DOCUMENT RESUME

ED 027 148 RE 001 341

By-Wark, David M.

Designs for Measurement of Status.

Pub Date 25 Apr 68

Note-16p.; Paper presented at International Reading Association conference, Boston, Mass., April 24-27, 1968.

EDRS Price MF-$0.25 HC-$0.90

Descriptors-Educational Research, Instrumentation, *Measurement Instruments, Research Design, Research Methodology, *Research Problems, *Sampling, *Test Reliability, *Test Validity

An examination of typical situations concerning the status of a variable is followed by a discussion of the measurement and empirical problems involved in measuring current status. Arguments for more consideration of variability, particularly individual variation, are presented. Among the problems discussed are (1) describing large samples through randomized and controlled sampling, (2) screening subjects into special categories for some special programs, (3) diagnosing an individual's abilities in different skills, (4) establishing a basis for measuring and describing behavioral change, (5) determining the validity and reliability of instruments used and of individuals tested, and (6) solving empirical problems of motivation, practice effect of multiple testing, disruptive effects on test results, and of the time involved. References are given. (NS)

DESIGNS FOR MEASUREMENT OF STATUS

David M. Wark

University of Minnesota

INTRODUCTION

This paper represents half of a two-part symposium on research design. In this paper, the emphasis will be on the measurement of current status. That is, we will be concerned with designs to take a cross sectional sample from the longitudinal flux of behavior. These cross sectional samples may be of various lengths. All that is required is that the period be treated logically as a unit for the purpose of study. Such samples are, of course, conceptually prior to any designs for measurement of change, the topic of the other half. We shall examine here typical situations in which the focus of a problem is

---

A paper presented at the International Reading Association Meeting, April 25, 1968, Boston, Mass. To be published in the IRA Proceedings. This paper was made possible by a grant in aid from my wife.

the status of a variable. Next, we shall turn to the empirical and logical problems involved in such measurement. Finally, we turn our attention from concern for status to the complementary but often ignored problem of variability as a worthy topic of study.

TYPICAL PROBLEMS IN THE MEASUREMENT OF STATUS

1. Descriptive Summaries of a Large Group

It is often desirable to provide a statistical picture of large groups of students. For example, one may wish to describe the skills of an entering freshmen class, or of the total eighth grade or of all the students in an Honors Program. These studies would be useful for planning remedial programs, establishing priorities for new purchases, and noting trends in ongoing programs.

If the numbers of students to be summarized is large, it is often desirable to sample, rather than to test every student in the potential population. In such a situation there are two possible approaches that could be used, each with different design problems. One approach is to take a random sample. In this design students are selected in such a way that every individual has an equal and independent opportunity of being included in the sample (1). These characteristics are vital if the sample is to be representative of the total group. The classic technique for this procedure is to assign a different number to each student and then use a random number table to draw numbers and therefore, students to be tested. As a common alternative, student cards can be filed in alphabetical or numerical sequence. Randomness is not violated if every Nth student from the list is selected and tested (2).

However, it may happen that some students have no chance of appearing in the sample. Such a constraint violates the requirement of equal opportunity for

sampling. For example, if it were desirable to test an incoming freshmen class it might be convenient to sample small orientation groups as they appear on campus. However, if for administrative reasons only the first or last groups to appear were sampled, there would be a distinct biasing factor. Many students would not have a chance to be tested. The students who for one reason or another have early or late orientation dates would make up the sample. They probably would not be representative of all freshmen. We might suspect that those coming to campus early might be from upper socio-economic levels and would not need summer employment or from lower levels and could not find jobs. Either situation would give a biased picture of the freshmen class. That is, some students contribute more to the averages than others. It would be more desirable to randomly sample groups all through the summer. Thus, if students do not all have an equal chance of appearing in the sampling, the picture of status may be inaccurate.

Even though all the students have an equal chance of being selected, some may not have an independent chance, and thus invalidate a sample. (1) That is, the selection of one person may autom atically raise the selection probability of another who is associated with him. Let us assume a school with 1,200 enrollment. We desire a 10% random sample of students to take an attitude survey. But having 120 individual students to come in for testing is too difficult. Will the design problems be solved by selecting six classes of 20 students, if the classes are chosen by random number table?

Yes and no. The requirement for equal opportunity of selection will be met. Thus, we would expect the sample to be unbiased. But what is being sampled, students or classrooms? Clearly, the latter. Because of the violation

of independence, the sample unit is groups in rooms, not individuals (6).

There is another problem. Some investigators would assume they were sampling 120 students from a population of 1,200. They might not realize that they were actually drawing six classrooms from a population of approximately 60. They would compute the standard error of their statistics using the inappropriate N of 120 students, instead of working with a replicated N of six classes. Since the standard error of a mean is reduced by using an increased sample, the incorrect computation has two effects: First, the smaller error term allows a fallaciously narrow confidence interval. The results would be an apparent precision of measurement but totally unjustified (2). Secondly, because the errors of measurement are so small, the differences between sample statistics may not exceed the conventional levels. With the correct N's and the correct standard error, the differences that are there may show up. Violations of the independence assumption lead to conservative inferences (9).

The second approach to the sample-for-summary problem is to draw a controlled, non-random smmple. In that situation, one would specify and select intentionally samples which were known to be representative of the population in some way. If a population contains a known percentage of male, freshmen, arts-college students living off campus, the sample would be controlled to contain a similar percent of such individuals. It is quite important that the actual students making up the sample be randomly selected.

The controlled sample is more efficient than a purely random procedure if the controlled sub-groups are known to have less variability than the population from which it is drawn (9). In the case of our hypothetical freshmen, the controlled sample is preferable if the male freshmen, etc., have more homogeneous

scores than the total class. If not, the controlled procedure lacks efficiency and may be biased. In general, the more precisely the sample can be controlled, the greater the efficiency but the greater the cost per respondent. The extra cost is the frequently necessary "call backs" to test students who have been identified by random procedures. Once a student has been so identified, he must be tested or the results will be biased.

2. Screening

Another reason to carry out status studies is to screen students into separate or special classes. A common example here is to screen college freshmen and take the bottom 10% for some special program such as reading or study skills. Unfortunately, this may be a very inappropriate procedure if the wrong type of test is used. When screening to cut-off below a certain point, such as the bottom 10%, it is best to use a test that is known to have a 50% difficulty index at that cut-off point. Tests are most discriminating at the point where half the students fail. In cutting off the bottom 10% of a college freshmen group it would be inapprppriate to use a test normed on college freshmen. The test would lack sufficient differentiating power at the bottom end of the distribution. In this case, it would be better to use a test normed on high school seniors or even high school juniors and to take everyone below the mid-point for the high school population. As a matter of fact, in terms of the Triggs Diagnostic Reading Test national norms in order to cut off the bottom 10% of a college freshman sample, one should use the tenth grade median.

It is well to remember that in any kind of special screening studies, students selected as atypical, either high or low, will, when tested a second

time, regress toward the mean of the first test (1). Thus, students selected for special treatment because they are low would normally show much improvement on a post-test even if the treatment were totally irrelevant.

3. Diagnosis

A third situation which can properly be considered a study of status is the diagnosis of individual skills. In this case, one tests to find comparative scores for different abilities: rate, retention, vocabulary, etc., rather than for a total overall score on reading. In fact, in a diagnostic situation the total test score is relatively unimportant (5). This means that in setting up a test battery it is important to be quite sure that the sub-scores used in the diagnostic categories are uncorrelated. If there is a high intercorrelation between sub-scores there is no separation and very little diagnostic validity to the test.

4. Establishing Base Rates

A fourth reason for carrying out status studies is to establish a solid base for experimental manipulation in the future. This blends into the question of designs for the measurement of change. Before assessing this change it is necessary to have a solid base of measurement. If treatment designed to increase a variable such as rate is applied when the measure is on an upswing anyway, it is possible to get an artificially high score. In such a case, the researcher has a highly increased probability of rejecting the hypothesis of no difference when, in fact, he should not do so. For example, giving a lecture on use of the library and finding that students increase book checkouts may not allow one to conclude that the lecture was effective. We would have to show that students

were not increasing their use of the library perhaps because of an eminent term paper. We need to know what the status of the behavior of interest was over a long period of time. David Yarington has done probably the most thorough study of college student reading behavior by examining reading rates and amounts week by week through an entire academic semester. He reports upswings in amount of time on reading prior to examinations but significant variations from week to week (12). Thus, it would be necessary to gather base rate information prior to any experimental studies designed to manipulate the amount of reading students do.

In such a situation of base rate gathering an experimentor might be inclined to briefly measure the behavior of interest and then block students into groups with similar scores. He might use an analysis of variance or even co-variance technique to equate for the measured pre-experimental behavior (9). But note that sophisticated analysis would be of little help in this case. The design question is one of long-term base rate. The period of observation must be long enough for the behavior of interest to stabilize. The question of an adequate criterion for stability is almost untouched in the field of reading. What is "stable comprehension?"

Murrey Sidman in his book "The Tactics of Scientific Research" states "The descriptive investigation of steady-state behavior must precede any manipulative study. Manipulation of new variables will often produce behavioral changes, but in order to describe the changes, we must be able to specify the baseline from which they occurred; otherwise we face insoluable problems of control, measurement and generality." (1,pp238)

PROBLEMS WITH THE NOTIONS OF STATUS

1. Measurement Problems

There are certain measurement problems involved in the notion of status. First of all, there are the universal concerns which apply to any testing situation, whether it be status or change. There is the problem of validity. In some cases this is no problem. We can develop a test of dart-throwing and measure directly students' effectiveness. But in the field of the readings such is hardly the case. The validity of a test of comprehension, for example, depends on a whole series of assumptions that we must make about the nature of comprehension. There is no obvious overt criterion. The use of such quasi poetic, non-behavioral objectives as "grasp the main idea" does not in any way solve the problem. Comprehension is internalized behavior, not open to the scrutiny of the researcher. We must infer the behavior concerned. The ultimate criteria in reading, the objective behavior, is not examinable with current technology. If it is ever to be brought out for scrutiny, it will be through the efforts of other behavioral and physical scientists working jointly with reading specialists. And until that cross collaboration begins, we are doomed to reading tests of questionable validity.

Another universal problem is reliability. Any measurement, physical, biological, social or educational will have some error built into it. Thus, we can expect slight differences between scores for the same skill measured on two different occasions. Each test has a reliability quotient which gives us some notion of how discrepant two measurements will be. These reliability quotients may be based on the various types of correlations-part scores, test re-test, parallel forms, etc. (5). Each has certain advantages. Hopefully,

the test manual accompanying a particular test will describe the type and extent of reliability, and give the logic for its selection.

In the measurement of status we must be concerned with more than just the reliability of the test. We are in a sense asking "what is the reliabiiity of the student." He will change from test situation to test situation yet we are purporting to measure something stable about him. We are able to talk about the reliability of a test because that is what is printed in the manual. But we should not forget that we are also concerned about the reliability of the human who takes the test. Think how we could increase the value of our research and service if each student had his own reliability quotient, perhaps stamped on his forehead.

There are several non-universal concerns for measurement of status. These would be specific to a particular situation. We might, for example, ask in a particular study what is the student's motivation for taking the test? Another way of saying that is what is the student's reward for taking the test. Couching that question in the language of learning theory (reward) immediately raises some interesting research questions. In what way does manipulating the "reward" affect the test behavior? Another particular question is, what is the effect of multiple testing in a battery? What is the pile-up effect, in other words, of taking several tests. Wark and Kolb (10) found that instead of test sensitivity and test experience, repeated testing can produce test fatigue and boredom over a short period of time. These provincial, unique questions can raise clear havoc in a stable baseline design. And of course, there is always the question of unplanned disruptive affects. Noise outside a testing room can affect testing. We might call this the brass band effect.

2. Empirical Problems

Quite aside from the measurement questions, there are certain empirical problems with the notion of status which must be faced. In fact, they are critical for research of any sort. Testing for status of any skill takes time. Time to test, to score, to interpret, to act upon those interpretations. Yet the skill measured is assumed to be stable throughout the period. What a student was during the time he took the test, he is presumed to be now when he is admitted to a reading program or exempted from an English course. The assumption that student skills are stable through time demands that at least one of the two following situations hold.

1) Strong assumption. The output measure will be constant throughout the time period that the student is being tested. This is not a safe assumption.

Figure 1 Here

Figure I shows the fluctuation in reading rate during five minutes of measurement using Form A of the DRT, upper level survey. Students were told to mark the line being read at the end of the two through fifth minute. We note that some students (S3, S4, S5) tend to show a gradual increase through the period but that the more rapid readers (S1, S2) tend to fluctuate rather independently and quite widely. We can't assume that even such a thing as reading rate, easily measured though it is, is stable through the testing period.

2) Weak Assumption. The output will show some variability but the

amount will be random. Thus, variability will balance or cancel out over time. Here again there is some doubt that the assumption is always true. Humphrey (3) reports that variability in rate is not random. It may interact with level of rate in a very complex way. Fast readers were most variable and mid-range readers were least variable. Slow readers were intermediate in variability. The results are based on three to seven samples of four minutes. Unfortunately, the published data report only group means. There is no indication whether or not individual students had achieved stable rates at some time during testing.

VARIABILITY AS DATA

In addition to the empirical questions, there are some conceptual problems with the notion of status. The very word "status" suggests a limited, if not fixed value. It suggests well defined quantities in the data. It biases us--this word--from thinking about variability as a datum in and of itself. When one considers status one treats variability as a nuisance factor. Yet there are many questions under the rubric of status problems that would justify a closer look at the extent and determinents of variance. These include but are not limited to:

1) Flexibility. Typically this is considered to be a function either of the purpose or the personality of the reader (8). It is indeed possible to affect variability in rate by appropriate instructions,(11) or what we could call purpose. What then, are the determinents of flexibility conceived as a variability in rate? How does one increase, decrease and shape it?

2) Readability. We can think of readability as a problem of variance. Again, we tend to think of the readability as a fixed value measured once and determined solidly (4), yet within material of specified readability there may be fluctuations that would be worth examining.

3) Study-Type Reading. We might consider the approach to study-type reading as a problem in variability. Students should skim some material and slow down and read intensively other material within the same text. Thus, study-type reading is an area where variability and its determinors should be of basic interest.

4) Typographic Clues. In what ways do graphic and typographic gimmicks affect rate, and what are the limits of such an effect?

5) Concentration. There is the question of concentration conceived of as variability rather than a clinical or internal problem. What determines the rate at which someone scans a text and then "daydreams?" Do students slowly peeter out in rate, or do they read at a stable speed and then suddenly drop off? Informal study at the University of Minnesota would suggest that it is extremely difficult to get data on this topic. When one asks students to examine their daydreaming, the variable vanishes. Apparently, when one concentrates on it, it is gone. This makes for much difficulty in analysis.

Another troublesome problem with variability is that when one is oriented towards study state or status studies one tends to design one's investigation in order to rule out variability. But there is no design that will make variability go away. It is possible to use large samples and thereby decrease variance, since variance is a function of the number of cases in the study. Or it is possible to balance out differences by

counterbalancing the sequence of tests. Thus, in a test of audio-visual methods of instruction in which one needs to get both audio and visual pre-tests it would be possible to give half the students audio first and half the visual first. Does this design solve the problem of sequencing and variability? It does, if one is willing to settle for the whole group as an experimental unit. But if you are really interested in individual people and their responses, it doesn't. Counterbalancing as a way to remove fatigue, variability or any nuisance variables in a status study is a little bit like a magician making cards disappear into a black box. Everybody in the audience knows that the cards are still around some place, but no one is quite sure where they are hidden. The analogy holds for counterbalancing as a way to remove variance in a status study. Anybody who thinks about it knows that the variability still exists, but it is hidden, screened by a lot of statistical legerdemain. When one is interested in the status of individual behavior one should look at that behavior and recognize questions of status and variability must be handled together.

This is perhaps a strange note upon which to end a paper on status. Yet, as I have tried to demonstrate, status and variability are like two sides of a coin. I feel that too much consideration has been given to the side of designs for group stability. We can profitably turn to the questions of individual variation. If we use our heads, we may have some interesting new tales to tell.

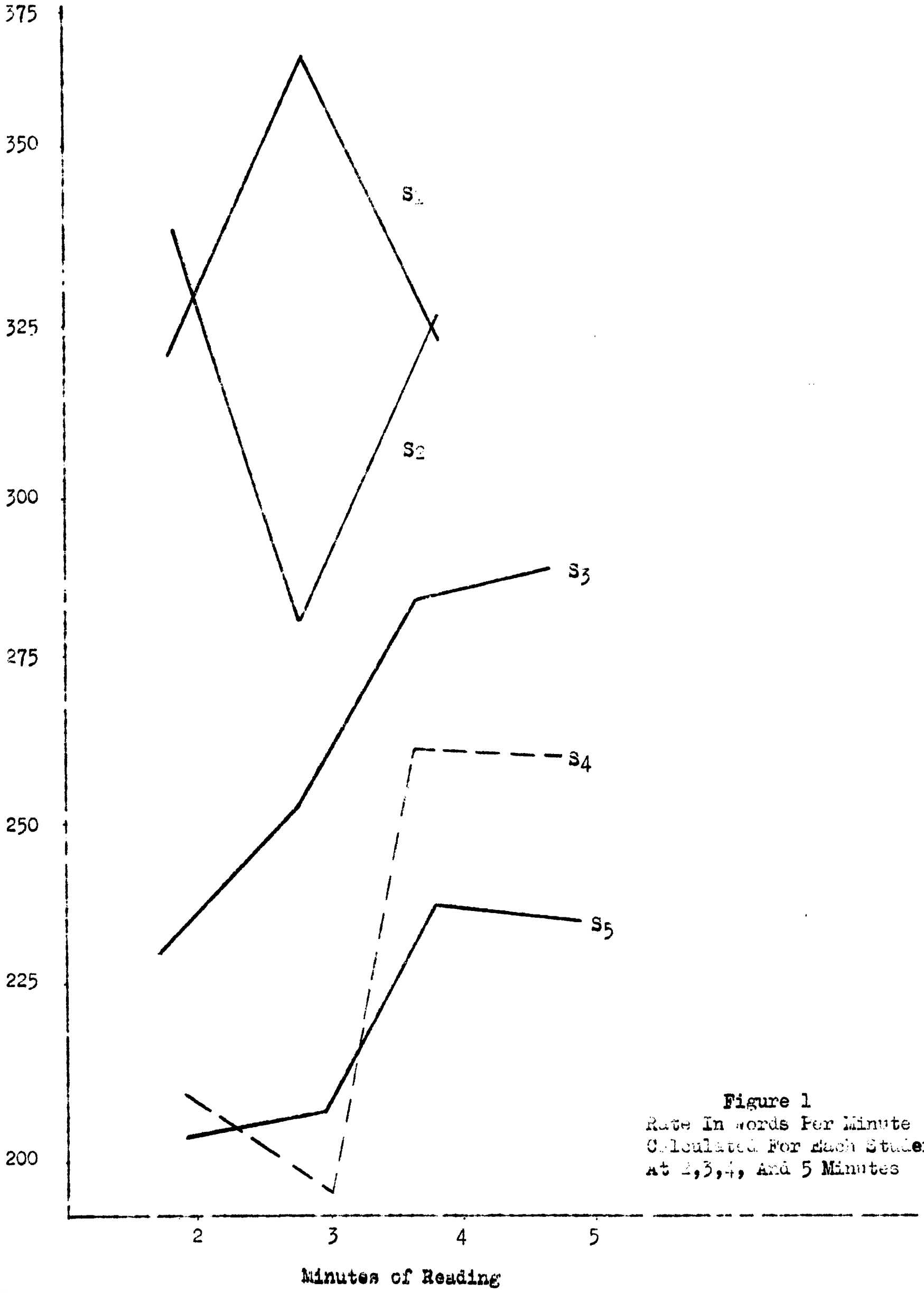

Figure 1
Rate In Words Per Minute Calculated For Each Student at 2,3,4, And 5 Minutes

BIBLIOGRAPHY

1. Campbell, D.T., and Stanley, J.

   Experimental and Quasi-Experimental Designs for Research on Teaching. In Gage, (Editor) Handbook on Research in Education New York: Rank McNally, 1960.

2. Edwards, A.

   Experimental Design in Psychological Research.

   New York: Rinehart and Company, Incorporated, 1956.

3. Humphrey, K. H.

   An Investigation of Amount-Limit and Time-Limit Methods of Measuring Rate of Reading. Journal of Developmental Reading. 1 (1957) 41-54.

4. Klare, G.

   The Measurement of Readability. Ames, Iowa: Iowa State University Press, 1963.

5. Lindquist, E. F. (Editor)

   Educational Measurement. Washington, D.C.: American Council on Education, 1950.

6. McCall, William A.

   How to Experiment in Education. New York: Macmillan Company, 1923.

BIBLIOGRAPHY
(cont'd)

7. Sidman, M.

Tactics of Scientific Research. New York: Basic Books, 1960

8. Taylor, S.E.

Eye Movements in Reading: "Facts and Fallacies". Reading Newsletter #30 F. P. L., Incorporated, November, 1963.

9. Walker, H. and Lev, J.

Statistical Inference. New York: Henry Holt and Company, 1953.

10. Wark, D. and Kolb, M.

An Experiment in High-Pressure Reading Instruction. Journal of Reading. 11 (1967)

11. Wark, D. and Raygor, A. R.

Operant Conditioning Techniques for Reading Instruction.

In press.

12. Yarington, David

A Study of the Relationships Between the Reading Done by College Freshmen and Aptitude and Scholastic Achievement.

Coop Research Project, #5-8421, Athens, Ohio: Ohio University.