternative term "meaningful".

In the above example, the 80 students should, collectively, be considered to have selected both correct answers jointly at a "meaningful" level (O>E). That is, the observed frequency is "meaningfully" (rather than significantly) higher than expectation. If only 40 had chosen these two items jointly then the observed frequency would be "meaningfully" below expectation (O<E).

in this case it is more likely for these students to get only one of these two items right than to get both right. The third case where the cell Chi Square is <2.4 implies that the joint occurance (both items right) is determined by the proportional tendencies to get either right independent of the other rather than upon some systematic property of the behavior of the respondants with these two items in interaction (O=E). In this latter case, such meaning as occurs is to be found among the frequencies rather than the interactions.

INTERACTIVE RELATIONSHIPS EXAMINEO

There are nine possible relationships among all the alternatives on multiple choice achievement items.

1. Between the two right answers (R*R)
2. Between one of the right answers and the wrong answers in the other item (W*R)
3. Between the wrong answers in both items (W*W)

These three can be tabulated across the O>E, O=E and O<E relationships. It should be noted that these categories are different from those assumed on a personality test where no answer is considered to be correct and between item dependencies and/or within item scales may be deliberately included when the test is constructed.

For a 4 alternative per item test there will be 1 of the (R*R) class; 6 of the (W*R) class; and 9 of the (W*W) class in each item. In theory then, if interactions are purely random, then the equation $R*R=\frac{W*R}{6}=\frac{W*W}{9}$ should hold true in all but about 30 of the 600 tables to be examined. Also if all events are assumed to be about 2.5 percent will be of the O>E type, the same for O<E and 95 percent will be of the O=E type. If actually meaningful results are being derived from

this study then systematic departures from these patterns should be evident.

With this information the null hypothesis can now be expressed in mathematical terms.

NULL HYPOTHESIS:

$$Ho: \frac{f(O>E)R*R}{N_1} = \frac{f(O<E)R*R}{N_1}$$

$$= \frac{f(O>E)W*R}{N_2} = \frac{f(O<E)W*R}{N_2}$$

$$= \frac{f(O>E)W*W}{N_3} = \frac{f(O<E)W*W}{N_3}$$

$$= .025$$

Where the N_k values are the maximum possible number of these types of interaction whose frequencies are described in the numerator. Support for H_o implies that the only source of meaning in the test under study would be the frequencies of answers. In this case, current practice would probably be supported.

ALTERNATIVE HYPOTHESIS 1:

$$H_{11}: \frac{f(O>E)R*R}{N_1} > .025 > \frac{f(O<E)R*R}{N_2}$$

$$H_{12}: \frac{f(O>E)W*R}{N_2} > .025 > \frac{f(O<E)W*R}{N_2}$$

$$H_{12}: \frac{f(O>E)W*W}{N_3} = \frac{f(O<E)W*W}{N_3} = .025$$

In this case, right answer aggregates and appropriate right answer subtests built from the R*R interactions would form the necessary and sufficient summary of achievement status information.

All possible altemate hypothesis can be built in this manner. Most of the remainder, however, imply that right answers are not sufficient to account for achievement status. The only exception to this statement would be the case where R*R and W*W are random but W*R shows the right and wrong answers polarizing. This latter would express the "Linear Dependency" proposition which is a logical deduction from the Know-Guess hypothesis and which has already been refuted elsewhere (See: Powell and Isbister, 1974). It should be noted, that these refutations have been upon behavioral rather than mathematical terms. The statistical relationships found among live data do not support the observations which should have occurred if these models were appropriate. The mathematics of existing test theories in current general use are logico-deductive systems, which as such are refutable only upon the basis of errors in the deductive processes. The refutation of concern here is to the applicability of these models to the particular kinds of data being studied.

COMPARISONS MADE

In addition to the testing of these hypotheses made, the following comparisons were made.

1. Frequencies and proportions of "meaningfully" events separated by item comparisons and aggregated by age level in each of these 3 categories. (W*W, W*R, and R*R)

2. Frequencies and proportions of "meaningful" events separated by age level and aggregated across item interaction in each of the 3 categories.

3. Frequencies of opposite events by item and category.

4. Frequencies of age equivalent pairs and student equivalent group sequences between the two administrations.

5. Continuities between pairs of alternatives across age levels.
In this latter case no aggregation was used.

Looking at the comparitive distribution of proportions of W*W, W*R, and R*R across age may give some idea of the degree to which each contributes meaningfully at different age levels. Bock (1972) found that wrong answers added meaningfully to the results of those students below the median but not above it. However, he used vocabulary items which are essentially "know-guess" in format or at the Knowledge Level in Bloom's Taxonomy (1956). The Proverbs Test contains "translation" items which are at least at the Comprehension level of Bloom's Taxonomy. If the pattern Bock found continues for items at a higher level of cognitive processes, then for evaluative purposes right answers would be necessary and sufficient. Wrong answers might still have potential diagnostic value particularly for the lower scoring students. If this pattern is reversed with W*W > R*R at the higher ages, then wrong answers may need to become an important part of evaluation.

In addition, of course, cross-referencing events by both item and age level makes it possible to establish the accuracy of both tabulations by the match of the totals.

The frequencies of opposites by item and category should give some indication of the amount of "noise" in these observations both within item pairs and by category. In this case, a "noisy" solution for wrong answer interactions would be $\frac{f(O>E)W*W}{N_3} = \frac{f(O<E)W*W}{N_3} > .025$.

A look at the age equivalent pairs and student equivalent group sequences should give some idea of the degree of consistancy among these observations and whether age equivalent or student equivalent events are more meaningful developmentally. If age equivalents is the more stable then developmental sequence is supported over intragroup stability, the reverse implies the opposite conclusion.

CONTINUITY OF INTERACTIONS ACROSS AGE

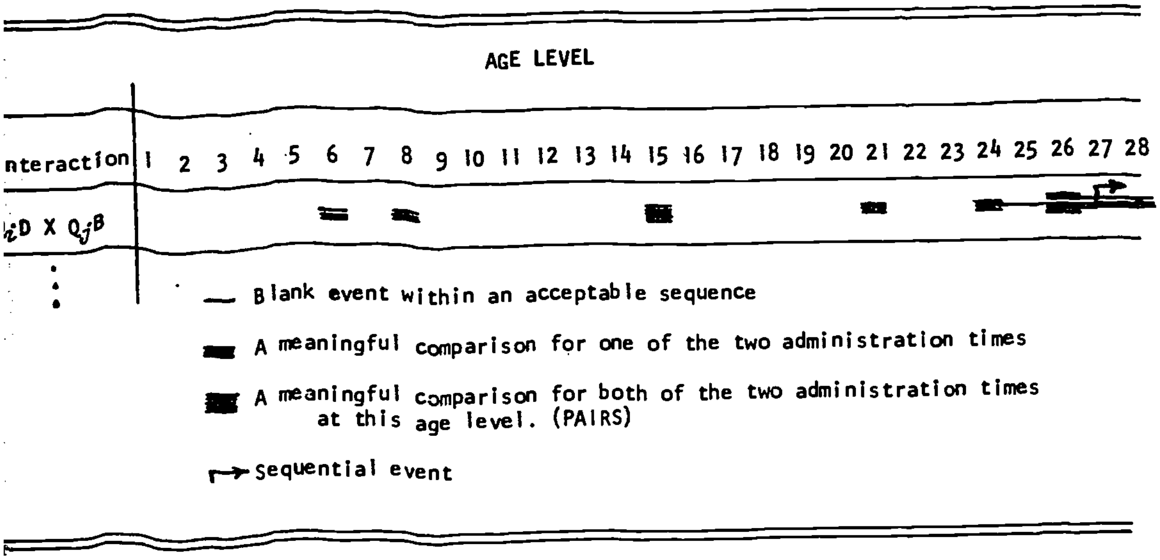
Continuous sequences operationally defined using the following algorithm.

1. The total occurance of meaningful events on any alternative pair across all age levels must be at least 6 to form a continuity. This means that 10 percent of the possible 60 must be represented.
2. Of these at least half of them or 3 events must be close enough together that not more than one age level separates any two events.
3. If more than one age level separates any two elements in a sequence, the sequence is assumed to be discontinuous at that point.
4. Less than three elements at different ages are not considered as a continuous sequence.
5. A pair is considered a single age event for the purposes of establishing a continuous sequence, but is considered with respect to the density of that sequence.
6. Overlap of no more than 3 events into one of the four "cognitive processing level" subdivisions developed in this study is assumed to place the sequence into the level of greatest density.

Here is an example to illustrate these rules.

Figure 4

"Example of a Continuous Sequence"


AGE LEVEL
nteraction
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
iD X QjB
Blank event within an acceptable sequence
A meaningful comparison for one of the two administration times
A meaningful comparison for both of the two administration times at this age level. (PAIRS)
Sequential event

In this case there are 10 meaningful events at eight different age levels including 2 pairs. Had there been less than 6 instead of 10, none of these would have been considered further. The Age Level 6 BLANK 8 sequence is close enough but not long enough to be considered to form a continuous sequence. The pair by itself at Age Level 15 is not close enough to two others to be continuous. The event at Age Levél 21 is separated by two age levels from its nearest neighbor and is, therefore not part of the continuous sequence. The pattern 24, BLANK, 26, 27, 28 meets all criteria and is, therefore considered to form a continuous sequence. The balance are assumed to be statistical artifacts and ignored. Thus the density of this "continuous sequence" in 4 out of 9 or .50. The arrow above 27 toward 28 means that this interaction was found in the first administration for age level 27 and in the second administration for age level 28. In this case, this event would seem to be more a characteristic of this group than of the age levels.

If continuous sequences are found in any number, two important questions can be considered. First, do these tend to spread out across the entire age range? If they do, then development would seem to involve changes within a relatively stable pattern and a single answer sub-group key can be used for all age levels. On the other hand, if continuous sequences appear to be relatively short and confined to specific age level spans or to form a "stairway" across the age levels then a single key for all age levels is not sufficient to describe development in this case from these answers on such tests. In this latter case (i.e. the appearance of a "stairway") strong support for a development sequence among answer patterns will be found. The presence of a "stairway will mean that development not only would be affecting which alternatives are chosen, but the between choice relationships as well. Should these effects vary with the age of the learner, a complex curved-line pattern would be implied. The MULTIQUAL results (Powell, 1978a) already has been presented provided the implication that rectilinear models may be inappropriate for evaluating learner performance.

The appearance of a "stairway" would further support this non-linearity, since in this case between alternative relationships would be changing with age.

Here then are the four basic questions addressed in this study. 1) What can the pattern of answer selection across age, using only frequencies as the raw data, tell us about the statistical and/or psychological (behavioral outcome) properties of learner development? 2) What process/product inferences (if any) might reasonable be drawn from these outcomes? 3) Is the total correct score a sufficient basis for evaluating learner achievement? 4) Is the "Cumulative Learning" hypothesis valid as a behavioral description of the nauture of learning events?

RESULTS

Considering the observations obtained, how do these compare, with the null hypothesis being explored by this study? Table 2 gives these results.

INSERT TABLE 2 ABOUT HERE

The expectation that all interactions would be random, or that patterning would favor the use of right answer aggregates over any other procedure was not supported by these findings. The relative sizes of the cell Chi Squares speak for themselves.

TABLE 2

FIT OF THE FREQUENCIES OF MEANINGFUL EVENTS TO THE RANDOM ASSUPTION

	INTERACTION		
RELATIONSHIP	R*R	W*R	W*W
0>E	0 = 60 E_1 = 15 x_1^2 = 135.00 E_2 = 30 x_2^2 = 30	0 = 12 E_1 = 90 x_1^2 = 67.60 E_2 = 180 x_2^2 = 156.8	0 = 633 E_1 = 135 x_1^2 = 1837.07 E_2 = 270 x_2^2 = 488.03
0=E	0 = 540 E_1 = 570 x_1^2 = 1.58* E_2 = 480 x_2^2 = 7.5	0 = 3368 E_1 = 3420 x_1^2 = 0.79* E_2 = 3240 x_2^2 = 5.06	0 = 4731 E_1 = 5130 x_1^2 = 31.03 E_2 = 4860 x_2^2 = 3.42
0<E	0 = 0 E_1 = 15 x_1^2 = 15 E_2 = 30 x_2^2 = 30	0 = 220 E_1 = 90 x_1^2 = 187.78 E_2 = 180 x_2^2 = 8.89	0 = 39 E_1 = 135 x_1^2 = 68.27 E_2 = 270 x_2^2 = 197.63
TOTALS	0 = 60 E = 600 x_1^2 = 151.58 x_2^2 = 67.50	0 = 232 E = 3600 x_1^2 = 256.17 x_2^2 = 170.75	0 = 672 E = 5400 x_1^2 = 1936.30 x_2^2 = 689.08
GRAND TOTALS	0 = 964	E = 9600	x_1^2 = 2344.05 x_2^2 = 927.33

* CELL CHI SQUARE NOT MEANINGFUL

There were two tests conducted here. In the case where the first range of expected values was considered, the random assumption fitting to the normal curve was that the ratio of 1:38:1. That is, O>E and O<E should be no more than 2.5 percent of the distribution. In the second case, since the critical value for the cell Chi Squares used was arbitary, the average proportion of departure 964 to 9600 was used instead. This gave a ratio of 1:18:1. That is, O>E and O<E should each represent about 5 percent of the total. In effect, this is a test of symmetry and of order. The frequencies are clearly assymetric and order R*R<W*R<W*W in order of meaning.

These observations leave little question that there may be meaning among interalternative interactions.

What are the distributive patterns of these meaningful events? The order presented on Page 10 will be followed. Table 2 gives the relationships among the frequencies of meaningful relationships (cell $\chi \geq 2.4$) among item pairs aggregated across age.

INSERT TABLE 3 ABOUT HERE

Only two of the R*R relationships might be considered large. Q_2 X Q_1 is 16 out of a possible 60 and Q_4 X Q_2 is 18 out of 60. Two others are close to 10 percent of the possible while the other 6 out of 10 have 4 or less meaningful relationships between right answer pairs. For most of these items, then, the "meaningful" relationships between them on the right answers are probably statistical artifacts even though all relationships occurred in only one direction.

TABLE 3

SUMMARY OF
THE REQUENCIES OF
MEANINGFUL RELATIONSHIPS BY ITEM ACROSS AGE

	ITEM			
ITEM	1	2	3	4
2	W*W = 57 W*R = 32 R*R = 16 PAIRS = 14 SEQUENCES = 3			
3	W*W = 67 W*R = 23 R*R = 4 PAIRS = 13 SEQUENCES = 11	W*W = 76 W*R = 22 R*R = 4 PAIRS = 12 SEQUENCES = 8		
4	W*W = 43 W*R = 16 R*R = 6 PAIRS = 5 SEQUENCES = 4	W*W = 59 W*R = 23 R*R = 18 PAIRS = 12 SEQUENCES = 9	W*W = 89 W*R = 22 R*R = 1 PAIRS = 14 SEQUENCES = 12	
5	W*W = 70 W*R = 29 R*R = 5 PAIRS = 11 SEQUENCES = 13	W*W = 51 W*R = 18 R*R = 1 PAIRS = 5 SEQUENCES = 3	W*W = 85 W*R = 23 R*R = 2 PAIRS = 8 SEQUENCES = 10	W*W = 75 W*R = 24 R*R = 3 PAIRS = 9 SEQUENCES = 7

TOTALS 1. WRONG BY WRONG = 672
POSSIBLE (N_1) = 5400
PROPORTION = .1244

2. WRONG BY RIGHT = 232
POSSIBLE (N_2) = 3600
PROPORTION = .0644

3. RIGHT BY RIGHT = 60
POSSIBLE (N_3) = 600
PROPORTION = .1000

4. AGGREGATE = 964
POSSIBLE = 9600
PROPORTION = .1004

5. PAIRS AT SAME AGE = 103
POSSIBLE (N_4) = 472
PROPORTION = .2182

6. SEQUENCES BETWEEN:
FIRST & SECOND TESTING = 93
(LESS CAUSED BY PAIRS) = 80
POSSIBLE = 472
PROPORTION = .1695

DIFFERENCES OF PROPORTIONS:

1. $\frac{W*W}{N_1} - \frac{W*R}{N_2}$; z = 9.444; p .000

2. $\frac{W*W}{N_1} - \frac{R*R}{N_3}$; z = 1.877; p .05

3. $\frac{R*R}{N_3} - \frac{W*R}{N_2}$; z = 3.236; p .01

4. $\frac{PAIRS}{N_4} - \frac{SEQ.}{N_4}$; z = 2.411; p .05

For most of the item by item interactions, the frequency of occurance of wrong answer pairings is substantially larger than 9 to 1 in favor of W*W, (i.e. the wrong-by-wrong answer interactions). In fact only 3 of the 10 pairs show a ratio smaller than this. The difference between proportions between W*W and R*R is almost big enough in favor of W*W as being statistically significant by larger than R*R. If the effects of Q_2 X Q_1 and Q_4 X Q_2 were removed, then this difference would be highly significant.

For most of the 10 item interactions of the 5 items used wrong answer combinations would seem to tend to more "meaningful" than right answer conbinations. Since only about 1/80th of the possible number of interactions have been considered here, a continuation of this current trend would be very likely to make this difference significant.

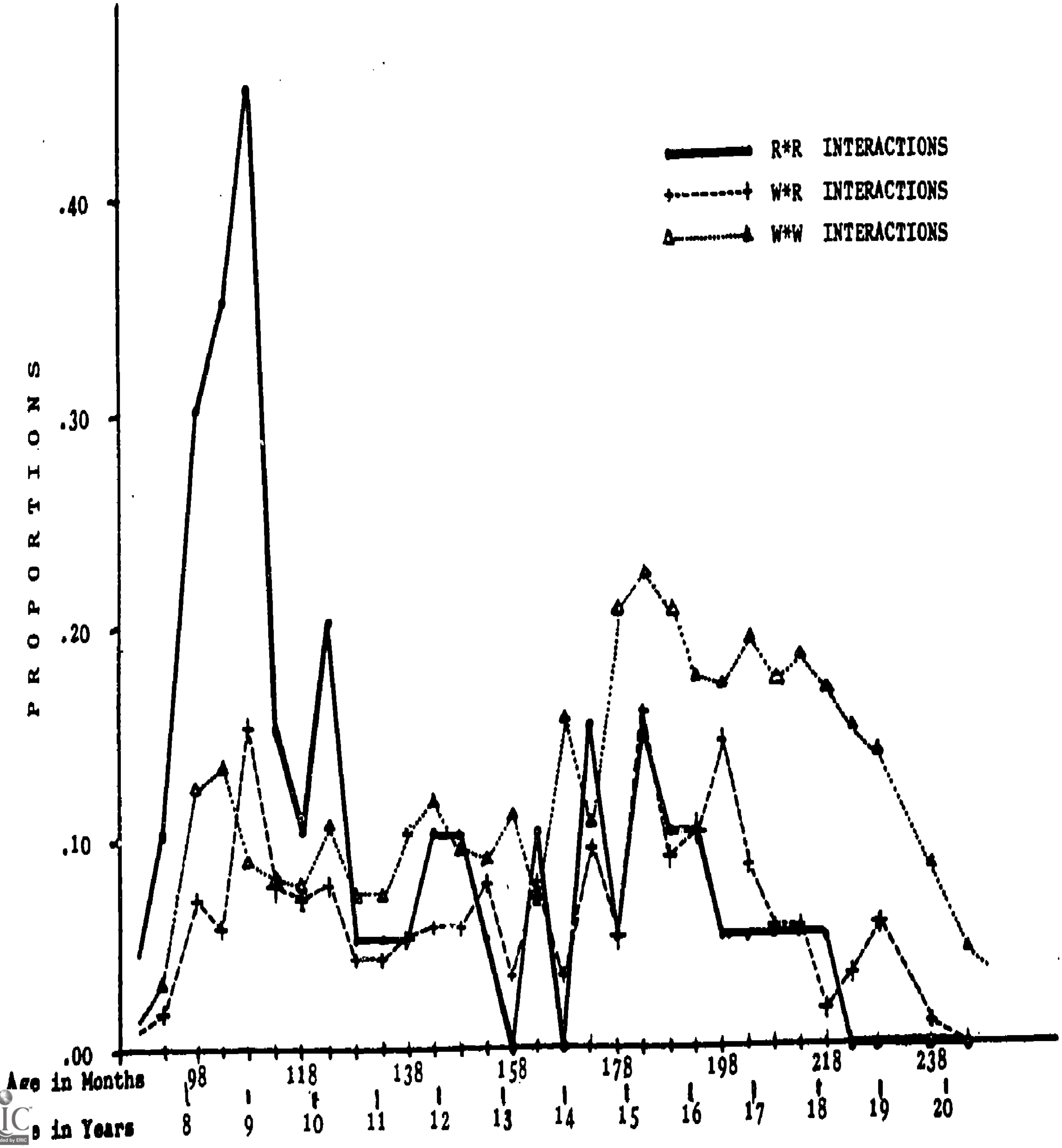
The second consideration was the pattern among these three categories of events by age level aggregated across item interactions. In this case the maximum possible R*R in each cell is 20, W*R is 120, and W*W is 180. To make for easier comparison the three patterns are shown on one graph in proportions of the total possible for each category. Figure 5 gives this information.

INSERT FIGURE 5 ABOUT HERE

Two observations are worthy of note. First, the R*R interaction frequencies are heavily loaded at the low end of the age scale. Second, from about AGE 11½ YEARS onwards, the W*W relationships are

FIGURE 5

COMPARISONS AMONG PROPORTIONS OF OCCURANCES OF INTERACTIONS ACROSS AGE



almost always above R*R or W*R or both. The single exception occurred at AGE 13½. This observation implies that the wrong answers contain more meaning for older but not younger children. This observation is the opposite to the findings reported by Bock. However, his items were vocabulary items which implies "know-guess" strategy involved in answering. If the "know-guess" phenomenon was also operative for the younger children in this study, then there is no contradiction between thesa results.

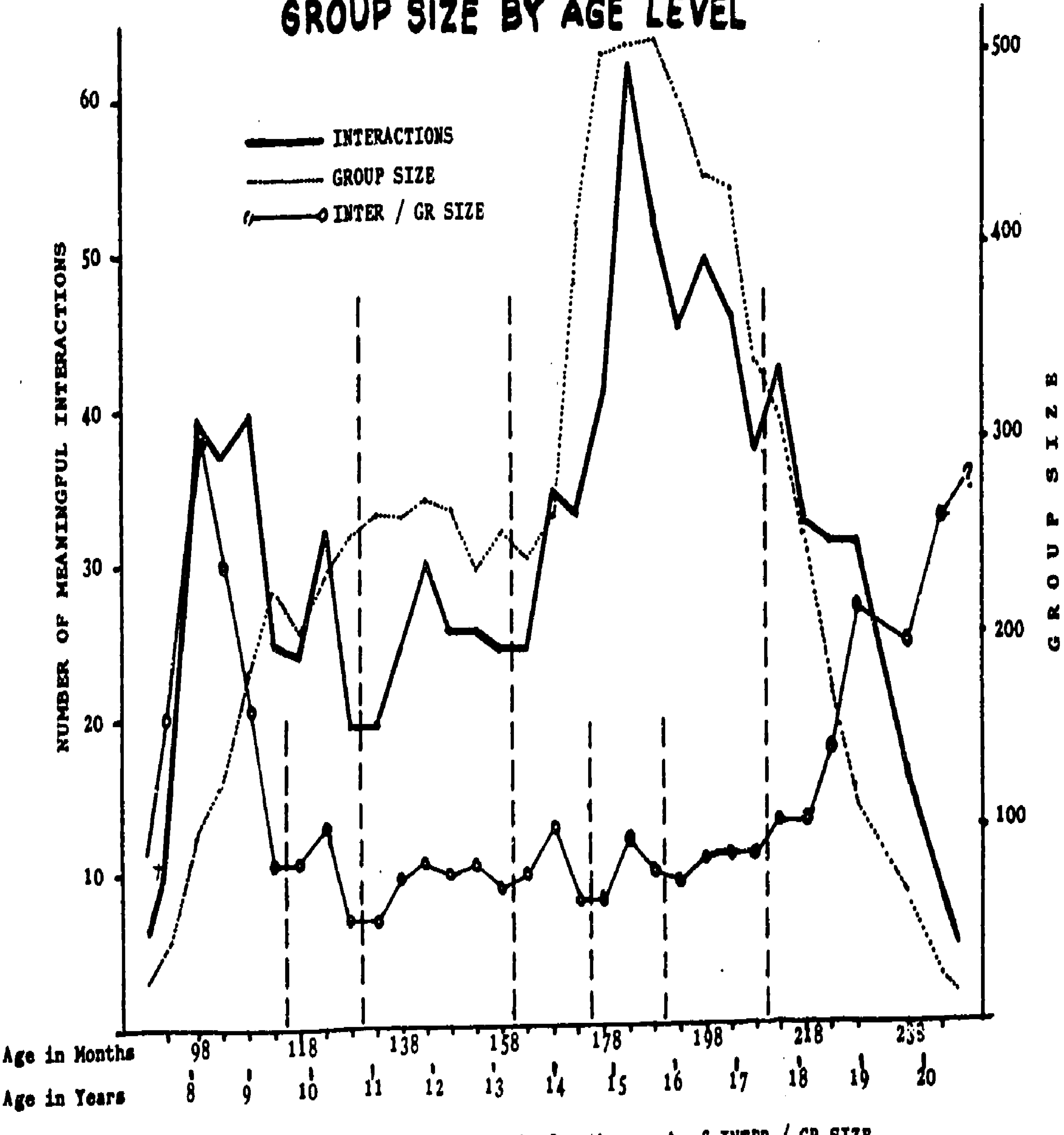
Another item of interest arises from the comparison between the aggregate frequency of relationships by age across item intereactions and the sample size for each age group. Figure 6 gives these comparisons.

INSERT FIGURE 6 ABOUT HERE

It is evident from Figure 6 that sample size does directly influence the frequency of meaningful relationships thoughout most of the range. However, it is equally evident that this is not the only influence. There seems to be at least three transition points where the relationship frequency is lower than sample size influence would predict. These transition points were used to divide the sequence into four "process" levels which roughly correspond to Piaget's stages (if Concrete Operations comes in two parts) and with the stages found elsewhere (See: Powell, 1977). Such transition points suggest that at least one aspect of development may involve cyclic

FIGURE 6

A COMPARISON BETWEEN MEANINGFUL INTERACTIONS & GROUP SIZE BY AGE LEVEL



NOTE: Read the left hand scale as a ratio for the graph of INTER / GR SIZE.

increase and decline of the degree to which interactions may be meaningful. The points of lowered meaningful interaction seem to coincide with the transitions between stages. This observation seems to correspond with the transitions between stages. This observation seems to correspond to the increase in the variety and frequency of errors found to occur at similar transition points (See: Powell, 1976). It also seems to be related to Piaget's inference that the psychological Schemata of a learner may need to be restructured at each transition point. These observations would seem to reinforce the possibility that development may be a non-linear phenomenon.

The third curve shows the pattern when the frequency of interactions is divided by the group size. The resulting proportions approach .05 where the dashed lines are located. The peak at the left is composed in large part of R*R interactions while the one on the right is largely W*W interactions. Apparently, in contrast to Bock's (1972) observation. In this study using a "higher mental process" test, wrong answer interactions seem to increase in meaning with age. This phenomenon seems particularly evident once the impact of group size is removed from these data.

In the later age levels (See: Figure 8) the number of right answers seems to fall off sharply, while the wrong-by-wrong interactions continue to increase in importance.

The frequencies of all six possible relationships (either O>E or O<E for R*R, W*R, and W*W) are given in Table 4.

I%SERT TABLE 4 ABOUT HERE

This table is self explanatory. Well under 1 percent of the total frequency of observations and 10 percent of the actual observations occur in relationships whose direction would imply meaninglessness to wrong-by-wrong answer interations. Using the cell Chi Square $\geqslant 2.4$ criterion produced very little "noise". None of these occur with the R*R interactions. The frequency of statistical artifacts would appear to be satisfyingly low. Such low noise levels would support the findings of high levels of variance accounted for reported elsewhere (See: in particular Powell, 1976).

The frequencies of pairs vs. sequences was given in Table 2. Pairs exceed sequences. Some 13 sets of pairs were consecutive. In this case the sequence was caused by the pairing. If these are moved then the proportion of sequences is singificantly smaller than the number of pairs. This observation would seem to have two implications. First age level seems to be more important than group membership in the formation of a continuous sequence. Second a span of 5 months duration may be sufficiently large to make a significant discrimination using this procedure. Using Total Correct scores alone, an age span three times as big would be needed before group differences become significant.

Figure 7 gives the "sontinuous sequences" which were found among these data. These continuous sequences are arranged

TABLE 4

FREQUENCIES OF ALL
POSSIBLE MEANINGFUL RELATIONSHIPS

		O>E	O<E	TOTALS
	Q2 X Q1	51	6	57
	Q3 X Q1	65	2	67
	Q4 X Q1	41	2	43
	Q5 X Q1	65	5	70
	Q3 X Q2	74	2	76
W*W	Q4 X Q2	56	3	59
	Q5 X Q2	48	3	51
	Q4 X Q3	88	1	89
	Q5 X Q3	85	-	85
	Q5 X Q4	60	15	75
	TOTALS	633	39	672
	Q2 X Q1	-	32	32
	Q3 X Q1	-	23	23
	Q4 X Q1	1	15	16
	Q5 X Q1	2	27	29
	Q3 X Q2	2	20	22
W*R	Q4 X Q2	1	22	23
	Q5 X Q2	3	15	18
	Q3 X Q3	1	21	22
	Q5 X Q3	-	23	23
	Q5 X Q4	2	22	24
	TOTALS	12	220	232
	Q2 X Q1	16	-	16
	Q3 X Q1	4	-	4
	Q4 X Q1	6	-	6
	Q5 X Q1	5	-	5
	Q3 X Q2	4	-	4
R*R	Q4 X Q2	18	-	18
	Q5 X Q2	1	-	1
	Q4 X Q3	1	-	1
	Q5 X Q3	2	-	2
	Q5 X Q4	3	-	3
	TOTALS	60	-	60

according to the sections of the four cognitive levels proposed above.

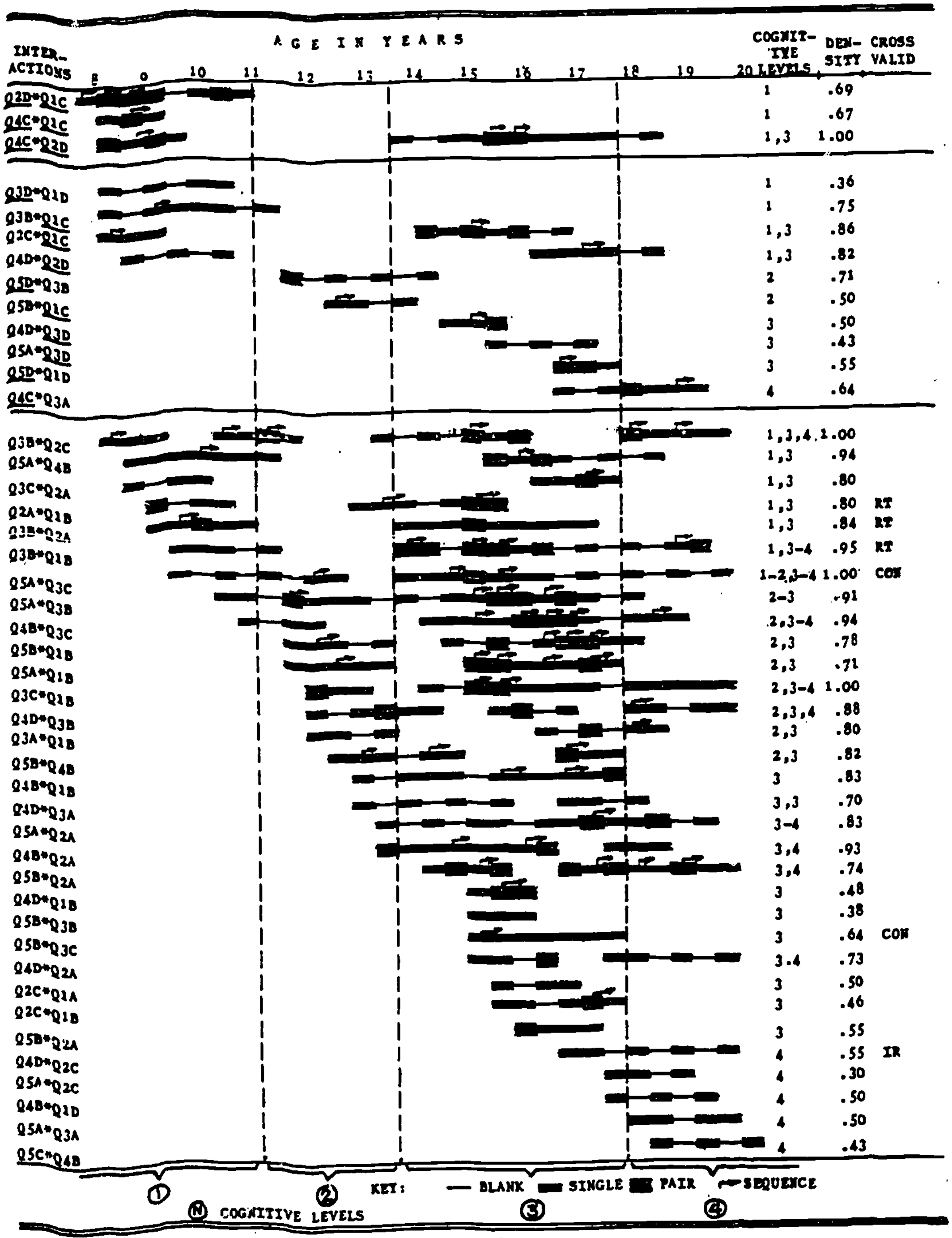
INSERT FIGURE 7 ABOUT HERE

The "stairway" pattern discussed above is clearly evident among many of these relationships. In the case of the W*W, many of the continuities either spread across or discontinously represent more than one level. Only a very few (5 out of 32) represent three or more levels and a larger proportion (about one third) only one level.

More than 86 percent of these sequences have a density of at least .50. That is, they account for at least half of the number of meaningful comparisons for that interaction. At a ration of 32 to 3, the W*W seems to be more important than the R*R and also at least as complex. Unravelling the hierarchal structure of this complex pattern will probably prove to be very challenging. Finally the marginal notations CON,RT and IRO refer to the classification of these comparisons in the earlier study (Powell, 1977), of the 9 which should have been present, 6 appeared. Only the classification OS, which should have had 3 meaningful relationships did not appear. In other words, two thirds of the homogeneous subgrouping of items have replicated from one sample to the second.

In addition, several wrong alternatives show meaningful sequences with two or more other wrong alternatives. It may be possible, if this

FIGURE 7
PATTERN OF CONTINUOUS SEQUENCES FOUND



INTERACTIONS	COGNITIVE LEVELS	DENSITY	CROSS VALID
Q2D*Q1C	1	.69	
Q4C*Q1C	1	.67	
Q4C*Q2D	1,3	1.00	
Q3D*Q1D	1	.36	
Q3B*Q1C	1	.75	
Q2C*Q1C	1,3	.86	
Q4D*Q2D	1,3	.82	
Q5D*Q3B	2	.71	
Q5B*Q1C	2	.50	
Q4D*Q3D	3	.50	
Q5A*Q3D	3	.43	
Q5D*Q1D	3	.55	
Q4C*Q3A	4	.64	
Q3B*Q2C	1,3,4	1.00	
Q5A*Q4B	1,3	.94	
Q3C*Q2A	1,3	.80	
Q2A*Q1B	1,3	.80	RT
Q3B*Q2A	1,3	.84	RT
Q3B*Q1B	1,3-4	.95	RT
Q5A*Q3C	1-2,3-4	1.00	CON
Q5A*Q3B	2-3	.91	
Q4B*Q3C	2,3-4	.94	
Q5B*Q1B	2,3	.78	
Q5A*Q1B	2,3	.71	
Q3C*Q1B	2,3-4	1.00	
Q4D*Q3B	2,3,4	.88	
Q3A*Q1B	2,3	.80	
Q5B*Q4B	2,3	.82	
Q4B*Q1B	3	.83	
Q4D*Q3A	3,3	.70	
Q5A*Q2A	3-4	.83	
Q4B*Q2A	3,4	.93	
Q5B*Q2A	3,4	.74	
Q4D*Q1B	3	.48	
Q5B*Q3B	3	.38	
Q5B*Q3C	3	.64	CON
Q4D*Q2A	3.4	.73	
Q2C*Q1A	3	.50	
Q2C*Q1B	3	.46	
Q5B*Q2A	3	.55	
Q4D*Q2C	4	.55	IR
Q5A*Q2C	4	.30	
Q4B*Q1D	4	.50	
Q5A*Q3A	4	.50	
Q5C*Q4B	4	.43	

observation proves stable, to use such biforkations to identify meaningful subgroupings of a sample or a population.

Another comparison which is illuminating is between the MULTIQUAL study and this present one. Figure 8 gives the distribution of the answers by age level on Item 1 of the Proverbs Test, with the raw data distribution on the same item (See: Figure 3a) superimposed. This comparison is also presented in Figure 8.

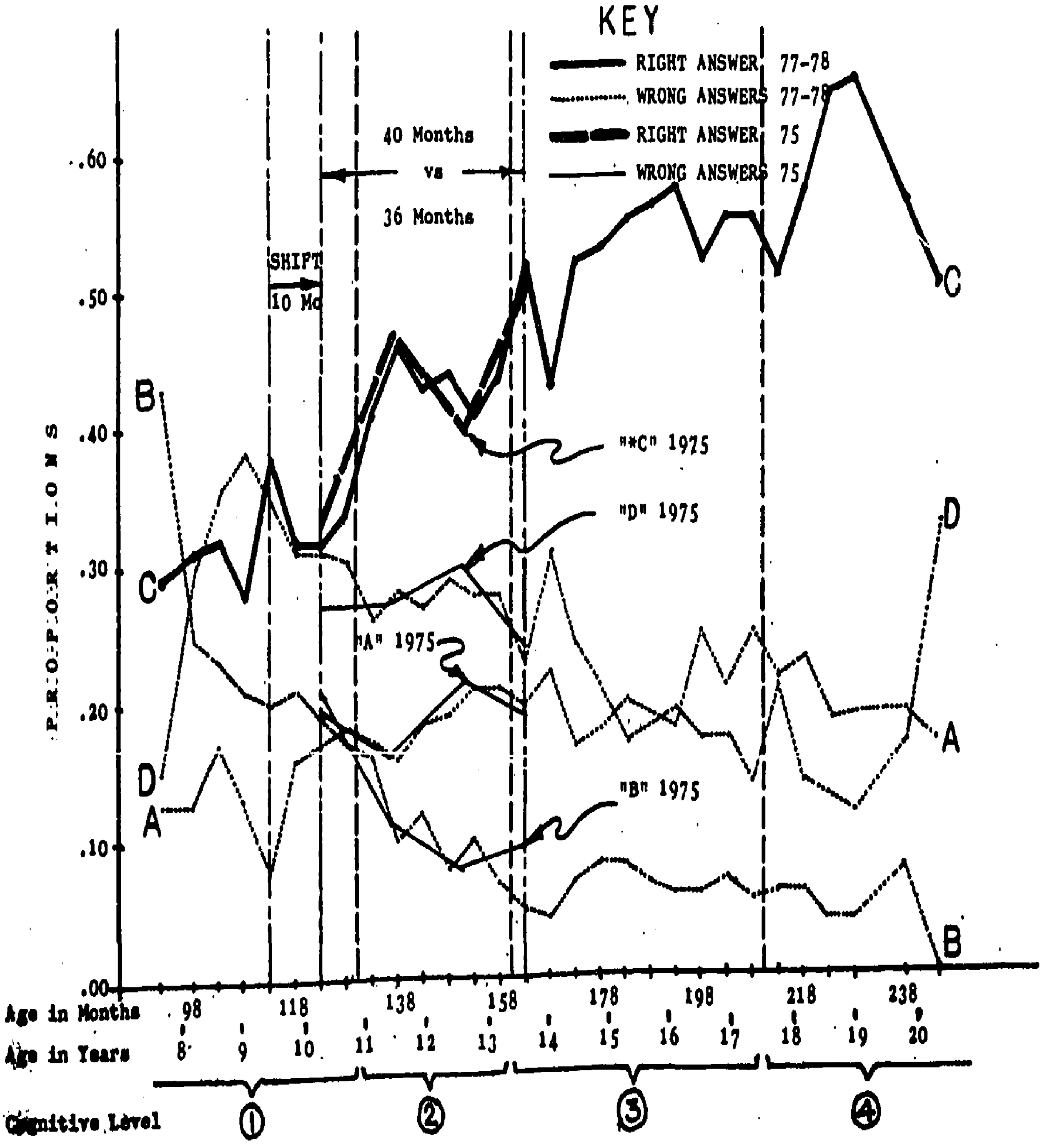
INSERT FIGURE 8 HERE

The fit is remarkable but it is not identical since it is necessary to shift the component curves to the right and to change their relative verticle placement and to expand the time interval. Their shape is not changed to get this fit. In mathematical terms, a phase shift, a change of amplitude, and a change of period are required to achieve this effect. Since the new sample is 50 percent city center and the original one was entirely suburban and this sample was collected in October/March of 77/78 and the earlier in May 75, a fit as good as this would be unlikely by chance. These data would seem to present a reasonably good replication if the position shifts are a legitimate procedure for such comparisons.

Are these positional shifts a Grade 5 event, a city center event, or a time of year event? If they are related to population mix then the patterns of answer interactions may be useful for identifying the response selection behaviors of specific subgroups in a population. In this case it may be possible to distinguish

FIGURE 8

ALTERNATIVE SELECTIONS ON ITEM 1 WITH THE PATTERN FROM 1975 OVERLAID



between cultural difference and mental retardation by using inter-alternative interactions.

The MULTIQUAL model version does not replicate nearly as well as the original proportions. This implies that although curved lines are necessary, quadratic and cubic curves may not be appropriate, a fact already evident when the extended range is presented as in Figure 8, and from the study of these same data conducted by Yu (1977).

Another feature of this pattern is that the right answers decline and the wrong answers increase with each of the three "transitions" which were identified from criteria independent of selection frequency. This observation supports the possibility of non-linear transformations may be a fundamental characteristic of development. This latter conclusion is further supported by the replication validation of these configurations mentioned earlier. In effect, irregularities in answer distributions may be meaningful and perhaps should not be removed, arbitararily, by replacing curves such as those found in Figure 8 by the "best fitting" straight lines. Straight line approaches to these data, such as Total Correct Scores may not be appropriate.

Also of interest is the appearance among the oldest learners of yet another drop in right answer frequency combined with a rise in wrong answer frequency. Does this observation imply that "Formal Operations" the high point for Piaget's theory, is not the last stage in development, as he assumes it to be?

DISCUSSION AND CONCLUSION

Of all the predictions made out the developmental patterns of answers, particularly among wrong answers all compa:ison supported a non-linear developmental pattern. Wrong answers remained generally superior to right one for meaningful information. It is important to point out once again that this information is related to the interactions between alternatives and is independent of selection frequency of particular alternatives. Selection frequencies may well be important as well. The point here seems to be that a large portion of the dynamics of the developmental processes may be lost when selection frequencies within items and particularly selection frequencies of right answers alone are the sole criterion for evaluation of learner performance.

Using the interactions among alternatives instead of cumulative frequencies it appears that learning may not be cumulative as the counting of right answers or addition of subscores implicitly assumes. Instead, development would seem to be a complex sequence of subtle transformations among thought processes which seem, from these data, to be strongly characterized by sequences of relationships and systematic relationship changes. Changes among the "wrong" answers appear much more frequently than among the right ones. This current study is now the third, using different age groups and tests in which wrong answers have appeared to be more "meaningful" than right answers. In the other two studies (Powell 1970, 1976) wrong answers were first to appear, and in aggregate accounted for the largest proportion of the total variance, when multiple regression procedures were used to predict independent achievement measures.

The most interesting of all these observations was the fact that W*W interactions were the most meanigful for the higher age levels. The implication drawn from this observation is that Knowledge level items may function adequately under the "know-guess" hypothesis and classical test theory may be appropriate for such items. However, for the so called "higher mental procees" items, the "know-guess" hypothesis may become increasingly invalid. This event may arise, because in a question requiring thought, there may be more than one reasonable answer. This possibility has already been explored with the 8 year olds discussed above.

The four questions posed earlier can now be answered. This analysis of the interactions among five out of 40 questions, represents 1/80th of the possible number of interactions on this test. Even this small portion of these data has already shown one statistically significant discrimination using a 5 month age interval over an age range in excess of 12 years. Such fine tuning has generally proven to be impossible for a group test of so few items using Total Correct scores to assess learner status.

First the patterns of answer selection across age suggest that learning may involve a complex hierarchically ordered sequence of interactive (non-linear) events.

Second, it would appear that, where "higher mental processes" are concerned, process may possibly take precendence over product. A stronger picture of the dynamics of development seems to emerge when the within event frequencies are suppressed in favor of between event interactions than when frequencies are considered by themselves.

Pattern analysis would seem to be much more meaningful than frequency counts.

Third, Total Correct Scores are probably not adequate for the evaluation of learner progress at levels above simple recall.

Fourth, for the "higher mental processes" the "Cumulative Learning" hypothesis would seem to be refuted. In any case, since a single contrary event is necessary to refute an "all" hypothesis, it can now be conclusively said that not all learning is cumulative.

It is now possible to address the question raised in the title of this paper. Developmental information is indeed available from wrong answers. Apparently, Total Correct scores by themselves are insufficient to describe the developmental status of a learner. Not only is it necessary to know how many answers of specific types are chosen, but which answers and perhaps even why these were chosen. There is evidence not reported here (See Powell, 1978b) to suggest that this latter information may be available indirectly from a parallel administration of a personality measure.

It appears, particularly for higher process tests that counting the right answers oversimplifies the situation, either ignoring or obliterating critical information. An approach to test interpretation which considers the pattern of particular answers (both right and wrong) selected would seem to be necessary to determine the developmental status of a learner. It also appears that a single administration of a content oriented achievement test may not be sufficient to this task. How a problem was solved may be directly available from the answer chosen but why it was chosen may require more information. This information is obtainable from interview (Powell, 1977) or from self-report (Powell, 1968). It may also be available indirectly from the parallel administration of personality meansures (Powell, 1978b).

In any event, analysis of test results on an answer-by-answer basis seems to be more meaningful than analysis on an item by item basis. Item by item analysis would seem to be more meaningful than any form of aggregation and particularly more meaningful than Total Correct scores. The fine tuning which may be possible using pattern analysis could hold considerable psychometric promise.

Although the current sample is large and the findings strongly indicitive in these particular directions, there are still many unanswered questions. It may be true that wrong answers are more "meaningful" than right answers without this conclusion being particularly useful to educators. Of what benefit might wrong answers be to educators if they were to try to use them?

Are the patterns found here characteristic of all children or only of this locality? Would the same pattern replicate in New York or Terra Haute or London, England?

If the irregularities in these curves are meaningful as well as the general trends what do these irregularities mean? Further exploration of the current data set which contains a personality test and is a repeated measures design, could possibly throw much light on these issues. In fact, it is already doing so, but this is the subject of another paper.

The most intriguing aspect of the results of this current study to the educator and to the measurement aspect is the implication that only Knowlege level questions may be stable under the "know-guess" hypothesis. Does this observation imply that Total Correct Scores may be valid only for recall and direct recognition items?

Does this implication also mean that all studies which used total correct scores for basic data (analysis of variance studies for instance) where more than recall was the concern, will need to be reworked?

Have educators, in pursuit of "right" answers forced inainity and triviality upon the learning process in order to get stable test results? Have these educators been aided and abbetted by measurement theorists who have built their theories upon the random normal variate model and upon the Total Correct Scores model from classical test theory? Has this problem been further compounded by the testing experts who have built their standardized tests and normed or outcome referenced these tests based upon Total Correct Scores, and/or total subtest scores?

Specifically, has the nature of the observations being made in educational measurement -- namely, the aggregation of frequencies of one arbitrary class of events -- prevented us from observing the most important learner-environment transactions in the learning process? Has this restriction in the observational framework employed effectively restricted educational outcomes to the recall of inainities and trivialities in order to achieve stable test results? Are the claims of critics correct -- for the wrong reason? Does this possibility explain many of the puzzling outcomes in educational research?

For instance, do these observations explain the low level of success on intelligence measures often experienced by individuals displaying subcultural differences? Do these findings help to explain why adults tend to do less well on intelligence measures than do adolescents? Do they help to explain why the profound thinkers tend to be less successful at formal schooling than more convergent thinkers? Do these outcomes help to explain why the expected differential effects from different educational interventions have failed to appear with any consistancy?

Do the subgroupings reflect cultural variables, learning style variables and the like in such a manner that it might be possible to get a better match between learner characteristics and teaching strategy than is now possible?

Would educational procedures focussing upon how people solve problems be more motivating than telling learners the solutions others have found? Could educators improve their ability to track learner progress using answer pattern analysis over Total Correct Scores? If so would this improvement be enough to warrant the extra difficulties involved obtaining this additional information.

In any case, the patterns among answers would seem to be far more complex than a Total Correct Score either implies or provides information about. A strong but complexly interactive developmental pattern seems evident. Test theory and evaluation procedures may be back to square one, but this time the outcomes from the available information would seem to be profound rather than superficial. The possibility of determining how learners attack, problems, of identifying learner subgroups and of getting considerable status information from relatively short tests may emerge from these findings.

In part, the current observations seem to explain why current measurement practice seems to produce superficial results, by implying that Total Correct scores may, at best, be superficial in evaluating learner progress, and at worst may be invalid to that process because an inappropriate mathematical model may be being employed.

Much more research is needed into this new area before the details of the ramifications of these findings are clarified. However, this present study may well be a good begining.

In this case, where do we go from here? It is already apparent that there are too many meaningful interactions among wrong alternatives within one item for a unique classification of the whole group. The multiple representation of alternatives within items would imply that alternative interactions may identify subgroups as well as positions in the developmental sequences. Further exploration of this problem is needed before some form of scoring procedure or pattern analysis proceudre can be developed which will extract the useful information from between alternative interaction in a manageable form.

Using Item No. 1 only it appears that relationships between wrong alternatives and personality variables are much more common than between right alternatives and these same variables. If this situation persists throughout, perhaps personality factors can be identified which may help to distinguish between the subgrouping of subjects mentioned above.

Once a clear basis for some form of scoring procedure or answer pattern descriptions which distinguishes among subects has been found, attempts can be made to predict the patterns in the second administration from the patterns in the first.

If "good enough" predictions can be established (say $r \geq .707$) then patterns in the second administration may be matched with the next higher age level in the first administration. Hopefully, this "leap frog" approach may reveal fairly clearly definable developmental pathways. These pathways might persist through several age levels; perhaps even to points of school exit (about which data are available).

The branch points and other critical characteristics of these pathways may be determinable. Once the developmental pathways (if such can be found) are mapped, attempts can be made to use this same instrument package with new subects about whom additional information can be obtained. The impacts of various intervention procedures upon pathway progress could, at this point be studies in depth. Penetration into the critical and/or central aspects of teaching/learning interactions may be possible. Hopefully, the findings from this data base using the proceudres employed in the present study can actually be pushed as far as this speculation proposes.

Judging from the fact that most age levels in this study are represented by at least one commencement or termination of a sequence, perhaps a response pattern analysis on this 40 item test could identify meaningful changes over relatively short time intervals. One 80th of the item interactions on this test have produced more meaningful sequences (32) than age levels (30) used in this study. There are nearly three fourths of a million potentially meaningful alternative-by-alternative interactions on this one test if using both administrations of 5 month age aggregates. If current ratios continue, throughout, much accuracy of placement and tracking may be well within reasonable possibilities.

The ultimate goal is to try to identify the impacts of intervention procedures upon developmental process outcomes. Since wrong answers seem to be a powerful source of process information, and inter-event relationships seem to add information to unrelated frequency aggregates, this approach would seem to hold promise. Could "effective teaching" get a better than current definition in this way? Could learning be improved, and by how much? These are all questions which may be answerable, at least in part, from the alternative procedures described in detail in this paper. All or most of these problems might be attacked from this present data set.

This research plan is the direction the present author intends to proceed. Anyone who is interested in pursuing any part of this complex problem using the data set employed herein is welcome to do so for the price of a computer tape, some postage, perhaps some phone calls and a willingness to share findings. Please join the team.

REFERENCES

Bloom, B. S. (Ed.), A Taxonomy of Educational Objectives: Handbood 1; Cognitive Domain, David MacKay, New York, 1956.

Bock, R. Darrell, Estimating Item Parameters and Latent Ability When Responses are Scored in Two or More Nominal Categories, Psychometrika, 37, (1), (March 1972), pp 29-51.

Bock, R. Darrell, MULTIQUAL: Log-linear analysis of nominal or ordinal qualitative data by the method of maximum likelihood, National Educational Resources, Chicago, 1973.

Glaser, Barney G. and Strauss, Anselm L., The Discovery of Grounded Theory: Strategies for Qualitative Research, Aldine, Chicago, 1967.

Gorham, D. R., The Proverbs Test, Psychological Test Specialists, Missoula, Montana, 1956.

Hakstain, A. Ralph and Kansup, Wanlop, A Comparison of Several Methods of Assessing Partial Knowledge in Multiple-choice Tests: II, Testing Procedures, Journal of Educational Measurement, 12(4), 1975, pp 231-254.

Kansup, Wanlop and Hakstain, A. Ralph, A Comparison of Several Methods of Assessing Partial Knowledge in Multiple-choice Tests: I, Scoring Procedures, Journal of Educational Measurement, 12(4), 1975, pp 219-230.

Powell, J. C., The Interpretation of Wrong Answers From a Multiple Choice Test, Educational and Psychological Measurement, 28(2), 1968, pp 403-412.

Powell, J. C., Achievement Information From Wrong Answers, (Short Title), Unpublished Ph. D. Dissertation, University of Alberta, Edmonton, Alberta, May 1970.

Powell, J. C., Evidence for a Phase and Stage Developmental Sequence Derived From Response Patterns on Multiple Choice Tests, Paper presented to the Annual Convention of the American Psychological Association, Washington, D.C., September 7th, 1976, p 62.

Powell, J. C., The Developmental Sequence of Cognition as Revealed by Wrong Answers, Research report presented to the Annual Convention of the Ontario Educational Research Council, December, 1975a, Alberta Journal of Educational Research, 23(1), 1977, pp 43-51.

REFERENCES Cont'd

Powell, J. C., Wrong Answers on Multiple Choice Tests: Blind Guesses or Systematic Choices?, Research paper presented to the Annual Meeting of the Psychometric Society, Hamilton, Ontario, August 1978a.

Powell, J. C. Cognitive Development from Wrong Answers, Research paper presented at the 13th Congress of the International Association for Applied Psychology, Munich, West Germany, August, 1978b.

Powell, J. C. and Isbister, Alvin G., A comparison between Right and Wrong Answers on a multiple choice test, Educational and Psychological Measurement, 34 (3), 1974, pp 499-5C9.

Shuford, E. H. Jr., Albert, A., and Massengill, H., Admissable Probability Measurement Procedures, Psychometricka, 1966, 31, pp 125-145.

Walker, Decker F. and Schaffarzik, Jon, Comparing Curricula, Review of Educational Research, 44 (1), Winter 1974, pp 83-111.

Yu, Kenneth, Developmental Patterns of Multiple Choice Tests of Elementary Students, Unpublished major paper, Department of Mathematics, University of Windsor, June, 1977.