ED 194 623 TM 800 754

AUTHOR Powell, J. C.
TITLE The End of an ERA: REQUIEM for the GLM. RIP.
INSTITUTION Windsor Univ. (Ontario).
SPONS AGENCY International Business Machines Corp., Armonk, N.Y.: Ontario Educational Research Council, Toronto.
PUB DATE May 80
NOTE 40p.: Paper presented at the Annual Meeting of the Psychometric Society (Iowa City, IA, May 26-28, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Higher Education; Logical Thinking; *Multiple Choice Tests; Reading Tests; *Scoring Formulas; Statistical Bias; *Testing Problems; *Test Interpretation; *Test Theory
IDENTIFIERS *General Linear Model

ABSTRACT

Current Scoring practices for multiple-choice tests are rooted in early Associaticnist Theory and are based on a two-step procedure: (1) right answers counted as ones and wrong answers are zeroes, and (2) number of right answers form a total-correct score. The author contends that if either step is invalid, the use of the general linear model (GLM) after tests are scored would invalidate its results. Based on a series of studies, evidence is presented to indicate that the scoring procedures are incorrect. Invalidation on psychological grounds is based on the observation that wrong answer selection is caused by item interpretation which may lead to considering logically correct answers to be wrong. Statistical invalidation is based on the develcpment of new procedures for interpreting contingency tables. It is suggested that present scoring procedures are nct applicable for higher-order multiple choice tests and that even with low-level skills, these scoring procedures may be misapplied when the discrimination level of the test and the performance levels of the learners do not match. When systematic curvi-linearity is found, current scoring procedures are inappropriate until valid transformations for the data have been found and can be applied. (Author/MH)

The END of an ERA

REQUIEM for the GLH

RIP

by

J. C. Powell
Faculty of Education
University of Windsor
WINDSOR, Ontario
N9B - 3P4

ABSTRACT

This paper presents substantive proof that the scoring procedure in current use with multiple-choice achievement tests is invalid upon both psycholgical and statistical grounds.

This paper has been prepared for presentation at the Annual Meeting of the Psychometric Society to be held in Iowa City, IA, May 1980. The research reported herein has been supported by grants from the International Business Machines Corp. (IBM) and from the Ontario Educational Research Council (OERC).

The END of an ERA

REQUIEM for the GLH*

RIP

As the United States entered into the First World War, and Associationism was in its heyday, Otis was asked to prepare a set of easily administered quick-screening tests for the U. S. Army. We will remember that Pavlov's experiments were barely out of the news at that time, and Watson was getting cats to do amazing things in his puzzle boxes. It was natural for Otis to assume that learning involved the establishment of associations or "mental connections" between a stimulus and a response.

The rest is history. The multiple-choice test which he adapted, if not invented, to this purpose proved very successful. A new era in measurement technology had been launched.

In keeping with good associationist principles, he developed a scoring procedure in which the frequency of "right" answers was counted. The logic behind this procedure was that the right answers were assumed, from their design, to be correct associations. The wrong possibilities were to be "plausible" but were expected to be chosen by "trial and error" in the absence of the correct associations. In other words the respondant either KNOWNS the answer or GUESSES. Since these "guesses" were considered to be "blind" (the product of trial and error) it was assumed that no meaningful information would be available from them.

In addition to this, since the average pattern across all wrong answers was expected to be random, there was also expected to be a certain number of the "right" answers which represented "lucky guesses." In this case, it would be impossible to determine which particular right answer

* GLH refers to the General Linear Hypothesis.

was meaningful, and which was not; so that item responses could not be interpreted. Only some sort of accumulation of responses could be useful, when attempting to assess learner status.

The result of these assumptions about the way in which the respondant to the test would behave when taking the test, was a two step scoring procedure:

STEP ONE: Scoring the Items

In this step the procedure was to make a pass through the test and to compare the respondant's answer with the one keyed in the predetermined control pattern as being the RIGHT answer. The respondant's particular answer was then changed to a "one" (1) for a match, and a "zero" (0) for a mismatch. If a "correction-for-guessing" were to be used then "omitted" answers were left BLANK.

Mathematically:

(1) $x_{ij} = \delta\ (1,0)$ where x_{ij} is the actual response of the ith individual on the jth item and $\delta\ (1,0)$ is the resulting binary conversion.

STEP TWO: Scoring the Test

In this step the procedure was to add the vector formed in step one for each respondant. When a "correction - for - guessing" was used, this was the third step.

Mathematically:

(2) $X_i = \sum_{j=1}^{n} \delta_{ij}$ where n is the number of items on the test.

This information is very well known, but is repeated here for several reasons. First, it should be noted that a good translation of the Associationist model into mathematical terms would conform precisely to these procedures. Second, it is well known that whenever the GLM is applied to test data, step one is almost always applied FIRST, ie. BEFORE any work with the GLM is attempted. Third, this second step is also generally applied BEFORE the use of the GLM, whenever analyses other than item analysis is contemplated. If there has been a predetermined subtest then equation (2) is applied to the scoring of this subtest before further analysis as well. Fourth, it is also well known, and logically obvious that if either of these two scoring procedures can be shown to be invalid, then this demonstration also INVALIDATES the use of the GLM after these precedures have been applied.

Finally, if step one is invalid, then the GLM can not be legitimately used upon the basic data set, because it is a nominal (categorical) scale, and the assumptions of the GLM require an INTERVAL scale to be fully functional.

Most psychometricians would agree that the above two and a half pages are obvious, and as such are probably not worth repeating. I would agree, except that I now propose to show that BOTH step one AND step two may be INVALID upon both psychological grounds AND upon statistical grounds. I have repeated these "obvious" facts precisely because they are obvious to the point that we often pay little attention to them, and to stress the fact that invalidating these two steps is equivalent to invalidating the USE of the GLM once these scoring procedures have been applied.

Of course, once we have some other procedure of transformation which validly converts these data into interval scales, then the use of the GLM becomes valid once more. However, if this alternative procedure accounts for nearly all of the available explainable variance by itself, then the use of the GLM would be valid, but UNNECESSARY.

It is within this context that I am arguing that we have come to the end of an ERA. It is indeed possible that the use of the GLM upon test data AFTER these two steps have been applied has yielded so many ambiguous and null results simply because these two steps removed so many of the available discriminations and so much of the available variance from the data set that little was left for the most powerful procedures to find.

In order to explain why such a possibility was not discovered before, we need only realize that curvi-linear phenomena yield strange results under linear analysis. It is desirable to avoid curvi-linear problems, if possible, since unlike with the GLM, a general solution for these problems does not exist. The few attempts which were made to include all answers, have all been "one shot" attempts which have not clearly pointed in any direction. It is quite reasonable to find people abandoning such attempts with more pressing problems at hand.

I have an unusual advantage over most researchers in the respect, because as a secondary-school classroom teacher I used test analysis on my tests for years before becoming a researcher. I therefore already knew that wrong answers had diagnostic value from practical experience long before I became interested in the statistical and the other psychological properties of this part of the data set. As a result, when the diagnostic properties did not jump out at me from my linear analyses of these segments of the distribution, I had the experience based motivation to persist.

In Search of the Esoteric

The term "esoteric" means "hidden." And the properties of answer selection distributions which I have been seeking have certainly been well hidden. At first glance, it would appear that I have been trying to process the "noise" in the system. In fact, this possibility has been raised upon several occasions. My rejoiner has been that it is <u>not</u> a property of NOISE to show consistant patterns with different tests and different age groups across what is now approaching a dozen studies.

What is even more interesting is the fact that although my findings seem paradoxical to the researcher, inclining him or her to discount these results, these findings do not seem so strange to the practicioner. In brief, I have found that "wrong answers", those strange little fellows we have been turning into "nothing" these many years, are actually more meaningful that the "right" answers we have been using to determine the achievement status of learners!

How I finally came to this paradoxical conclusion is too long a story for this paper. I propose, therefore, to try to capture the flavour of these events by extending the study I reported to the NCME in Boston last month (April 1980).

To begin with, I did not begin to get clearly definitive results until I returned to the disaggregated basic data patterns. It was only from the cross-tabulated contingency tables of the relationships between item pairs and single item repetitions that the underlying properties of these distributions became clear. Earlier efforts using wrong answers as dummy variables in multiple regression equations showed the superiority of wrong answers, but did not show the SAME wrong answers to be more meaningful in the different studies using a variety of age groups. Not only were wrong answers consistantly better, but also they were consistantly inconsistant. Shakespeare once said, "A pox on both your houses." which was precisely the way I was feeling until I hit upon the use of the contingency tables.

These tables present a serious problem for interpretation, however. We can easily determine whether or not a table is homogeneous from the size of its aggregate χ^2 wherein we determine the expected values from the marginal proportions. A non-homogenious table would reflect non-linearity. Too many more than a random number of such tables in a data set, and the data itself is non-linear and inappropriate for use with the GLM.

However, where do we go after establishing non-homogeneity? If the non-homogeneity is not consistantly in particular rows or columns, then we are in very real difficulties. The only other place for this departure from linearity to occur is within the cells of the table themselves. The critical values for cell χ^2 values are indeterminate because of a zero in the denominator of the equation.

To get around this problem, my colleagues and I (Powell, Shklov and Rahim, 1980) developed an alternative approach. Using the same assumptions as Otis for the generation of a data set, we simulated the data we were using in these current studies. The regression lines for each item was assumed to be the trend for selection for that item, and the standard deviation for the scatter of observed points about that line was assumed to be the measurement error in that item. The regression value for an age level ± a random value distributed as error* was set as the probability that this answer would be "right". Wrong answers were distributed as equiprobable across the difference between p_R and one. A second random number, rectangularly distributed between zero and one determined the simulated answer. A data set which duplicated the original in age group level frequencies was then generated, and the contingency tables for the simulation were struck. The frequency distribution for averages of the frequencies of cell χ^2 values was determined. The cumulative proportions for the final frequency polygon were used for the critical values for cell χ^2. A Monte Carlo approach to verify these values has been conducted. These values are reported in the work cited, but the more important ones are : $p = .10$; $\chi^2 = 1.4$; $p = .05$; $\chi^2 = 2.1$: and $p = .01$; $\chi^2 = 3.8$. These values also closely approximate the extrapolated values we could obtain from standard

* These random numbers were normally distributed with a mean of zero and a standard deviation of the assumed s_e.

tables.

From this combination of simulation and Monte Carlo procedures we generated a useful new tool for the interpretation of the cells in a contingency table. We could now look for meaning where it had previously been unavailable.

Comparisons by the Million

To give you some idea of the scale of these studies, I gave a 40 item test in reading comprehension (The Proverbs Test, Gorham, 1956) to more than 4000 students in the age range from 7 years to 20+ years. The test was administered twice with a 5 month gap between administrations. This procedure netted me nearly 3000 students who had taken the test twice. By dividing the age range into intervals of 5 months on each child's age in months, I obtained 30 age levels with an average of about 100 in each group. The 5 month grouping had two purposes. First, it would not matter how I blocked or combined these subjects, I would not get duplicate representation in any grouping. Second, the 5 month interval represented half of a 10 month school year.

Table 1 gives the break-down of the subgrouping of the sample.

INSERT TABLE 1 ABOUT HERE

The sample presents a representative selection of the schools in an urban industrial city in the mid-West of about a quarter million population.

With 30 age levels, 2 administrations, and 40 items, and two other tests with which to make comparisons, the 4 x 4 contingency tables will generate millions of cells to examine. As a result, only a small inroad into the properties of these data have been made since 1977-'78 when they were collected.

As already indicated this large scale study was mounted because earlier indications of the value of wrong answers and of the possibility of an underlying curvi-linear pattern in these data distributions warranted trying to get enough data points to plot some patterns.

Thus far there have been three major cuts into the mass, the first one (Powell, 1978) reported the fact that replication between this sample and a previous one from 1975 (with 550 subjects) was found. The replication, though not yet complete enough to be inequivocal, was strong enough to suggest that what is being reported here represents general properties of the test as well as the specific properties of this particular sample.

The second study (Powell, 1979) considered the interactions among the first 5 items on this test across the age range. Considering a "meaningful" interaction to be related to a cell χ^2 of greater than 2.3, two types of interaction were expected. The level of 2.4 minimum was conservatively based upon extrapolation before the simulation study. These two types of interaction were; when the observed frequency meaningfully exceeded the expected frequency, and the reverse of this relationship. Where $0 > E$ an event similar to a +ve correlation was assumed and the interaction was considered to be "joint." The reverse was "mutually exclusive."

Making the same assumptions made by Otis, several hypotheses could be considered. Right answer by right answer interactions should balance toward the joint type (which they did; 60 to 0 out of 600 possible), should be more frequent that 5% of the events (they were 10%) and should increase with age as less guessing occurred (they DECREASED with age). The right by wrong answer interactions should be exculsive rather than joint, and otherwise show the same patterns as for R * R. They were found to be mutually exclusive (12 to 220 out of 3600); represented 6.44% of the events; and did NOT change with age. Wrong by wrong interactions, taken as random events, should show about equal numbers of joint and exclusive events (joint events were FAVOURED 633 to 39 out of 5400); should be less that 5% of the possible (they were 12.44%) and should show no age pattern (they INCREASED in frequency with age).

As we can see from these paradoxical results, the events related to the right answers generally support these hypotheses, but the ones related to the wrong ones completely REFUTE them. Clearly wrong answers show systematic properties beyond the level of being "noise" in the system. It is equally clear that an approach to this problem which considered only the right answer relationships would not reveal this fact. A researcher would probably explain this DECREASE in R * R interactions as a property in the increase in the frequency of right answer SELECTION, and let it go at that point. With the joint W * W pattern exceeding the R * R interactions by more than 10 to 1 (633 to 60) and the R * W events near the chance level, we can begin to see both how much is being lost when we convert the wrong answers to zeroes and why this has been missed in other studies.

It is now clear that the way to derive the distribution properties of answer selection is to use a large sample, segmented cross-section contingency table analysis. The crucial part of this analysis involves some cell-by-cell interpretation and comparison, for which the obtaining of the critical values for cell χ^2 seems to be essential.

Before going to the third study (Powell, 1980) of which this one is an extension, we should turn briefly to the psychological properties of answer selection.

Paradoxes
Paradoxes
Paradoxes
Paradoxes

The first analytic study I conducted into the area of wrong answer selection (Powell, 1968) involved this same test being used here, except with college Juniors and Seniors. The right answers showed a strong single factor with good separation using principal components analysis, so that I was certain that I was using a "good" test before I began. In addition to collecting answers, I asked these students to explain their selections in a separate booklet. I used their reasoning to classify the four wrong answer factors, which showed simple structure, which I studies further. For the most part, these wrong answer selections reflected reasoning errors such as linking only part of the proverb to a translation of it (Over-simplification.) The general quality of the reasoning from one group effectively described the reasoning within all members of a factor in another class of students about two thirds (.64) of the time. As a result, the diagnostic use of wrong answers I had witnessed as a Math/Science teacher in secondary school seemed to be present at the college level in this language-based test, but not as strongly as expected.

The surprising finding was that one type of "wrong" answer (Irrelevancies; defined as true statements unrelated to the problem) I found clear evidence of multiple reasons for the selection of the same answer. Those students in the middle of the range by total-correct score indicated their choice was based upon the truth of the statement. There was also a smaller group, who scored in the 80th percentile range, who interpreted the question differently, but legitimately, and chose this so-called Irrelevancy upon a basis which was logically CORRECT, once we admit the appropriateness of their alternative interpretations. Here was evidence that some "wrong" answers which should be considered RIGHT, and which identified those who had "over-read" the question.

To probe this issue further, I collected the reasoning, using trained interviewers, from about two thirds of the 550 children (ages 8 to 16) to whom I gave this same test in May of 1975. I used a non-parametric procedure to develop a set of twelve wrong answer subtests, and then classified these using the reasoning protocols. This classification was supported by between 50% and 60% of the reasoning reports for all 12 subtests. This test, by the way, has two "right" answers scores. One for "concrete" and the more usual one for "abstract" answers. As such I hoped that the transition between concrete and abstract reasoning in children's verbal recognition patterns might be illuminated.

I used a simplex approach to ordering these subtests: that is, the more closely related subtests were to each other the nearer they were in the resulting sequence. The big surprise was "no surprise." All 14 subtests arranged themselves into an order which reproduced the age sequence without exception. This is the closest I have come to seeing a perfect correlation in all of my years of working with live data. A very strong developmental influence seemed to be present among these data. No stretch of the imagination could conjure this observation as "noise" in the system.

When I looked at the interpretations, a sequence strongly reminiscent of Piaget's accounts was present. Many of these wrong answers displayed CORRECT REASONING once we took into account the DEVELOPMENTAL PERSEPECTIVES in their cognition of these learners. These answers were NOT WRONG when the alternative world-view which characterizes their developmental stages is taken into account. Figure 1 below illustrates this point.

INSERT FIGURE 1 ABOUT HERE

Of the three influences (interpretation, procedure and information content) the order of precidence seemed to be as just given.

The eight year olds who chose alternative "C" seem to be interpreting the proverb (Quickly come, Quickly go) in terms of their own physical movement about the classroom. Their typical reasoning (that's what the teacher always says) links the concept of "being quick about it" quite accurately to the answer they selected (Always do things on time). They have apparently not yet decentered and are still interpreting such statements in terms of their personal experiences. Once we know what is happening, the connection between the choice of answer and the reasoning behind it becomes quite clear. These linkages are NOT "trial and error" in any sense, random or not, but come directly from the emerging logic of the learner.

I could continue to make the same point with each of the other "wrong" answers, but such would be redundent and has already been done elsewhere (Powell, 1977). These observations when contrasted with the proposal attributed to Association Theory at the beginning of this paper (learning involves forming "correct" associatons and that in their absence "trial and error" will be observed) leave little doubt about the psychological invalidity of that theory. QED.

It would not be unreasonable to expect to find the mathematical translation of a behavioural theory to be behaviourally invalid if the theory itself is shown to be invalid. Hence the disappearing Paradoxes in the title of this section.

We can now return to the mainstream of our discussion about the distribution properties knowing that an effective analytic procedure should expose sequences of answers rather than merely a transfer to "right" answers, and along with this there should be some sort of emerge-decline pattern of curved-line events to account for these transitions. We also need to find some reason why these patterns have not been more clearly evident before this.

The Silent
Jiggle

The third study in this series (Powell, 1980) also proved to be the most profitable to date. In this one I cross-tabulated the within-item pre- post- events for all of the learners who had taken the test twice. My purpose was to try to determine the patterns of change which occurred among answer selection. In this case an 0 > E event would be a stable selection, with more people than expected giving the same answer upon both occasions (for the events in the principal diagonal), and unstable for the reverse relationship. The change events would be shown by the patterns in the off-diagonal cells.

The major findings were that stability overwhelmed change and the wrong answers were significantly more stable than the right ones with this stability increasing with age for the wrong answers, and decreased for the right ones. This observation suggests that an important characteristic of development is increased diversity!

However, I am a bit ahead of myself. Before we start to consider the patterns among the cells, we need to satisfy ourselves that there is enough curvi-linearity among these data to trouble about. Since about two thirds of the 1200 contingency tables were NOT homogeneous, this issue was dispensed with quickly.

The balance of our discussions will center upon Item 18 since it was this Item which I reported in Figure 1 (page 12) so that we already know the expected order to be found from the psychological sequence (C → D → A → B*).

Because stability overwhelmed the off-diagonals, I will deal with these issues first. Among the 1200 tables to be considered there were 4800 diagonal cells of which 1200 were repeated choice of the right answer and 3600 were for repeated choice of the wrong answer. Of these 4800 more than 40% (or 1965) had χ^2 values which exceeded 2.0. Only one of these was a significantly unstable event. Of the stable events, the stability of the wrong answers exceeded that of the right answers by a factor of about 3.5 to 1 (or 1527 to 437; differences between proportions $z = 3.75$) and as a result most of the off-diagonal significant events had $0 < E$ as avoidances rather than changes.

If we look only at Item 18, as shown in Figure 2, the pattern of stability becomes clear.

INSERT FIGURE 2 ABOUT HERE

The verticle arrangement in Figure 2 is the "psychological" order given above. The number of horizontal lines indicates the level of significance with one line for $p \leq .10$, two for $p < .05$ and three for $p < .01$. Thus we can consider this horizontal density to reflect degrees of stability from "stable" to "extremely stable". Values less than this will be homogeneous and will indicate that the marginal frequencies are sufficient to account for the choices, and that the frequency of repeated choice on the pre-test should be considered to be independent of this same frequency on the post-test.

If we consider only the "wrong" answers (the lower three) for the moment, there is a visually evident progression of the density of stability from left to right (with increasing age). The persistance of alternative "C" is a bit surprising as is the long extremely stable period for alternative "D". It should be mentioned that although the age scale is in years, these data actually represent 5 month age blocks, hence the length of the representative coding is inconsistant with the age scale.

Considering the right answers, item 18 is unusual since the majority of right answers have their stability to the far left. The extremely stable section from about age 15 on would, most commonly, not be there. It is also evident that there are not very many sections of the age range without at least some stability. It is clear that the internal dynamics of this item seems to be, in general, more meaningful than are the marginal (aggregate) frequencies. Later on we will see that the marginals actually supply different information than does these internal dynamics.

In order to overcome the overwhelming influence of the stability factor, I used a procedure which may prove to be equivalent to the procedure in factor analysis which remove the first factor's influence in order to find the second one. I simply dropped the diagonal frequencies and recalculated the χ^2 on the assumption of homogeneous diagonal elements as shown in Figure 3.

INSERT FIGURE 3 ABOUT HERE

In this Figure I show the original print-out table, and the "doctored" table below it to show the impact of this procedure. Notice that the off-diagonal frequencies are identical on both tables but that the locations of significant events has changed considerably. With the actual diagonal frequencies in place, two of the three significant events are avoidances and are in the "row/column" relationships to the most highly stable event. This observation suggested to me that the stability was "overwhelming" the change pattern. With the diagonals remove, reducing the total frequency from 96 to 63, the three significant elements are all changes. The procedure achieved its objective but the impact of the violation of the assumptions for χ^2 calculations is uncertain at this time.*

The fact that this second-order contingency table achieved the purpose for which the procedure was designed, may have been fortuitous, but it made me bold enough to try to use it for a third order level of analysis. I collected all of the changes which emerged for the 30 tables in Item 18 and arranged these frequencies into a "doctored" 4 x 4 table like the one just discussed. The resulting significant "changes" from this third order analysis recovered the psychological sequence which we already have seen

* There seems to be no mathematical problem with the cell χ^2, since this is merely an alternative model for the "expected" values. However, for the overall χ^2 this procedure forms in "incomplete" model which creates problems in the determination of the number of degrees of freedom.

from the logic of the reasoning and now have from statistical procedures. The full details are elsewhere (Powell, 1980) so need not be repeated here (both ERIC and the Library of Congress have copies).

For this present study, I have taken this analysis one step further. I found the average for the linkage points for the changes which were shown to be meaningful. From this information I prepared Figure 4.

INSERT FIGURE 4 ABOUT HERE

The number of lines in each arrow* follows the same code for the strength of change in Figure 4 as was used for stability in Figure 2 (page 14.) The vertical arrangement of responses was obtained by the average of the changes to that choice. It is identical to the psychological order from Figure 1. It seems to follow an accelerating upward pattern like the bottom of a growth curve. Each point in this curve seems to be associated with a period of stability of response selection.

The downward arrows are also interesting. The shift from "A" to "D" predates the reverse trend. Perhaps "D" is more powerful at this age. The very strong trend from "D" to "C" starts a period of stability (interrupted once in 20 months) which is then followed by a 20 month gap. This gap, which begins at age 16, coincides with the youngest legal school - leaving age in this system. The return to "egocentricity" ahead of early school leaving is an intriguing possibility which makes intuitive sense. If

* Two arrows are not shown in this Figure. In these two, although significant, they represented "avoidances" (O < E) not changes. These were "C/B" and "A/C."

supported when I process the "school - leaving" data, it would suggest clear signals that they may be "high risk" may be being sent out by these young people several, if not many, months ahead of time. Since this particular pattern involves a change from one "wrong" answer to another, using "zeroes" obliterates it. This information is clearly NOT AVAILABLE when only the "right" answers are considered.

The downward movement from the "right" answer at the top would also not be anticipated from the Associationist model, but if it represents those those who have progressed so far that they are beginning to "over-read" some questions, then this downward vector may actually be a reflection off of the ceiling of the test and may represent an upward continuation of development. This change results, when "right" answers alone are being considered, in a LOWERING of the learner's SCORE.

I did not try to put these two diagrams completely together since the visual complexity would have reduced the impact of these observations. It is quite clear that "wrong" answers seem to contain a good deal of developmental information which is not available from the "right" answers. These curved-line patterns, with reversing directions, suggest that development may follow multiple pathways. With this much complexity, it is not surprising that trying to reach into it from the one-sided direction of the "right" answers has not proven to be very successful.

The last straw with respect to the Associationist "know-guess" hypothesis comes from the next pair of observations.

In Hand
with GOD

Having looked at the statistically meaningful change patterns within these tables, I decided to explore two other sources for change patterns which were available from these data. One of them was the changes in the marginal totals for each alternative on the pre- to post- test transitions. The other was the changes in these same values from one 5 month age block to the next. Very strong support for this complexity found in the comparisons between these two changes would further invalidate current practice.

Figure 5 gives a schematic version of the first of these two change patterns, showing the general pattern with small irregularities removed, and Figure 6 does the same for the other pattern.

INSERT FIGURE 5 ABOUT HERE

These linear pathways were derived from an end-to-end set of vector drawings from the within-group pre- to post- aggregate changes. The pattern is actually somewhat more irregular since I have attempted to capture only the major trends here. (See: Powell, 1980 for the actual results.)

The right answers appear to be a "step function," with a spurt in the eleventh and the fourteenth years. The one in the eleventh has been noticed by several practicioners of my acquaintance, although I have not seen it discussed in the literature. The one at fourteen may be associated with the transition from concrete to formal operations discussed extensively by Piaget and others. The relationsip between the increase in right answers at age 11 and the decline in alternative "C" is puzzeling because "C" to "B" transformations is one of the two "avoidance" events.*

* The other one was "A" to "C".

The up-turn in both "C" and "A" at the extreme right are also of note. The one for "A" is probably related to the cascade effect from the right answers already noticed. The timing is off for the "D" to "C" change. Within-group aggregate patterns do not seem to coincide as closely to the internal item dynamics as we could expect.

When we turn to the between-group dynamics, yet another picture emerges. This pattern is shown in Figure 6.

INSERT FIGURE 6 ABOUT HERE

The surge-plateau pattern characteristic of the within-group dynamics has gone, to be replaced by the beta patterns found among the 1975 data as well. (See; Yu, 1977.)

Comparison on an event-by-event basis, thus, has shown that the within-group dynamics is significantly DIFFERENT from the between-group dynamics (sign test for inter-point direction equivalence; z = -2.60). The progression seems to be from cell-by-cell to within-group to between-group dynamics, giving an interpretation clue. The marginal changes for the within-group are actually a composite of three sub-groups (those who stayed, who arrived, and who departed).. However, with the marginal proportions REMOVED in the homogeniety comparisons for the cell-by-cell dynamics it is partly coincidental for the marginal and the internal changes to occur in the same direction.

In addition, it appears from these observations that development may go in more than one direction. In this case, populations may not be homogeneous, but have a complex sub-structure instead. From these considerations, it appears that aggregates may reflect sub-population mix rather than development.

In the replication study* it was clear that sub-population mix was important since the selection proportions had to be changed in the pattern fit. Also, the pattern from the suburban group of 1975 had to be raised in average age level by a full year to fit the community cross-section sample of 1977-'78. Since the two studies were an average of 30 months apart, both coming from the same community, the replication would seem to have been achieved from an averaging effect with the developmental dynamics and the sub-population mix. I would probably not have achieved as good a replication had differing communities been used, or had a longer time-span elapsed. Whatever strange coincidence of natural events occurred to produce these same configurations (although at different selection levels) for all four alternatives between these two samples we may never know.

This much seems to be clear, however, that the aggregates seem to be more sensitive to cross-sectional events from the sub-population mix, and the internal events in each item seem to be more sensitive to the longitudinal events related to the development of cognition and achievement. It is entirely possible that the use of the scoring procedure derived from Associationist theory has been removing from our data set the information we were seeking before we began our data analysis!

In keeping with good culinary traditon, I have saved the best for the last.

Out of

Nowhere

* The use of η to determine explained variance for fitted curves gave an explained variance in excess of .60 for each of the four curves separately. I do not know how to get the block-fitting of all four curves as a unit from these results. Perhaps the replication was as good as .80 explained variance.

We now turn our attention to the pattern of departure from homogeneity within Item 18, which produced a finding so startling as to all but confound the senses. Figure 7 gives the basic observations.

INSERT FIGURE 7 ABOUT HERE

There is nothing particularly out of the ordinary in this Figure, except that 17 of the 30 tables are non-homogeneous. Little doubt is left in this observation that complex curvi-linearity is present.

It is also evident (visually) that the major influence in the size of the departure from homogeneity is related to the group size which I indicate with the dotted line in the background. The next question would reasonably be, what would be the pattern if the impact of this aggregate were to be removed? Figure 8 gives this result, taken in two steps.

INSERT FIGURE 8 ABOUT HERE

In part A of Figure 8, I used the simple expedient of dividing each χ^2 value by its corresponding group size. This step seemed to generate what appeared to be a cyclical pattern. I assumed a two year interval, for a reason I will indicate in a moment, and sketched such a pattern as background, and inserted the center-line. It appeared from this procedure that the use of the linear transformation of dividing by the group size overcompensated for its effect.

Rather than trying to determine the exact transformation (perhaps a square root or a logarithm) I simply rescaled to make the center-line straight. Part B of Figure 8 gives the results. In the hypothetical cyclical pattern extracted by "removing" the impact of group size, the "eleventh year spurt" is clearly evident. There is also the possibility of some sort of second order oscillation since the oscillations seem to increase in magnitude to the right in two separate stages.

This pattern would not be particularly remarkable, and in point of fact could be "noise" in the system, had I not found a very similar pattern from an independent source. We must remember, that this pattern has the impact of the marginal frequencies removed when homogeneity was being determined. As a result, if a similar pattern is found using the marginal frequencies, then this pattern can NOT be noise.

In the study of which this present one is an extension, I reported an unusual observation. To begin with, the replication results pointed to the possibility that the irregular variations about the regression lines may not be "noise" as is typically assumed. To explore this possibility, I assumed these patterns to be multi-modal. I had a student (Alison Caird) tabulate the frequencies of the primary and secondary modes for both the right and the wrong answer selection proportions for all 40 items and all but the very lowest and highest age levels (where the group sizes were too small to be representative of the cohort). These counts reflect the marginal proportions, a factor which is removed when considering the degree of non-homogeneity. Figure 9 shows these results with the same cyclic phases superimposed.

INSERT FIGURE 9 ABOUT HERE

The heavy line is the pattern for the modes for the right answers, and the dotted line is the same pattern for the wrong answers, made more similar in appearance by using a 2 to one rather than a 3 to 1 for the abscissa. The surprise I had in the earlier study (Powell, 1980) was that the wrong answers lagged the right answer cycle by about 5 months rather than being contra-cyclic as would be expected. The surprise this time was that if we consider the right and wrong cycles as a single pattern, then the non-homogeneity pattern is closely in the opposite phase to the modes-of-modes pattern for half of the cycles. Assuming, as seems reasonable, that a fifth level of modes would most likely peak at age 11 for the right answers, and 19 for the wrong answers, this pattern may resolve even more closely. In modes-of-modes pattern, we once again see a tendency for the oscillations to increase in amplitude to the right.

It appears that these "independent" sources mutually reinforce the evidence of non-linearity among these data. The cyclic pattern seem to support Piaget's phase and stage "clinical" model, except that rather than converging upon "formal operations" as a unifying entity, learners seem to DIVERGE AS THEY LEARN TO THINK.

The END of
an ERA

I began this discourse by suggesting that multiple-choice tests and our current practice for scoring them arose from early Associationist theory, then in the forefront of psychological thinking about learning. These inventions instituted a new era into educational testing.

I also suggested how the twin concepts of "mental connections" and "trial and error" combined into the "know-guess" hypothesis to lead directly to current scoring practice. The mathematical translation of this latter hypothesis produced a two step procedure. In step one, the "right" answers were scored "one" as a "good association" and the "wrong" answers were scored "zero" as a "guess." Some of the right answers would be "lucky guesses" so particular answers could not be interpreted.

The second step counted the number of "right" answers to form a "total-correct" score which was assumed to reflect how much the learner "knew." The total was sometimes modified to remove the "lucky guesses." Hence current scoring practice.

If either or both of these procedures are invalid, then the use of the General Linear Model (GLM) AFTER tests are scored using these procedures would invalidate the results from the GLM. I then presented evidence which demonstrates both procedures to be invalid.

The invalidation upon psychological grounds was based upon the observation that the major contributor to wrong answer selection was item interpretation which frequently leads to the considering of LOGICALLY CORRECT answers to be wrong. Diagnostic and other information is also present among these "wrong" answers. Few such answers are "blind guesses."

Statistical disconfirmation, which required the development of a new procedure for the interpretation of contingency tables, was dependent upon several considerations. The psychological pattern for development which suggested that learners moved from one wrong answer to another before reaching the "right" one, was derived statistically from these data using these new procedures. Information about learners not available from the right answers thus becomes available.

Bock's (1972) study showing little gain from Knowledge level wrong answers, and these which show considerable gain for Comprehension level wrong answers (with other evidence showing even more gain at the Analysis level) seems to suggest that present procedures may be appropriate for test items requiring low level skills like recognition or recall. The observed increase in diversity as thinking skills increase suggests their misapplication for higher order tests. Other evidence suggests that even with low level skills, the current scoring procedure may be being misapplied when the discrimination level of the test and the performance levels of the learners do not match. Both of these shortcomings can be substantially reduced by the simple expedient of considering all answers in our interpretation attempts.

It appears that much useful information about learners and the learning process, including answers which are LOGICALLY CORRECT may be lost by scoring the "wrong" answers as "zero."

From replication attempts, from internal dynamics analysis and from within-group/between-group comparisons, it appears that the internal dynamics of an item, when full disaggregated, reflects the longitudinal and other developmental properties of learning and achievement, while aggregation of these data seems to reflect the cross-sectional properties of sub-population mix.

Current educational research seems to show that the cross-sectional properties of groups seem to overwhelm the longitudinal properties inherent in these data sets. Perhaps, however, the scoring procedure has been systematically removing these longitudinal properties from these data before the analysis. As just one example, research to date has seemed to show little advantage favouring one approach to teaching over another. However, once population diversity has been controlled, the internal dynamics approach may show considerable differential effect of a sub-group specific nature.

A final nail in the coffin of current scoring practice was driven in, when systematic curvi-linearity at a second level and perhaps even a third level was found and independently supported (to a degree) among these data. With so much curvi-linearity, current procedures (including the GLM) are not appropriare until valid transformations for these data have been found and can be employed.

We are indeed at the end of an ERA, not of multiple-choice testing, these are more powerful than experience seemed to show, but at the end of the use of a scoring procedure which has served us less well than we have thought these past 60 years.

The implications from this research can be summed up in one sentence. All of the educational research which has used the present scoring procedure BEFORE commencing other analyses will need to be reworked. Via con Deos! Rest in Peace!

REFERENCES

Bock, R. Darrell, Estimating item parameters and latent ability when responses are scored in two or more nominal categories, Psychometrika, 37 (1), March 1972, 29-51.

Gorham, Don R., The Proberbs Test, Missoula Montana, Psychological Test Specialists, 1956.

Powell, J. C., The interpretation of wrong answers from a multiple-choice test, Educational and Psychological Measurement, 28 (2), 1968, 219-230.

——— The developmental sequence of cognition as revealed by wrong answers, Alberta Journal of Educational Research, 23 (1), 1977, 43-51.

——— Wrong answers on multiple-choice tests; Blind guesses or systematic choices? Paper to Psytic, Soc. 1978, (ERIC)

——— Can developmental information be obtained from wrong answers? Paper to NCME, 1979. (In ERIC)

——— Patterns of change among answer selection, Paper to NCME, 1980. (In ERIC)

Powell, J. C., Shklov, N. N., and Rahim, M. A., A Monte Carlo approach to the determination of critical values for cell Chi Squares. Paper for Psychometric Society, Iowa City, 1980.

Yu, Kenneth, Patterns of multiple-choice tests of elementary students, Major Paper, Dept. of Math., Univ. of Windsor. Avbl from Powell, Faculty of Education, Univ. of Windsor. WINDSOR, Ont.

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	OVERLAP OCT./MAR.	GRAND TOTALS
1	AIM < 96		43		3	25	46
2	96 - 100	8	68	MOSTLY	32	32	100
3	101 - 105		70	SAME	55	40	125
4	106 - 110	9	130	GROUP	53	102	183
5	111 - 115		101		120	62	221
6	116 - 120	10	127		78	95	205
7	121 - 125		137		101	100	238
8	126 - 130		142		114	102	256
9	131 - 135	11	145		118	110	263
10	136 - 140		129		125	97	254
11	141 - 145	12	165		106	119	271
12	146 - 150		135		131	90	266
13	151 - 155		138		100	88	238
14	156 - 160	13	152		104	99	256
15	161 - 165		114		132	73	246
16	166 - 170	14	163		101	110	264
17	171 - 175		262		150	189	412
18	176 - 180	15	264		237	194	501
19	181 - 185		258		247	201	505
20	186 - 190		251		255	177	506
21	191 - 195	16	249		228	162	477
22	196 - 200		219		220	145	439
23	201 - 205	17	210		219	117	429
24	206 - 210		171		173	88	344
25	211 - 215		186		130	84	316
26	216 - 220	18	125		131	58	251
27	221 - 225		87		81	40	168
28	226 - 230	19	47		66	18	113
29	231 - 240	20	20		43	9	63
30	240 < AIM		10		14	4	24
	TOTALS —		4319		3676	2830	7995

FIGURE 1

AN EXAMPLE OF THE PSYCHOLOGICAL BASES FOR ANSWER SELECTION

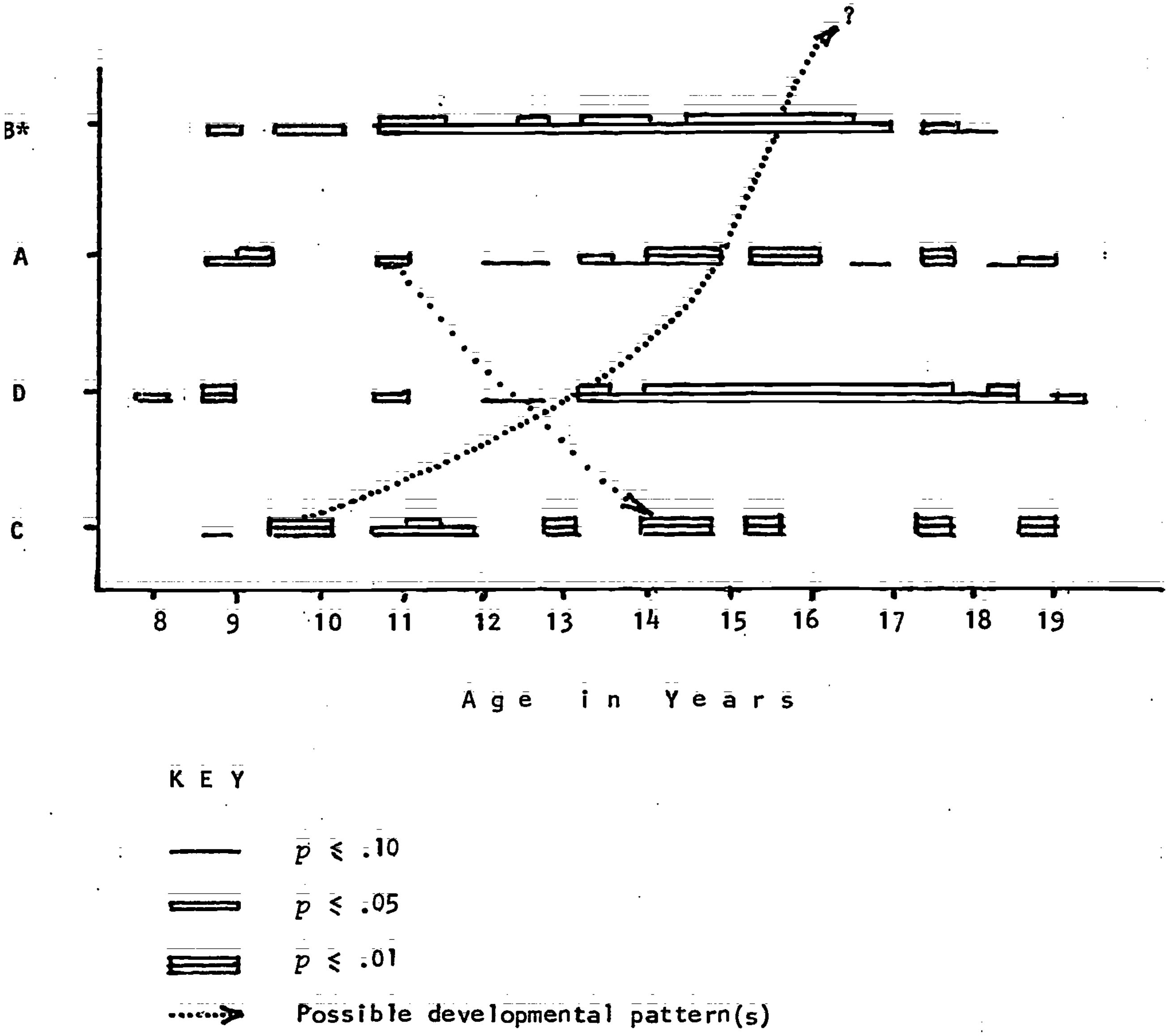
Proverb: QUICKLY COME, QUICKLY GO. (EASY COME, EASY GO.)

Alternative	Age of most Common Choice	Reported Reasoning
a. ALWAYS COMING AND GOING AND NEVER SATISFIED.	13	You should stick to a job 'til it's finished.
b. WHAT YOU GET EASILY DOES NOT MEAN MUCH TO YOU.	Adult	Keyed as the RIGHT Answer.
c. ALWAYS DO THINGS ON TIME.	8	That's what a teacher always says.
d. MOST PEOPLE DO AS THEY PLEASE AND GO AS THEY PLEASE.	10	It talks about coming and going.

Source: Item 18 from The Proverbs Test by Donald R. Gorham, Missoula Montana, Psycholigial Test Specialists, 1956. Reproduced with permission.

FIGURE 2

STABILITY PATTERNS
AMONG ANSWER SELECTIONS
FOR ITEM 18



AN EXAMPLE OF A PROCEDURE TO OBTAIN HIGHER ORDER RELATIONSHIPS FROM CONTINGENCY TABLES

BASIC DATA: Item 18 Age level 10 (136-140 Months)

FREQUENCY EXPECTED CELL CHI2		PRE					
		A	B*	C	D	TOTALS	%
POST	A	9 10.0 0.1	4 5.4 0.4	3 4.2 0.4	13 9.4 1.4[a]	29	30
	B*	5 10.3 2.7[b]	13 5.6 9.7[c]	4 4.4 0.0	8 9.7 0.3	30	31
	C	3 3.8 0.2	1 2.1 0.5	4 1.6 3.6[b]	3 3.6 0.1	11	12
	D	16 8.9 5.6[c]	0 4.9 4.9[c]	3 3.6 0.2	7 8.4 0.2	26	27
TOTALS		33	18	14	31	96	
%		34	19	15	32		100

OVERALL $\chi^2 = 29.0$; $p < .005$; $df = 9$

"Doctored" table with main diagonal removed.

FREQUENCY EXPECTED CELL CHI2		PRE				
		A	B*	C	D	TOTALS
POST	A	////////// ////////// //////////	4 1.6 3.6[b]	3 3.2 0.0	13 7.6 3.8[c]	20
	B*	5 6.5 0.3	////////// ////////// //////////	4 2.7 0.6	8 6.5 0.3	17
	C	3 2.7 0.0	1 0.6 0.3	////////// ////////// //////////	3 2.7 0.0	7
	D	16 7.2 10.8[c]	0 1.5 1.5[a]	3 3.0 0.0	////////// ////////// //////////	19
TOTALS		24	5	10	24	63

OVERALL $\chi^2 = 21.2$; $p < .005?$; $df = ?$

NOTES: a. $p \leqslant .10$; b. $p \leqslant .05$; c. $p \leqslant .01$.

FIGURE 4

WITHIN - ITEM DEVELOPMENTAL PATTERN
FROM CONTINGENCY - TABLE ANALYSIS

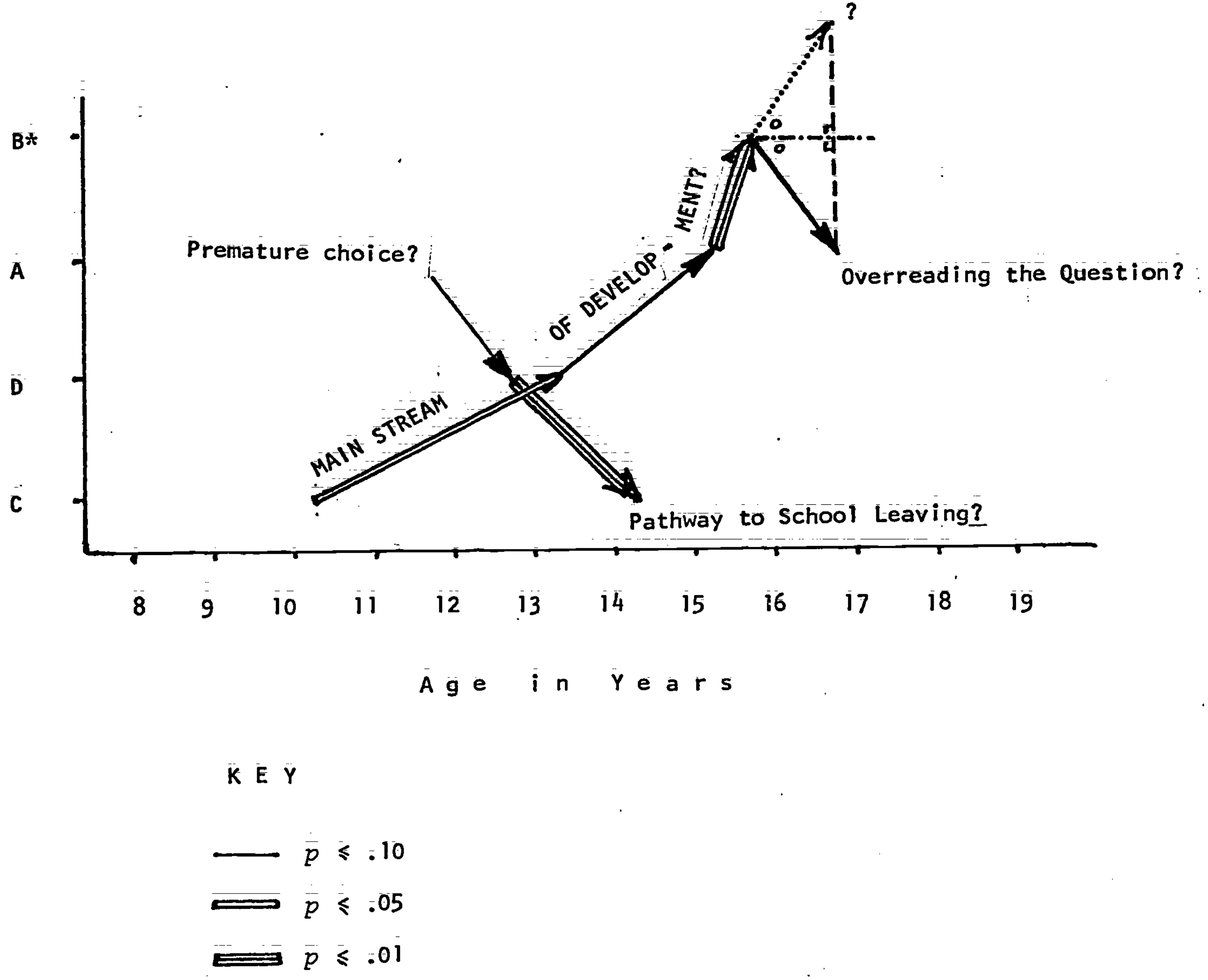
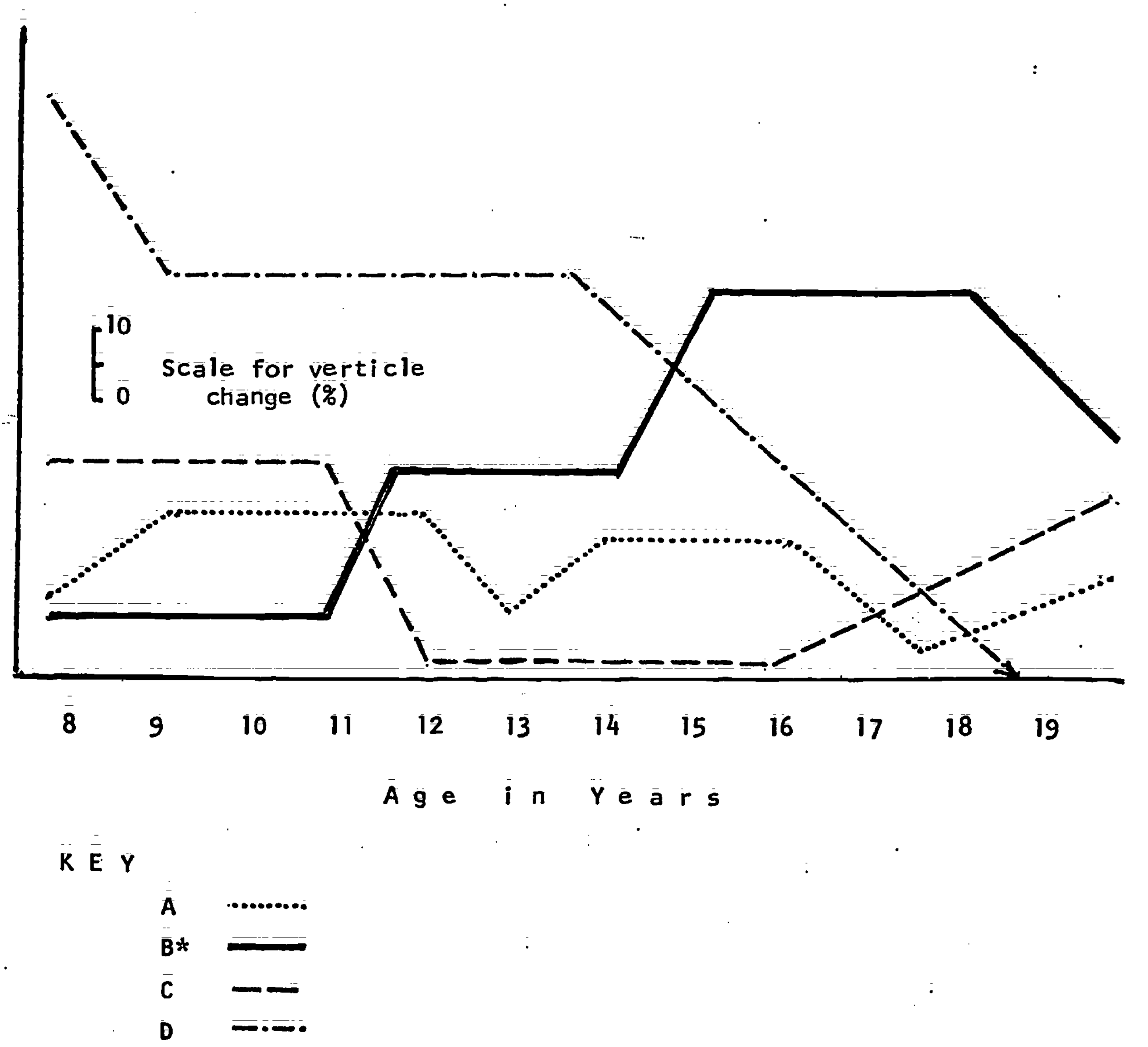


FIGURE 5

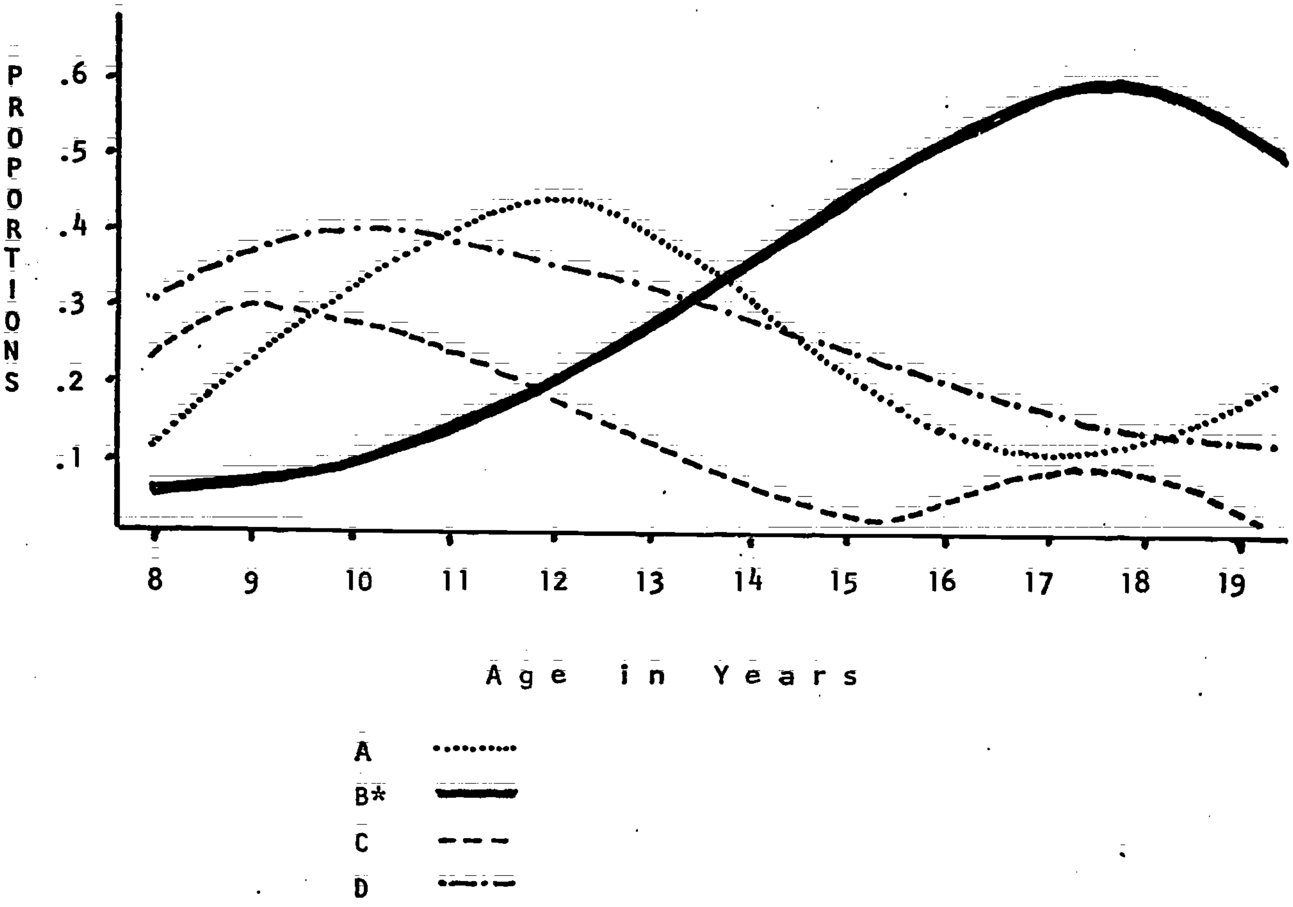
DYNAMICS OF WITHIN - GROUP CHANGES
USING MARGINAL PROPORTIONS



NOTE: This pattern has been simplified from the original vector pathways.

FIGURE 6

DYNAMICS OF BETWEEN - GROUP CHANGES
USING AVERAGE MARGINAL PROPORTIONS



This pattern has been simplified from the original proportion polygon.

FIGURE 7

COMPARISON BETWEEN THE OVERALL CHI SQUARES FOR EACH AGE LEVEL IN ITEM 18 AND THE GROUP SIZE

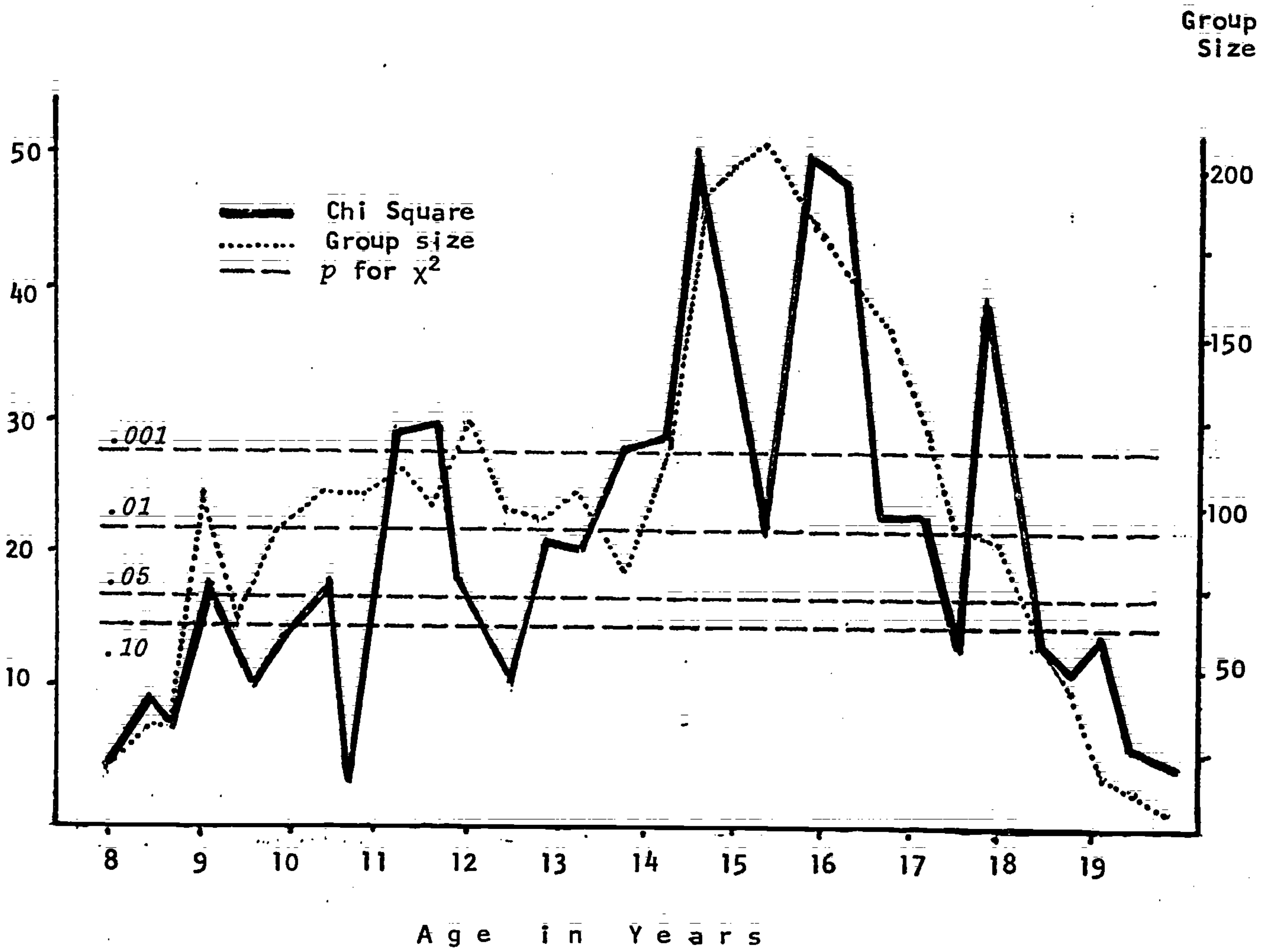
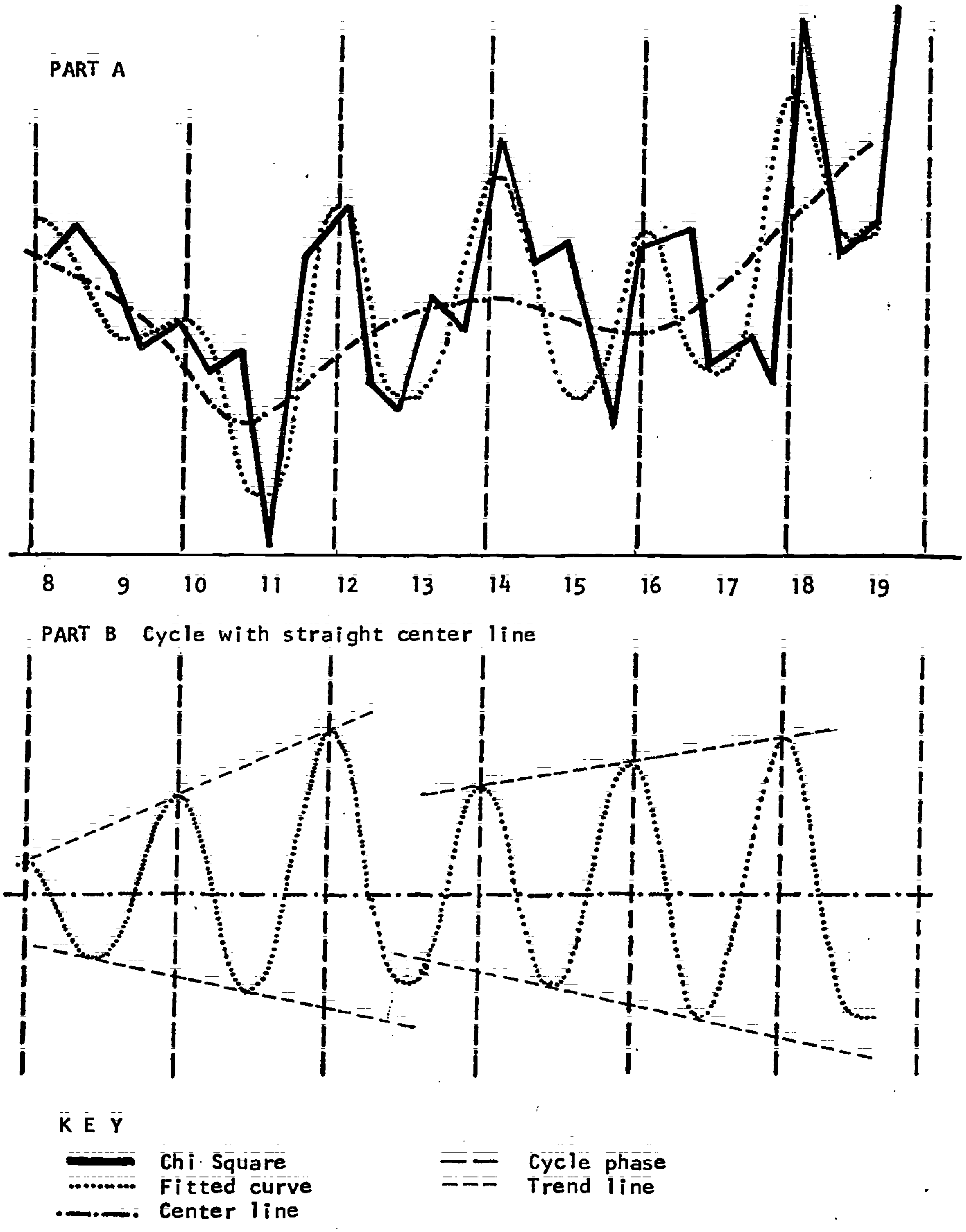


FIGURE 8

EXPLORATION FOR A PATTERN BY
REMOVING THE DOMINANT EFFECT



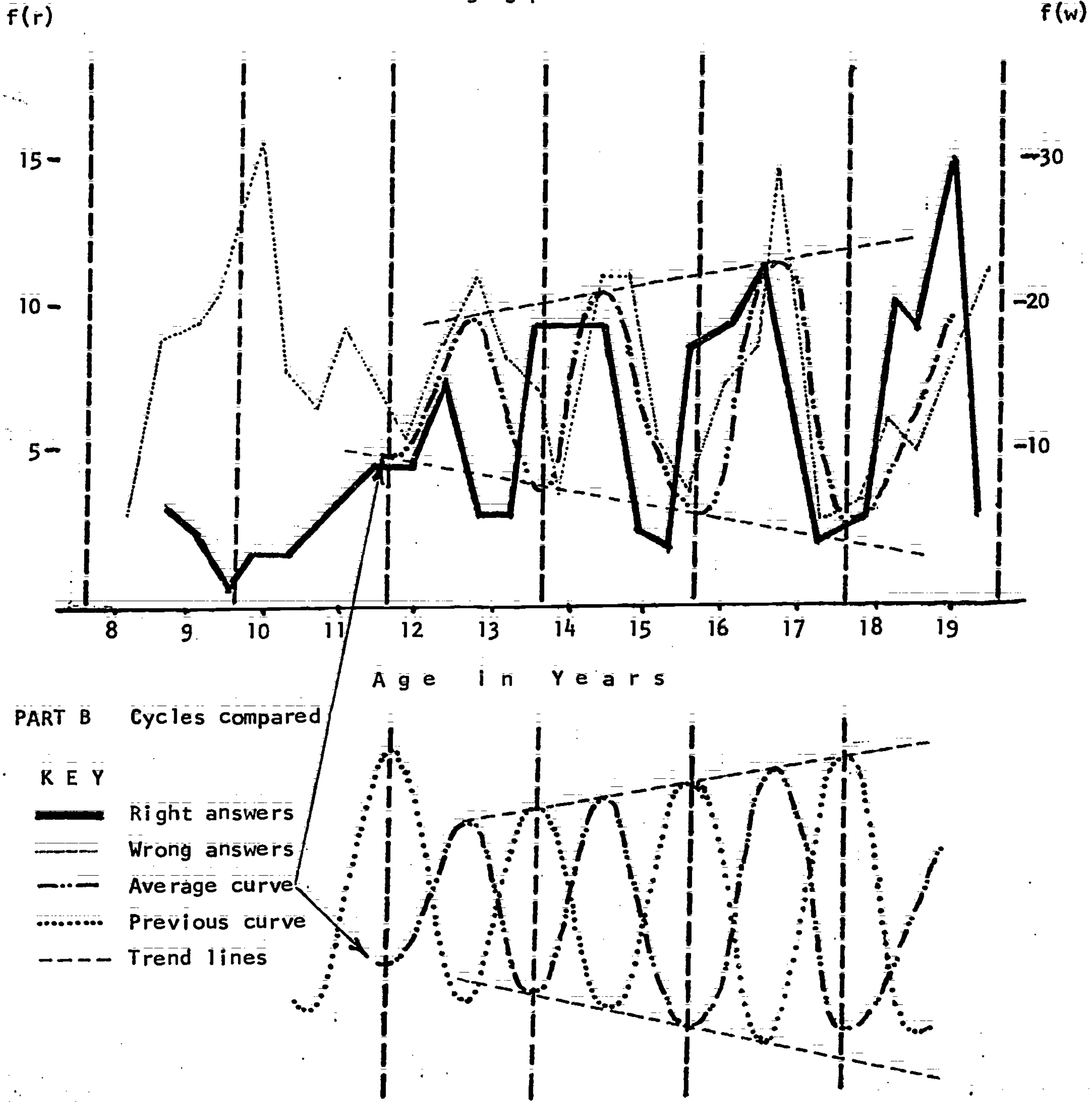
PART A

PART B Cycle with straight center line

FIGURE 9

MATCHING CYCLIC PATTERNS FROM
TWO INDEPENDENT SOURCES

PART A Modes-of-modes from all right and all wrong answers with cyclic phase and averaging pattern



PART B Cycles compared

NOTE: Trend lines taken from previous curve