Three Tasks for Use in Laboratory Small-Group Experiments

by

Thomas L. Conner

December, 1964

Three Tasks for Use in Laboratory Small-Group Experiments

I. Introduction

The design of laboratory experiments in order to construct and test formal models of social processes confronts the investigator with the responsibility of analyzing in more detail than is usual the nature of the activities experimental subjects are to be asked to engage in. This analysis must often include pre-experimental tests to see if the behavior of subjects engeged in these activities in the absence of experimental manipulations can be characterized in a theoretically specified manner. Accordingly, this paper reports on the development and test of three experimental tasks for use in such laboratory experiments. Although they were designed according to a pre-specified set of criteria related to the requirements of a particular series of experiments (Berger & Snell, 1961; Dornbusch, et al., 1962), it was felt that the experimental requirements were general enough that the tasks would be of interest and use to other investigators.

II. Test Criteria

The basic form that each task was required to take was that of a series of discrete decision stages in which at each stage subjects would have to evaluate and choose between two mutually exclusive and exhaustive alternatives. It was necessary that each choice, to the subject, be essentially ambiguous in the sense that standards for choice be ill defined or unaccessible; and that, consistent over individuals and decision stages, there be no bias toward one choice

alternative rather than another. It was also necessary that it be possible to convince subjects that a particular ability be associated with the making of "correct" choices. Finally, the choice at each stage had to be uninfluenced by the choice at any past stage.[1]

The above requirements were common to all three tasks which were developed. However, two of the three had to meet an additional requirement. Previous lack of success in creating tasks which met the requirement of a lack of bias over individuals and decision stages raised the possibility that it might be necessary for the choices to be not only ambiguous but as devoid of any "content" as possible in the sense both of observable principles for choice and of subjective feelings about choices. Hence, two of the tasks were constructed with the aim of meeting this latter requirement as well as those above. The third task remained "contentful" though still ambiguous.

In more precise terms, the above amounts to saying that if in each of a series of binary choices between choice alternatives labeled (arbitrarily) A and B a subject is asked to evaluate and choose between A and B then there exists a number p which is constant over individuals and over decision stages which specifies the probability that any particular subject will choose alternative A. Further, if

$q = 1-p$

m = number of subjects

n = number of decision stages (hereafter called trials)

---

[1] With respect to this last requirement, it is clear that most of the tasks in use in learning experiments would be inappropriate.

x = number of choices of alternative A on any particular trial by all subjects

y = number of choices of alternative A by any particular subject over all trials

then it can be shown that x and y are binomially distributed with respective probability mass functions as follows:

p(x) = probability that on any particular trial exactly x choices will be A

$$= \binom{m}{x} p^x q^{m-x}$$

p(y) = probability that any subject will make exactly y choices which are A

$$= \binom{n}{y} p^y q^{n-y}$$

The test of whether a particular task fulfills the requirements stated above consists in obtaining from the above statements a set of empirical consequences which can be shown to coincide with the data gathered from the actual choices of a group of subjects. If every empirical consequence is in fact verified then the task is without much question assumed to qualify. It would, however, be unusual for there not to be some particular consequence which either failed to be verified or which presented a borderline decision. Such cases if they occur will be examined individually to ascertain whether the failure might be attributed simply to random error or to some particular cause, whether the failure is even important, and whether some technique exists for compensating for the failure.

The actual decision about verification or failure to verify in a particular case is further complicated by the inadequacy of most criteria of evaluation. Statistical tests do exist in most instances although it is dubious in many cases that the existing

tests are really appropriate. This investigator feels that the basic issue to be resolved in each instance is at what point must the investigator resort to a subjective judgment. Some might feel that the choice of significance level is the proper place for such judgments. Others might feel that in some instances the appropriateness of statistical tests is sufficiently doubtful as to necessitate, if possible, judging the results themselves directly. This investigator leans toward the latter opinion and will act accordingly, although for the sake of those who feel differently the results of statistical tests will also be presented. The actual procedure which was followed is outlined below.

To say that no bias exists in favor of one alternative rather than another is to say that p has the a priori value of 0.50. Given the model above, the best estimate of p is the mean proportion of A choices over all subjects, all trials, which should approximate 0.50.

The assumption of independence of trials means that the proportion of A choices, controlling for the immediately previous choice or for any pattern of previous choices is also an estimate of p. The mean proportion of A responses over all subjects and trials considering only the immediately previous response is the only quantity which will be examined here and it will be discussed in terms of an "aggregate transition matrix" in which the rows of the matrix give the proportion of each kind of choice on trial n+1 given that the choice on trial n was of one particular type. Such a matrix is shown below:

choice on trial n+1

		A	B
Choice on trial n	A	p	q
	B	p	q

Independence specifies that the rows of the matrix be identical. A X^2 test for independence in a 4-fold table is supposed to roughly indicate whether they are in fact identical.

The use of the term "aggregate" indicates that the above matrix can be considered to be the end product of the combination of a set of more than one other similar matrixes. One such set specifies a separate one-step matrix for each choice of n (i.e., the choice on trial 2 given the choices on trial 1, the choice on trial 3 given the choices on trial 2, etc.). Another such set specifies a particular matrix for each subject ignoring when the transitions occurred. The examination of the aggregate matrix to test independence assumes that both sets are statistically homogeneous. If either assumption is in doubt, there is a test for independence which does not assume homogeneity (Goodman, 1962). This test involves the computation of X^2 for each matrix in a set and then summing these values over all matrixes in the set. This sum is also distributed as X^2 with degrees of freedom equal to the sum of degrees of freedom of each of the separate X^2 values. However, the examination of the aggregate matrix is the preferable procedure if the assumption of homogeneity is a reasonable one.

The assumption that x and y are binomially distributed will be examined by obtaining from the specified distribution of p(x) and p(y) the distribution of the number of trials on which x choices were A

and the distribution of the number of subjects who chose A y times. The *a priori* distribution for the former is given by np(x) and the *a priori* distribution for the latter is given by np(y). These theoretical distributions can be compared to the observed distributions using a X^2 test.

III. Description of the Tasks

A. The "Meaning Insight" test

In the Meaning Insight test (abbreviated MI in later sections) subjects were asked to choose which of two primitive, non-English words (actually artificial words) had the same meaning as an English word which was presented. They were instructed to make this choice on the basis of the *sound* of the primitive words and the *meaning* of the English word. For example, in the set of words shown below each subject would be asked to choose which word, A (KUL) or B (TUM), has the same meaning as the English word "bear".

BEAR

A B

KUL TUM

Subjects were led to believe that the artificial words were phonetically spelled words from an actual language or languages. They were also led to believe that individuals differed in their ability (i.e., Meaning Insight Ability) to intuitively choose the proper word. The choice of such an unusual kind of task was motivated in part by the requirement that subjects should have no pre-experimental estimate of their own ability based on previous performance in other activities.

It was intended that subjects who took the test would have feelings of association between the English word and the primitive words

and it was hoped that these feelings would be balanced between the items in each primitive word pair. This latter requirement is an especially difficult one to meet as there do not now exist any proven or objective criteria for balancing the words. Hence, it was decided to construct a large number of the word sets by use of some informal phonemic-semantic criteria to be described below and pretest them with the intention of selecting those sets which empirically seemed to display a lack of bias. This select group of word sets was then pretested again to see if they met the criteria explicated previously in section II. The second pretest is the one reported on here.

The informal phonemic-semantic criteria for word set construction were based loosely on the work of Charles E. Osgood, James J. Jenkins, Wallace A. Russell, and George J. Suci. These were utilized solely in the hope of perhaps increasing the yield of acceptable word sets. There was no serious theoretical interest in the criteria themselves.

With respect to the English words, the meaning of each word was rated as positive or negative on each of a set of five dichotomous dimensions listed below with the arbitrarily selected positive aspects listed first.

strong - weak

large - small

heavy - light

hard - soft

fast - slow

In the example given, "bear" would be rated as strong, large, heavy, hard, and fast. By contrast, the word "feather" would be rated as

weak, small, light, soft, and slow.

The attempt was then made to characterize the phonemic aspects of the vowels and consonants of the artificial words using the same dichotomous dimensions. Artificial words were then matched and selected to associate both positively and negatively with the English word. In the example given, m and l in the final position in a word were both rated (compared to other consonants) as strong, small, soft, slow, and neither heavy nor light. The consonants t and k were rated as weak, small, light, hard, and fast. The vowel u appears in both words and hence its rating is irrelevant. The details of the ratings are unimportant and are discussed only to generally indicate how the word sets were constructed. In the final analysis the empirical results of the pretest alone were what would be used to decide the usefulness of the task.

Two further examples of word sets from the test are given below.

SUDDEN		SHADY	
A	B	A	B
TO-KA	PA-TO	ZEM-PA	BEM-SA

Each such word set was photographed and a 35 mm. slide made. These slides were projected in timed sequence for the pretest allowing 10 seconds for examination of the words. At the end of each 10 seconds subjects were instructed to mark answer sheets and the slide was removed from the screen. Responses from the answer sheets were later punched onto IBM cards for analysis.

B. The "Relational Insight" tests

Two other tests similar to the Meaning Insight test were devised. As noted previously these tests were to be devoid of "content", which

meant that subjects were to have minimal feelings of association between the choice alternatives A and B and the single comparison stimulus. The Relational Insight tests utilized ideographic characters from ancient Japanese and phonetic spellings of ancient Japanese words. Subjects were instructed to match the sound or sounds of the Japanese word or words with the form of the ideographic character or characters. (The actual instructions to subjects were phrased as a matching of "sounds and symbols"). The real meanings of the words and characters were declared to be irrelevant to correct matching and subjects were instructed not to attempt to deduce meanings. An example of the first test using two ideographs and one word (labelled RI2) is shown below.

wazuka

(A) (B)

Subjects were to pick an ideograph, A or B which seemed to be "correctly" matched with the sound of the Japanese word.

An example of the second test using two words and one ideograph (labelled RI1) is also shown below.

azakeru asamuku

(A) (B)

In this test subjects were to choose one of the words, A or B, which matched the ideograph. Notice that in both cases the ideograph pairs or the word pairs are matched in some respects and unmatched in others. This matching was done according to purely intuitive criteria.

As in the previous test subjects were instructed that some individuals were better able to relate (because of "Relational Insight" ability) the forms of the ideographs with the sounds of the words. Both tests, as before, were photographed and pretested twice in the same manner as the Meaning Insight test. The second of the pretests is reported on here.

IV. Results of the Pretests

As indicated previously, a select subsample of slides from each of the three tasks was chosen on the basis of an initial pretest and then shown to groups of subjects in a second pretest to ascertain whether the tasks met all of the criteria specified in section II. The number of slides tested for each task is given below.

Relational Insight, two ideographs (RI2) : 61

Relational Insight, one ideograph (RI1) : 39

Meaning Insight (MI) : 37

The first data to be presented concerns the estimate of p for the three tasks. In the table below the mean proportion of A choices over all subjects, all trials is given for each task. For comparison purposes, the same estimate from the initial pretest is given, both for the entire set and the select subset.

	entire set first pretest	select subset first pretest	select subset second pretest
RI2	.49 (N=15,790)	.49 (N=10,492)	.49 (N=5760)
RI1	.49 (N=17,897)	.49 (N=7449)	.49 (N=3306)
MI	.51 (N=21,439)	.47 (N=8251)	.52 (N=3600)

TABLE I

Estimates of p from mean proportion of A responses over all subjects, all trials.

There seem to be no large departures from previous results or from the *a priori* figure of .50 in the above table. Hence, in this case the binomial model is taken to be confirmed.

The question of the independence of trials, as previously discussed, can be dealt with in two ways. The first is the examination of the aggregate transition matrix for each task. These are given below. Beside each matrix is the value of X^2 for the usual test for independence in a 4-fold table.

RI2

	A	B	
A	.46 (1298)	.54 (1503)	(2801)
B	.51 (1512)	.49 (1447)	(2959)
	(2810)	(2950)	(5760)

$X^2 = 13.05$

$p < .001$

RI1

	A	B	
A	.47 (774)	.53 (871)	(1645)
B	.52 (872)	.48 (789)	(1661)
	(1646)	(1660)	(3306)

$X^2 = 9.80$

$.001 < p < .01$

MI

	A	B	
A	.50 (946)	.50 (931)	(1877)
B	.54 (926)	.46 (797)	(1723)
	(1872)	(1728)	(3600)

$X^2 = 4.025$

$.02 < p < .05$

In each case the X^2 values are such as to require a rejection of the binomial model. This conclusion must be tempered, however, by the fact that in dealing with sample sizes of this size it is very easy

to obtain significant X^2 values. Should it be the case that the model fits the data in all other respects, whatever departures from independence these values indicate could easily be ignored.

The second way of dealing with the question of independence c examination of the matrixes which were combined to produce the aggregate matrix. Recall that there are two different sets of such matrixes. The first mentioned was the set which includes a matrix for each particular one-step transition. The sum of X^2 values for these matrixes together with the degrees of freedom and estimated significance level are given below for each task.

	$\sum X^2$	d.f.	sig.level
RI 2	95.93	60	.002
RI 1	40.45	38	.37
MI	51.60	36	.04

TABLE 2

Sum of X^2 values of one-step transition matrixes for each particular transition.

The second set of matrixes included a matrix for each particular subject ignoring when the transitions occurred. The sum of X^2 values for these matrixes together with the degrees of freedom and estimated significance level are given below for each task.

	$\sum x^2$	d.f.	sig.level
RI2	119.32	96	.05
RI1	85.83	87	.52
MI	105.45	100	.34

TABLE 3
Sum of X^2 values of one-step transition matrixes for each particular subject.

There would seem to be no clear conclusion one could draw from these tests as they give inconsistent results. The aggregate matrixes indicate one-step dependence for all three tasks. The sum of X^2 values for the separate one-step matrixes indicates that only RI2 and MI display one step dependence. The sum of X^2 values for separate subject matrixes indicates that only RI2 displays one-step dependence. These tests, of course, are not independent tests and this further complicates the case. Perhaps the safest way to approach the problem is to ask the following questions: if it were the case that a one-step dependence exists, what kind would it be and how might it be explained? With respect to the former question, the aggregate transition matrixes in each case indicate that the direction of dependence, if it exists, is toward too much alternation. With respect to the latter question, it must be recalled that the labels A and B are arbitrarily assigned. Hence, either there is something intrinsic in each task which produces the dependence and the assignment of labels happens to coincide; or, more probably, there is some small intrinsic bias in binary choice

sequences in general toward too much alternation. This latter explanation is the one which the investigator prefers, but, of course, any position at this point is speculation. In any event the evidence for one-step dependence is not overwhelming and for the present is not regarded as a serious problem.

The final set of data to be discussed is probably the most important although so far little has been said about it. The tasks really stand or fall on the distributions of np(x) and mp(y). Graphs 1-3 show the distribution of x vs.np(x) for each task. The dotted lines indicate a priori binomial distributions and solid lines indicate observed distributions. Here clearly there is a failure of agreement with the predictions of the binomial model, although RI2 certainly comes the closest to fitting. The X^2 goodness of fit test in each case gives a significance level beyond .001. Casual inspection of the graphs is also quite convincing in support of a lack of fit.

A reasonable explanation for this failure would be a lack of homogeneity and/or consistency from slide to slide. That is, p may not be constant over decision stages. If this were the case then it would have serious consequences for the interpretation of the results of any experiment. It is also possible that this lack of homogeneity could partially account for whatever indication of one-step dependence there was in the data examined previously.

The question which now arises is, given the above, what can be done, in terms of the criteria specified in section II, to salvage the tasks. The one solution which seems feasible is to randomize the order of presentation of the slides for each experiment

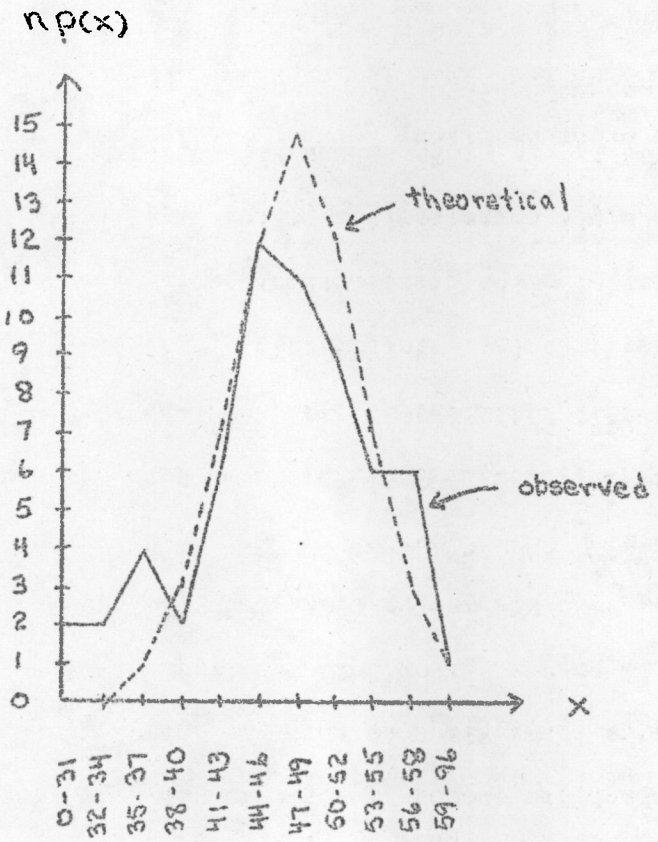
which is carried out. This increases within trial variability and would have to be compensated for by increasing the size of the sample of subjects who participate. No other solution, however, seems immediately obvious.

The final data are given in Graphs 4-6 which show the distribution of y vs. mp(y) for each task. In each case the a priori distribution (dotted line) is closely matched by the observed distribution (solid line). None of the X^2 values is large enough to indicate a lack of fit. Hence in this case the prediction of binomial model is taken as confirmed.

V. Conclusions

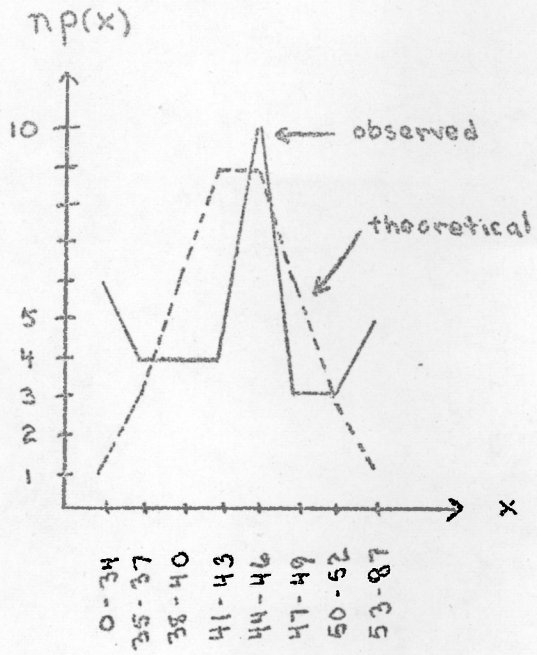
The pretest of the three tasks cannot be said to have been completely successful. All three failed the same important test, the distribution of x vs. np(x). Fortunately, however, a method exists (randomizing order of slide presentation) for compensating for this failure, and hence all three would seem to be at least usable. The question of which task is preferable would not seem to be decideable on the basis of the data given, although Meaning Insight seemed to come out worse than at least one of the other tasks on most tests. Other criteria will more than likely have to be invoked in the final analysis.

The fact that the pretest in some respects is inconclusive stems from two main facts. The first is that the job of constructing tasks which meet a rigid set of criteria is far from easy. The second, encountered throughout the analysis, is that the investigator is often faced with problems of evaluation for which there are no clear criteria or ready solutions. Only continued work and thought can reduce this margin of inconclusiveness.



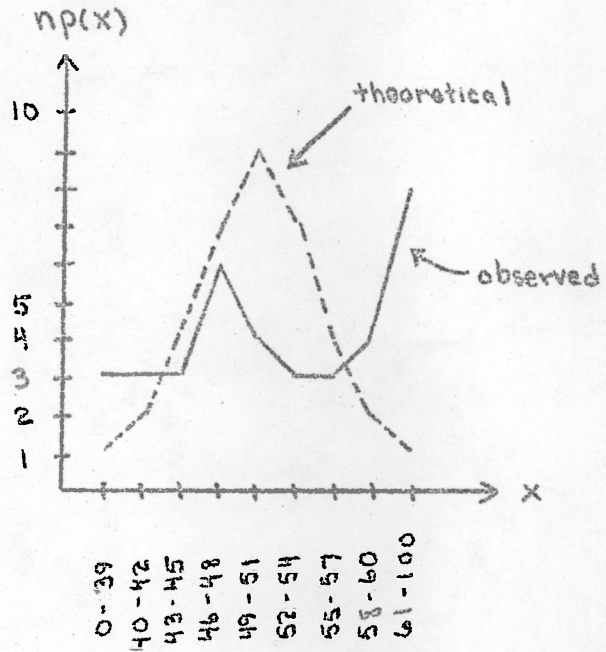
GRAPH 1

x vs. np(x) for RI2; $X^2 = 28.43$, $p < .001$



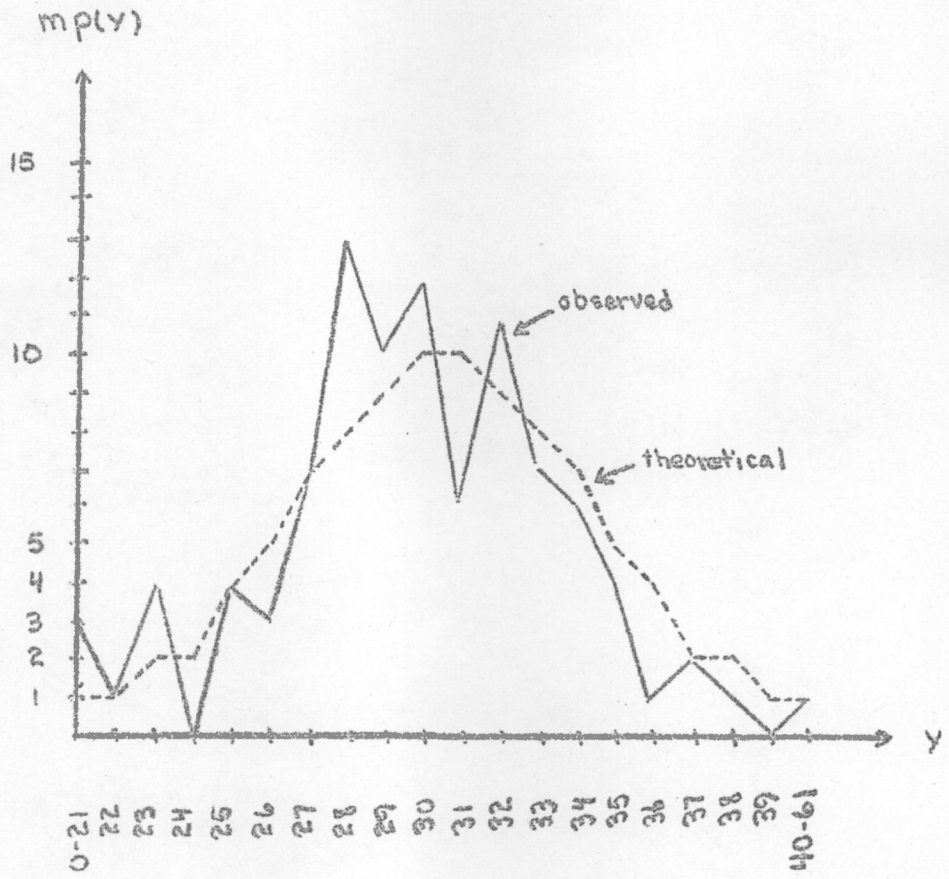
GRAPH 2

x vs. np(x) for RI1; $X^2 = 46.39$, $p < .001$



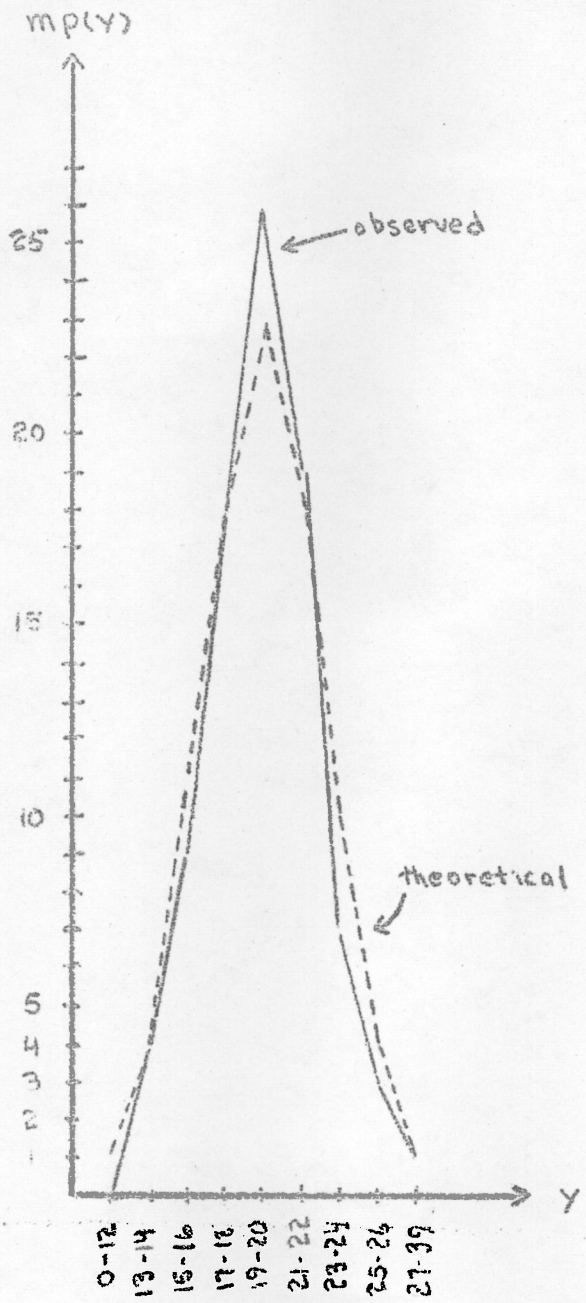
GRAPH 3

x vs. np(x) for MI; $X^2 = 61.21$, $p < .001$



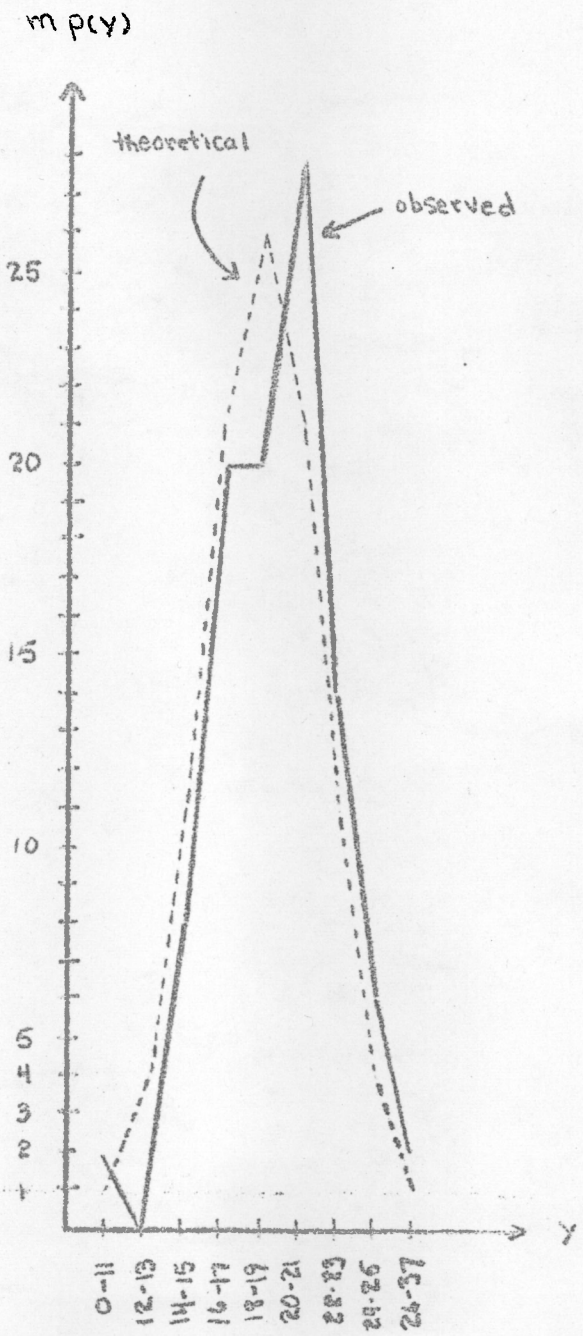
GRAPH 4

y vs. mp(y) for RI2; $X^2 = 18.20$, $.50 < p < .70$



GRAPH 5

y vs. mp(y) for RI1; $X^2 = 2.70$, $.95 < p < .98$



GRAPH 6

y vs. mp(y) for MI: $X^2 = 12.40$, $.10 < p < .20$