DOCTORAL DISSERTATION SERIES

TITLE AN EXPERIMENTAL STUDY OF SAMPLING PROCEDURES FOR THE DETERMINATION OF ACHIEVEMENT TEST NORMS IN A CITY SCHOOL SYSTEM

AUTHOR J. WILMER MENGE

UNIVERSITY OF MICHIGAN DATE 1948

DEGREE PH. D. PUBLICATION NO. 1163

1 2 3

U M UNIVERSITY MICROFILMS

ANN ARBOR · MICHIGAN

AN EXPERIMENTAL STUDY OF SAMPLING PROCEDURES FOR THE DETERMINATION OF ACHIEVEMENT TEST NORMS IN A CITY SCHOOL SYSTEM

by

J. Wilmer Menge

Committee in Charge:

Professor Clifford Woody, Chairman
Professor Irving H. Anderson
Professor Paul S. Dwyer
Professor Rensis Likert
Professor Theodore M. Newcomb
Professor Willard C. Olson

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the University of Michigan

1948

ACKNOWLEDGMENTS

This study was made possible through the cooperation of my colleagues in the Department of Instructional Research of the Detroit Public Schools, who made available to me all of the original data secured from a city-wide testing survey.

Acknowledgment of their encouragement and counsel is here made to Dr. Clifford Woody, Dr. Irving H. Anderson, Dr. Paul S. Dwyer, Dr. Rensis Likert, Dr. Theodore M. Newcomb and Dr. Willard C. Olson, the committee which directed the preparation of my dissertation.

Special acknowledgments are due to:

Professor Clifford Woody, who as chairman of the committee which supervised the conduct of this study, gave helpful advice in planning the form of the dissertation and directed me to sources of assistance in the solution of specific technical problems that arose during the course of the investigation

Roe Goodman, in charge of the Sampling Section, Survey Research Center of the University of Michigan, who helped me to gain new insight in the application of sampling theory to practical problems and who read the entire manuscript in preliminary form making many constructive criticisms and suggestions

Eugene Brown, in charge of the Tabulating Division, Central Payroll Department of the City of Detroit, who gave invaluable assistance in planning the details of machine tabulation procedures used

Virginia W. Menge who made all of the statistical computations which form the basis of the conclusions reached in the investigation

Mary Jane Boyd whose judgment, artistry and skill as a typist have given this report its final form.

J. W. M.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

TABLE PAGE

TABLE PAGE

CHAPTER I

INTRODUCTION TO PROBLEM

A test norm as commonly defined is a central tendency--usually the mean or median--of a distribution of scores from a defined population. The main purpose in determining test norms is to provide objectively defined reference points that may be used in interpreting scores made by different individuals and groups. A norm, in itself, represents neither a "minimum standard" nor a "goal" of achievement. In the light of present knowledge of individual differences it is obvious that a norm does not provide a very reliable prediction of achievement for a given pupil. It gives a somewhat better prediction of average achievement for a given group of pupils, but even here a considerable margin of prediction error is to be expected because of the systematic differences commonly found among groups. Achievement test norms are useful tools primarily because they provide a common terminology for describing levels of performance.

The descriptive terminology which constitutes the language of norms has a direct empirical reference--norms are to be discovered, not created, by research methods. Furthermore, the boundaries within which such discoveries may be made are restricted by the nature of the research methods used. A norm is only as good as the procedures used in deriving it.

The Usefulness of Sampling in Determining Test Norms. There are two ways to determine a norm. One way is to test every individual in the particular population under consideration and compute the central tendency desired. The other way is to test a sample, so chosen as to represent the total population, and compute the central tendency of the sample. This second method is nearly always used in deriving norms for relatively large populations of elementary and secondary school pupils. As a matter of fact most of the research in education involves the "...drawing of inferences about a population from what is known of a sample taken to represent that population."[1]

Every-pupil survey testing is expensive. Such census surveys are probably never worth what they cost apart from the instructional use teachers are able to make of the test scores. The problem of cost, therefore, is one of the reasons for using some method of sampling when the central purpose in administering a test is to determine a norm. Quite apart from the question of financial outlay required, the administration of a test requires an outlay of pupil time. It would be hard to justify expenditure of the countless hours needed to administer to all pupils all of the good tests

---

[1] E. F. Lindquist, Statistical Analysis in Educational Research, p. 1. Chicago: Houghton Mifflin Company, 1940.

available, even for the worthy purpose of determining adequate norms. Sampling is to be regarded, therefore, not as a second-choice substitute for a better method, but rather as a scientific means of getting information which would otherwise be inaccessible.

<u>The Difficulty of Securing Random Samples of Individual Pupils</u>. Populations of pupils dealt with in educational research are usually such that it is highly impracticable if not impossible to draw from them truly random samples of individuals. A random sample is one selected in such a way that (a) every unit of the population has the same chance of being selected, and (b) the drawing of each unit in the sample is independent of the drawing of any other unit.[2] This means, for example, that in selecting a random sample of 100 cases from all 12th grade pupils in Michigan high schools there would have to be 100 separate drawings. Secondly, the mechanics of selection would have to be set up in a manner that insured an equal chance of choosing any 12th grade pupil in any one of the several hundred high schools on each of the 100 drawings.

Lindquist suggests that a useful concept of 'randomness' is to think of a sample as one "...so drawn that all other

[2] George W. Snedecor. "Design of Sampling Experiments in the Social Sciences," Journal of Farm Economics, Vol. 21 (1939), p. 849.

possible combinations of an equal number of members from the population had an equal chance to constitute the sample drawn."[3] For instance, there is an almost limitless number of different combinations of 100 twelfth grade pupils in Michigan high schools. One combination consists of 3 pupils from Ypsilanti, 6 from Marquette, 37 from Ann Arbor, and 54 from Detroit. If the sampling is strictly random, this combination must have the same chance of being selected as any other. In actual practice of course, this particular combination might have no chance of being drawn. The only feasible procedure would be to secure the cooperation of a few schools in advance and then select a number of cases from each one.

Even in a large city school system the practical obstacles to securing a random sample of individual pupils are almost insurmountable. Pupils are accessible only as members of intact groups, i.e., schools and classes. It would be extremely difficult and costly to set up procedures whereby each of, say, 10,000 sixth grade pupils in 150 different schools would have an equal chance of being drawn in a sample of 500. If the investigation called for administering a test to each of the 500 pupils, a teacher from whose class two pupils were drawn would probably find it easier to give the test to the entire class rather than single out the two pupils

---

[3] E. F. Lindquist, op. cit., p.3.

and see that the test was administered to them under standard conditions while the rest of the class was engaged in some other activity. The conditions under which data are to be secured from members of the sample thus place a further practical limitation upon the use of methods of random selection of individual pupils.

The Mathematical Bases of Reliability Estimates for Sample Data. The concern here about randomization in selecting a sample is due to the fact that in deriving a test norm the central purpose of sampling is to secure results that will yield inferences about the larger population. Inferences, of course, may actually be made from any sample regardless of how it is selected. At the present time, however, there are no known methods for describing objectively the degree of confidence to be placed in such inferences from sample data unless there has been random selection at some point in the sampling process.

Some sampling methods do not employ the random technique but call for the selection of samples that conform to chosen criteria. Such methods are referred to as "purposive sampling".[4] Lindquist points out that any such method "...suffers from the serious disadvantage that it does not

[4] Jerzy Neyman. "On Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection," Journal of the Royal Statistical Society, Vol. 97 (1934), pp. 558-606.

permit any objective description of the reliability of the results obtained."[5] Smith and Duncan[6] also emphasize random selection as a prerequisite for drawing inferences from a sample. Samples obtained by any other method "...may be 'thought' to be good representations of the population, but just how good is indeterminate. Random sampling is the only method so far devised that permits logical inferences about a given population."[7] Neyman expresses the same idea more succintly in saying "there is no room for probabilities, for standard errors, etc., where there is no random variation or random sampling."[8]

The manner of selecting the sample is, therefore, the primary factor which determines whether there can be any objective estimate at all of the reliability of a norm. In discussing this point Snedecor[9] refers to the "widespread misinformation" about the function of randomness in sampling.

---

[5]E. F. Lindquist, op. cit., p.7.

[6]J. G. Smith and A. J. Duncan. Sampling Statistics and Applications, p. 154. New York: McGraw-Hill Book Company, Inc., 1945.

[7]Ibid.

[8]Jerzy Neyman, op. cit., p. 558.

[9]George W. Snedecor, op. cit., p. 852.

> One often hears the statement, 'I am not interested in a test of significance--it is only the average that I wish to know.' The hitch is that from a sample the population mean can never be known but only an estimate of it. How reliable is this estimate? In the absence of complete information about the true mean, one can only know the fiducial probability that such parameter lies within certain limits. But the validity of this probability rests on the assumption of randomness in sampling. In modern sampling theory, estimates of an average and of its variability are not two separate problems, but are rather two aspects of the same problem of estimation, and there is no theory of estimations for samplings other than random.[10]

Results of Research on the Sampling of Pupil Populations. Random sampling of schools or of classes is a feasible alternative to sampling of individual pupils. The sampling of schools is, in fact, a common practice in determining norms. It is easy to get large numbers of cases by taking all the individuals in a few schools. When selected in this manner, however, the mere number of pupils does not provide an index of reliability of a norm. The reason is that large achievement differences between schools are known to exist.

In 1930 Lindquist[11] reported a study of the average achievement of each of a large number of high schools on several objective tests. Data from the test in English and American Literature for Grade 12 will illustrate his findings.

---

[10] Ibid.

[11] E. F. Lindquist. "Factors Determining the Reliability of Test Norms." Journal of Educational Psychology, Vol. 31 (1930), pp. 512-520.

This test was given to 3,233 pupils in 172 different high schools. "The standard deviation of raw scores...for the total 3,233 pupils was 26.26 score units. A distribution of the 172 average scores of individual schools was 15.81 score units... The variability of raw scores of individual pupils was therefore only 1.66 times as large as the variability in school averages. In other words, the differences between schools were almost of the same magnitude as the differences between individuals in all schools."[12] Among the ten tests administered in more than 150 high schools there were only four tests which showed a variability in raw scores more than twice as large as the variability in school averages. In the case of none of the ten tests was the standard deviation of raw scores as large as three times the standard deviation of school averages. The data from this study also show that, contrary to expectation, the variability in school averages appears to be almost independent of the number of pupils tested in each school. A summary of average performances of schools classified into three enrolment groups showed that the large schools (over 500 enrolment) showed only slightly less variability in average scores than did the small schools (less than 125 in enrolment).[13]

Generalizing from the data secured in this study,

---

[12] E. F. Lindquist, op. cit., p. 514.

[13] E. F. Lindquist, op. cit., p. 515.

Lindquist[14] states that a norm based upon scores obtained from a small number of schools, regardless of the number of scores obtained is likely to deviate so far from the mean of the population that little if any reliance may be placed upon it. He adds, further, that because of the obvious impracticability of securing a strictly random sample of individuals from an entire population of school children it is essential that the school rather than the pupil be considered as the basic sampling unit.

Few research workers have reported investigations of the reliability of data secured from sampling groups of pupils rather than individual members of a pupil population. An exhaustive search in the published literature of educational research covering the twenty-year period from 1928 through 1947 yielded accounts of only two such studies, in addition to the one already described, dealing specifically with the problem of determining test norms.

The first of these two reports, published in 1940, is another study by Lindquist,[15] similar to the one he reported in 1930. The basic data presented are distributions of scores on an achievement test in 'English Correctness'

---

[14] Ibid.

[15] E. F. Lindquist. Statistical Analysis in Educational Research, pp. 66-72. Chicago: Houghton Mifflin Company, 1940.

for 9th grade pupils in each of 11 Iowa high schools. These schools were selected at random from all schools of 65 to 125 enrolment that participated in the 1938 Iowa Every-Pupil Testing Program. A total of 414 pupils was tested in the 11 schools. The mean of the combined distribution is 164.3 and the standard deviation is 29.3. Lindquist shows that if this were considered as a random sample of pupils the standard error of the mean would be estimated as 1.44. He goes on to show the fallacy of assuming this to be a random sample of individuals and states that the sample should be considered as consisting of 11 schools rather than 414 pupils. The remainder of the analysis of the data is devoted to showing that the use of appropriate statistical procedures gives 4.03 as the estimated standard error of the mean rather than 1.44, the original estimate based on the assumption of random selection of individual pupils.

The second of the two investigations dealing with analysis of sampling error of test means associated with selecting cases in 'groups' rather than as 'individuals' is described by Marks.[16] In this study Marks reviews the sampling plan used in collecting data for the revision of the Stanford-Binet scale. He first calls attention to the fact that the plan involved the use of 'cluster' methods of

---

[16]Eli S. Marks, "Sampling in the Revision of the Stanford-Binet Scale," Psychological Bulletin, Vol. 44 (1947), pp. 413-434.

sampling. The primary sampling unit was a community, each community selected consisting of a cluster of individuals. Since each individual was not selected independently from a total population, the conventional formula for estimating the standard error of the mean does not apply. Marks applies two alternative statistical procedures to the published data on the Stanford-Binet revision to illustrate the magnitude of the error that can be made by applying formulae based on unrestricted random sampling to data obtained by cluster sampling. He states that the sample used in this study is not an extreme case of such error even though it is found that the standard error for cluster sampling is three times the standard error for random sampling of the same number of individuals. "Many studies use data from one or two groups ...to draw conclusions about the whole population...or even all human beings. In this case the standard error obtained (from the proper formula)...may be 50 to 100 times greater than that obtained from...(the formula commonly used). Use of the "correct" formula...will make supposedly significant differences vanish more rapidly than a quart of ice cream at a children's party."[17]

Both Lindquist and Marks point out that systematic differences between groups or "positive intra-class correlation" with respect to the variable studied is the

---

[17] Eli S. Marks, op. cit., p. 426.

reason why cluster sampling methods give a larger sampling error than unrestricted random sampling of the same number of cases. An intact group of individuals tends to be more homogeneous than the same number of individuals drawn at random from the total population. Walsh[18] has shown that even a very small amount of intra-class correlation can have a very substantial effect on estimates of the reliability of a mean. The effect of this correlation increases with the sample size, i.e., with the number of individuals in the cluster. When the average size of clusters is moderately large (100 or more), serious misinterpretations will result from failure to take into account a seemingly insignificant degree of group homogeneity.[19] It is not at all uncommon in educational research to take groups of 100 or more cases from a single community or even from a single school. This method of sampling is extremely useful. Valid estimates of sampling error can be derived if the design provides for random selection of clusters.

The Need for Research on the Practical Application of Cluster Sampling Methods. If test norms are to be determined at all they will have to be derived from data secured by

---

18J. E. Walsh. "Concerning the Effect of Intra-Class Correlation on Certain Significance Tests," Annals of Mathematical Statistics, Vol. 18 (1927), pp. 77-96.

19Eli S. Marks, op. cit., p. 426.

sampling groups of pupils, not individual pupils. Secondly, if norms are worth establishing, they are worth establishing adequately. Just what constitutes 'adequacy' in a norm is to some extent, at least, a matter of opinion. Two of the specifications for a good norm, however, would not be questioned by research workers. The first is that the norm should represent a reasonably good estimate of the population to which it refers. The second is that there should be an objective definition of the margin of sampling error to be taken into consideration in making inferences from the sample to the population. Cluster sampling methods provide the only feasible means of securing cases that are widely representative of a population of pupils. The application of appropriate 'cluster formulae' for determining sampling error provides the means of getting objective estimates of error.

Apart from the previously mentioned studies of Lindquist and Marks the educational research literature on this problem is practically non-existent. Something is said in textbooks about the desirability of securing "unbiased samples" but the questions of how to select such a sample in actual practice and how to evaluate its efficiency are usually dismissed with little more than passing comment. For example, in a recently published text Odell[20] devotes approximately

---

[20] C. W. Odell. An Introduction to Educational Statistics. New York: Prentice-Hall, Inc., 1946. Pp. xii + 269.

two pages to the problem of selecting samples. In comparing alternative sampling designs he uses the following illustration:

> "...let us suppose that a superintendent of schools of a populous county wishes to measure the general achievement of pupils completing the elementary school, but cannot well expend the money and effort necessary to test more than one-fifth of such pupils. A good method of selecting those to be tested would be to ...arrange the names of all alphabetically and take every fifth one. Even better would be to apply the same method by schools or districts rather than to the country as a whole...intermediate between good and bad would be to select at random one-fifth of the schools and to test all graduating pupils in them."[21]

This third method (random selection of clusters) happens to be the very one which is most practical to use. Odell gives no directions for evaluating the efficiency of such a sampling design beyond the statement that the plan is "intermediate between good and bad." In commenting further on this type of design he states that

> "...a sample of 20 classes of twenty-five pupils each is much less likely to be representative of 10,000 high-school freshman in a large city system than is one of 500 chosen by taking every twentieth one alphabetically. This is true even if the 20 classes are a thoroughly random selection of all such classes... Such a sample will tend to be intermediate in representativeness between one of 500 pupils properly chosen and one of 20 pupils, also properly chosen."[22]

Again, the cluster design is said to be "intermediate in

---

[21] C. W. Odell, op. cit., p. 229.

[22] C. W. Odell, op. cit., p. 230.

representativeness" but there are no suggestions as to how to make an objective estimate of its efficiency. Odell has been quoted at length because he gives a more detailed discussion of the problem of selecting samples than is commonly found in standard texts on educational statistics.

The literature on specific sampling techniques in the related field of psychology adds little to what is found in education. When McNemar[23] set out to prepare a critique of the applications of sampling theory in social psychology, he found that detailed descriptions of specific techniques were so few such a critique would not be justified. Because of the lack of materials on the problem he undertook to study, his extensive summary deals instead with "...certain general principles of sampling, somewhat unfamiliar to psychologists which could be gleaned from the rather widely scattered literature of statistical methodology."[24] McNemar expresses disappointment at "the paucity of techniques for drawing and checking a representative sample," and concludes with the opinion that "...the confidence to be placed in the results of a study should vary directly with the amount of information concerning the sampling and experimental techniques rather than inversely with the

---

[23]Quinn McNemar. "Sampling in Psychological Research," Psychological Bulletin, Vol. 37 (1940), pp. 331-65.

[24]Quinn McNemar, op. cit., p. 362.

square root of the number of cases."[25]

The literature available at the present time thus gives little help to the research worker in a city school system on the task of designing and actually carrying out a sampling plan to determine average performance on an achievement test. Such city-wide surveys are usually conducted for one of two purposes: (a) to establish "local" norms to be used in the future by individual teachers in the city, or (b) to compare the average achievement of pupils in the city with some previously established norm. In either case it is essential to get a valid estimate of the precision of the sample results. Estimates of precision, however, have no meaning apart from control of the procedures used in drawing the sample.

It is the purpose of this study, therefore, to apply appropriate sampling theory to the dual problem of (a) selecting samples which consist of existing groups of pupils, i.e., schools and classes, and (b) deriving objective measures of the precision of sample results. The specific methods by which samples are drawn will be described in detail. The sampling designs to be used are relatively simple because they were developed with a view to future use in dealing with similar problems where one has to work under the usual practical administrative

[25]Quinn McNemar, op. cit., p. 363.

restrictions. In any large city school system it should be quite feasible to determine test norms, with a measurable degree of precision, by the use of sampling procedures similar to those described in this report.

CHAPTER II

STATEMENT OF PROBLEM

Background of Problem. Certain practical difficulties encountered in administering achievement tests on a city-wide basis in Detroit Public Schools provided the stimulus for making this study. Test data are collected periodically on a city-wide basis for two purposes: (a) to establish city norms for locally constructed tests designed for use in the regular instructional program, and (b) to determine the average accomplishment of pupils in the city on 'survey tests' (commercially published tests used in conducting achievement surveys).

Sampling procedures have been used in recent years in securing test data for both these purposes. The pupil population in a given grade has been sampled by selecting a certain number of schools and then testing all pupils in each of the schools chosen. A serious limitation of the procedures used in the past has been that the methods of sampling did not provide for any way of finding out how well the obtained results actually represented the population from which the samples were drawn. The sampling designs were "purposive" in type and consequently there could be no computation of dependable estimates of sampling error. For this reason it was impossible to determine the degree of significance of differences between the results of surveys conducted at different times. A central objective of the investigation reported here has been to apply procedures which overcome this limitation.

Definition of Problem. The problem under investigation may be stated as follows: To determine among several measurable designs for sampling a pupil population which one yields results most closely representing the population from which samples are drawn.

Three specific limitations are to be imposed on the study of this problem. First, the only sampling unit employed will be a normally constituted group of pupils, i.e., a school or a class. Second, the population will consist of pupils enroled in one grade in the public schools of Detroit. Third, only one variable will be studied--score on a reading achievement test.

Techniques to be Used. Four different methods of drawing samples will be applied in order to determine which one yields results most closely representing the total population. The four methods to be used are:

a. simple random sampling of schools

b. stratified random sampling of schools

c. simple random sampling of classes

d. stratified random sampling of classes

The mean score is to be used as the index of "representativeness" of each sample. Relative efficiencies of the several sampling methods will be determined by making comparisons among the respective standard errors of the means obtained by applying each method in turn to the designated population.

The standard error for each method of sampling will be derived by two different procedures. First, repeated sampling will be used to produce a distribution of means from which an estimate of the standard error of the mean may be obtained. Second, the appropriate cluster sampling formula will be applied to the population data. This formula[1] which gives a very close approximation to the true standard error is

$$\sigma^2_{\bar{x}'} = \frac{M - m}{(M - 1)m} \; \frac{\sum_{i=1}^{M} \left[N_i^2 \; (\bar{x}_i - \bar{x})^2 \right]}{\bar{N}^2 M}$$

$\sigma^2_{\bar{x}'}$ = the square of the standard error of the mean of the sample

M = the total number of clusters (schools or classes) in the population

N_i = the number of individuals (eligible for the population) in a given school or class

$\bar{x}_i$ = the mean score for the N_i individuals in a given school or class

m = the number of clusters in the sample

---

[1] Eli S. Marks, "Sampling in the Revision of the Stanford-Binet Scale," Psychological Bulletin, Vol. 44 (1947), p. 420.

$\bar{X}$ = the mean score for the population

$\bar{N}$ = the average number of individuals per school (or class) in the population

It will be observed that this formula gives the standard error for a sample of clusters from a finite population.[2] In the form presented here it can be used only where data are available for the entire population. A modification of the formula which gives an estimate of $\sigma_{\bar{X}}$, derived from the data of the sample itself will be presented later. A second modification of the formula to be given later is applicable to designs which employ stratification of the sampling units.

Description of the Population. The population to be sampled consists of all pupils enrolled in the Detroit Public Schools in February 1947. February was the first month of the second semester of the school year 1946-47. Official Monthly Membership Reports[3] from individual schools show that at that time a total of 8139 grade 8A pupils were listed on the class rolls of 237 different classes in 97 different schools.

The geographical distribution of these schools covers the entire city. The economic level of individual school neighborhoods thus ranges from the lowest to the highest to be found within the city limits. Detroit, in common with

---

[2] M. H. Hansen, and W. H. Hurwitz, "On the Theory of Sampling from Finite Populations," Annals of Mathematical Statistics, Vol. 14 (1943), pp. 333-362.

[3] Detroit Public Schools, Form 533, for February 1947.

with other large industrial communities, has many separate neighborhood concentrations of nationality groups and racial groups. One or more of the 97 schools serve the children of every such group in the community.

Compulsory attendance laws keep practically all children in school at least through grade 8 (the modal age for grade 8A is a little under 14 years). The 8139 pupils, therefore, represent practically all children in the city (not including those in parochial and private schools) eligible for membership in this grade. The parents of these pupils are factory workers, bankers, bus drivers, college professors, labor leaders, ministers, carpenters, musicians, barbers and all the rest.

The level of reading achievement for each member of the pupil population was determined by having special examiners administer the Stanford Advanced Language Arts Tests (sub-test 1: Paragraph Meaning, and sub-test 2: Word Meaning) of Form DM[4] to each of the 237 classes.[5] Every grade 8A pupil present was tested. The total number tested was 7724. This is 94.9 per cent of the total enrolment (8139). Because of the administrative problems involved it was not possible to have the special examiners return to the schools

---

[4] Stanford Advanced Language Arts Test, Chicago: World Book Company, 1941. (A copy of this test appears in Appendix B.)

[5] Detailed descriptions of how the test was administered and scored and how the original data were recorded are given in Appendix A.

later and give the test to the five per cent (415 pupils) who were absent on the original testing date. For the purposes of this study, therefore, the 7724 pupils tested are considered to constitute the total population.[6] Similarly, the pupils actually tested in each class and in each school will be assumed to constitute the "total populations" of the individual classes and schools respectively.

The distribution of test scores achieved by the 7724 pupils is shown in Table I. The mean of the distribution is 66.74 and the standard deviation is 9.05.[7] Individual scores range from 42, which is three points above the lowest possible "equated score" on the test, up to 91, one point below the maximum possible score.

According to the "national norms" for this test, reported by the publisher,[8] a score of 42 is average for pupils enroled in the upper half of grade 4. The relative achievement of the

---

[6] The chronological ages and previously obtained intelligence test scores for the 415 absentees were compared with similar data for the entire enrolment (8139). The absentees were, on the average, slightly older and they made slightly lower scores on an intelligence test than the population as a whole. It is probable that had they been given the reading test their average score would have been slightly lower than the average for the 7724 pupils.

[7] All computations, throughout this study, were made from ungrouped data. In showing distributions, as in Table I, it is usually necessary to group the data for reasons of convenience in presentation. However, all sums and all sums of squares used in various computations are derived from the original data for each pupil recorded on 7724 IBM cards.

[8] Directions for Administering: Stanford Advanced Language Arts Test, p. 6. Chicago: World Book Company, 1941.

TABLE I

Distribution of Total Scores
(Average of Test 1 and Test 2) Made by Grade 8A
Pupils on Stanford Advanced Language Arts Tests, Form DM

Total Score*	Number Pupils
90-92	12
87-89	79
84-86	167
81-83	304
78-80	457
75-77	629
72-74	728
69-71	797
66-68	996
63-65	965
60-62	825
57-59	678
54-56	540
51-53	340
48-50	153
45-47	43
42-44	11
Total	7724
Mean	66.74
S. D.	9.05

*These are "equated scores" derived from raw scores by means of a conversion table which appears on the test answer sheet. See Appendix B.

one pupil represented in Table I who made this score is four full grades below the grade 8A where he is actually enroled. The highest score for which a norm is reported[9] is 75, representing average achievement for pupils just beginning grade 11. The twelve grade 8A pupils, shown in Table I, with scores of 90 and above, almost certainly have a reading ability superior to that of the average high school graduate and perhaps superior to that of the average college freshman.

The mean for the 7724 pupils is a fraction of one score point above the reported published norm for pupils in this grade. About 27 per cent show achievement in reading which is two full grades or more above the norm and about 13 per cent show a level of achievement two full grades or more below the norm. Test surveys conducted in a single grade typically show this wide range of achievement among individual pupils.

The variance of pupil scores within an individual school is, in general, somewhat less than the variance for the total population of pupils. Summaries of data showing the mean and standard deviation for each school will be given in connection with descriptions of the different methods of sampling by school groups.

---

[9]Ibid.

CHAPTER III

RESULTS FOR SAMPLING BY SCHOOL GROUPS

There are certain practical advantages to be gained by using the school, rather than a sub-group within the school, as the primary sampling unit. One advantage is that every one of the separate units (schools) eligible for inclusion in the sample is known and can be located geographically before any field work is undertaken. This makes it easy to exercise complete control over the method of selecting the sample. The field workers, i.e., principals and teachers, do not exercise any choice in picking the membership of the sample. Since there is no sub-sampling within the school groups, there is no possibility of biased selection of individuals. Once the sample of schools is chosen, every individual pupil to be included is exactly specified. In this study, for example, the individuals constituting a given sample are all pupils enroled in grade 8A in the schools selected for that sample.

There is a second major advantage in using the school as the primary sampling unit. The school is also the primary administrative unit in a city school system. Arrangements need to be made with the administrative head of each school in which pupil data are to be secured. Usually it is no more difficult to make arrangements with a school principal for giving a test to all of the pupils in the school who are enroled in a given grade or subject than it is to get his cooperation in securing the necessary data on a limited number of individuals or sub-groups. Insofar as mere numbers of pupils contribute at all to the reliability of results obtained from a sample it is worthwhile to take into account the large numbers of pupils that may be secured from a relatively small number of schools. For instance, it is possible to get, say, 1500 grade 8 pupils from a sample of 20 schools by taking all pupils in this grade in each school. Arrangements would need to be made with 20 different school principals to furnish supervision in gathering the desired data. To get that number of cases by any feasible plan of sub-sampling within the schools would almost certainly increase several fold the time and effort involved in making the necessary advance arrangements with the additional administrators whose schools would be included in the sample.

In the light of these practical considerations the following four different designs were developed and applied using the school as the sampling unit:

1. Simple random sampling of schools
2. Stratification of schools by size of enrolment--random sampling of schools within each stratum
3. Stratification of schools by geographic location--random sampling of schools within each stratum

4. Primary stratification of schools by geographic location; sub-stratification of schools by size of enrolment--random sampling of schools within each sub-stratum.

In planning the solution of a sampling problem it is necessary to provide for the determination of the degree of error of sample results and to consider the cost in time, effort and money required to achieve the level of accuracy desired. Alternative designs (i.e. procedures for actually selecting the units to be included in the sample) are studied from the points of view of administrative feasibility, cost, etc. The design which is most efficient in a statistical sense is not always the one most economical to use in a given situation because of these other considerations.

It is very helpful to have in advance some detailed information about the population to be sampled so that alternative designs can be set up and evaluated before a sample is actually drawn. It is for this reason that the next section of this chapter gives (a) a detailed description of the organization of the pupil population in school groups and, (b) the distribution of mean test scores for the 97 schools.

I. SUMMARY OF DATA BY SCHOOL GROUPS FOR TOTAL POPULATION

Number of Pupils Enroled and Number Tested in Each School. The smallest grade 8A enrolment among the 97 schools was three pupils. The largest was 459 pupils. Among the 97 schools the range in number tested was from 3 to 435.

Table II gives the distribution of enrolments and the distribution of sizes of groups tested in the respective schools.

Six of the schools enrol fewer than 20 pupils each in this grade. In each of 58 schools there are from 20 to 59 enroled, 12 schools enrol from 60 to 99 each and in each of 21 schools the enrolment is 100 or more. Five out of the 21 schools enrol more than 300 pupils. The mean size of the grade 8A enrolments is 83.9 pupils.

The distribution of numbers of pupils tested in the respective schools as shown in Table II corresponds closely to the distribution of enrolments. In the school which enrols only three pupils in the grade, all three were tested. In the largest school 435 were tested (94.8 per cent) out of an enrolment of 459. The mean size of the school groups tested is 79.7 pupils.

The only marked difference in the two distributions--number enroled and number tested--is found in the case of the lowest step in Table II which shows that each of six schools has an enrolment of less than 20, whereas in each of 11 schools fewer than 20 pupils were tested. The large difference in the two frequencies for this step in the table is due to two facts. First, in the case of three schools enroling exactly 20 pupils each there was one or more absent on the testing date. Second, in each of two schools enroling 24 pupils there were five absent on the testing date giving a total of 19 tested for each of these schools. The percentage of absence in these latter two cases was far higher than in any of the other schools.[1]

TABLE II

Sizes of Grade 8A Groups in 97 Schools

Size of Grade 8A Group (No. of Pupils)	Number of Grade 8A School Groups	
	Enroled	Tested
440-459	1	
420-439	1	1
400-419	1	
380-399		2
360-379		
340-359		
320-330		
300-319	2	
280-299	1	1
260-279		2
240-259	2	
220-239	2	3
200-219	3	2
180-199	1	3
160-179	1	
140-159	4	4
120-139		1
100-119	2	2
80-99	3	3
60-79	9	8
40-59	24	23
20-39	34	31
1-19	6	11
Total Number of Groups	97	97
Total Number of Pupils	8139	7724
Mean Size of Group	83.9	79.7

[1]Data showing the number of pupils enroled and the number tested in each of the 97 schools is given in Table XLIV, Appendix C.

Because of the marked skewness of the distribution in Table II the mean number of pupils tested per school (79.7) is not a good representation of central tendency. In two-thirds of the schools the tested population consists of fewer than 60 pupils. The median number tested per school is 47.

Mean Test Scores for the 97 School Groups. Pupils in the school showing the highest achievement on the test made an average score of 74.0. Those in the school showing the lowest achievement made an average score of 54.0. This difference of 20 points between the means of the highest group and the lowest group is more than two times as large as the standard deviation of individual scores (9.05)[2]. The distribution of school means in Table III shows 11 schools with average scores of 72 or above and 15 with average scores below 62. The average of the 97 school means is 65.98. This is lower than the average of the 7724 individual scores (66.74)[3].

The standard deviation of school means, that is, the standard deviation of the actual distribution of the 97 school averages shown in Table III is 4.29. This is approximately one-half the magnitude of the standard deviation of individual pupil scores. The variance of school means is only slightly influenced by the factor of size of enrolment of individual schools. For example, among the 21 schools enroling more than

---

[2]See Table I, p. 24.

[3]Ibid.

100 pupils each the range in average scores was 59.5 to 72.4. Among the 20 schools enroling 30 or less the range in average scores was 54.0 to 73.2.[4]

TABLE III

Mean Scores on Reading Test for School Groups of Grade 8A Pupils in 97 Schools

School Mean Score	Number of Schools
74-75	1
72-74	10
70-71	10
68-69	13
66-67	22
64-65	15
62-63	11
60-61	8
58-59	4
56-57	1
54-55	2
Total	97
Average of School Means	65.98*
S. D. of School Means	4.29*

*Computed from ungrouped data given in Table XLIV, Appendix C.

II. RESULTS FOR SIMPLE RANDOM SAMPLING OF SCHOOLS

Method Used in Drawing a Sample. With the test data summarized separately for each of the schools, the application of a plan of sampling became a laboratory problem. First, each

[4]The number tested, sum of scores, and mean score for each school are given in Table XLIV, Appendix C.

school was given a two-digit code number. The numbers used were 01 to 79 inclusive and 81 to 98 inclusive--a total of 97 code numbers. The selection of a random sample of schools was then merely a task of choosing at random, from among the codes, the number of schools desired for the sample. A table of random numbers was used in selecting each code number (school) to be included in a sample. The specific manner in which "random numbers" were used will be described in detail.

Mean Scores for Samples of Ten Schools. A group of ten schools was chosen by means of Fisher's[5] table of random numbers. This six page table includes 7500 two digit numbers. In preparing to draw the required number of schools, the "starting point" in the table was selected in haphazard fashion, without looking at any number in the table. This starting point was specified as the number to be found in a designated row and in a designated column on a given page. It was further decided, in advance, that the first ten unlike two digit numbers, reading down from the starting point would constitute the sample of ten schools. The table of random numbers was then opened to the specified starting point. The first ten numbers reading down the column were: 53, 45, 23, 25, 11, 89, 87, 59, 66 and 50. Insofar as the numbers in Fisher's table are truly random, every school had an equal

[5] R. A. Fisher and F. Yates, Statistical Tables for Biological Agricultural and Medical Research, pp. 82-87, London: Oliver and Boyd, 1938.

chance to be chosen on each drawing and the selection of each school was independent of the selection of any of the other nine schools. The ten schools are therefore a random sample of the total population of 97 schools.

After the schools were chosen, the first step in finding the mean pupil score for the entire sample was to compute the sum of the "sums of scores" and the sum of the numbers tested for the ten schools. For example, the sum of scores for school 53 is 2794 and the number tested is 42; for school 45 the sum of scores is 1941 and the number tested is 29; and so on for the other eight schools in the sample.

Table IV shows that the sum of scores made by the 697 pupils in the 10 schools is 45,990. This sum divided by the number of pupils gives 65.98 as the mean for the sample. This mean happens to be exactly the same (to the second decimal place) as the average of the <u>means</u> of the 97 schools. It also represents an estimated average score for the 7724 <u>pupils</u>. However, the "true mean" for the total pupil population is known to be 66.74. The estimate of the true mean <u>pupil</u> score derived from the sample of 10 schools is somewhat low. The absolute amount of the difference between this estimated mean of the pupil population and the actual mean of the pupil population is thus 66.74-65.98 or .76. This obtained difference of .76 is a result of sampling variation.

TABLE IV

Numbers of Pupils Tested, Sums of Pupil Scores and Sums of Squared Pupil Scores for 10 Schools Randomly Selected from a Group of 97 Schools

School	No. Pupils Tested	Sum of Scores	Sum of Squared Scores
11	39	2,864	212,520
23	53	3,820	278,918
25	75	4,876	322,474
45	29	1,941	130,973
50	19	1,143	69,769
53	42	2,794	188,962
59	22	1,448	97,254
66	42	2,536	155,020
87	155	9,351	575,355
89	221	15,217	1,064,459
Total	697	45,990	3,095,704
Mean Score		65.98	
S. D. of Scores		9.37	

Another sample of 10 schools, drawn in a similar manner, might give a mean that differed either more or less than .76 from the average for the population. The important question is how much confidence can be placed in an estimate of the population average based on data from a single sample of 10 schools. If a number of samples were drawn consisting of 10 schools each, what would be the extent of variation among the sample means themselves?

An answer to this question was worked out empirically by actually drawing 19 additional samples of 10 schools each. After each drawing of ten code numbers the schools constituting that sample were figuratively thrown back into the "hopper" so as to be eligible for selection in the succeeding sample. The selections were made from the table of random numbers described previously by continuing *down* the column of digits from the starting point. When the bottom of the column was reached, the selection continued with the top of the adjacent column to the *right*, then down this column to the bottom, and so on. The second sample chosen in this manner consists of schools 09, 10, 13, 27, 38, 52, 54, 69, 71 and 77. The numbers, of course, were not found in this order in the table.

Summaries of data for 20 such samples are shown in Table V. Sample Number 1 in this table is the sample described in detail in Table IV. It will be noted, for example, that the number of pupils, the sum of scores and the mean score opposite "Sample 1" in Table V are taken from the totals at the bottom of Table IV.

The means of the 20 samples range from 64.6 to 69.5. There are thus approximately two chances in 20 that the average score for a sample of 10 schools will differ from the population mean (66.7) by as much as two points. The highest of the sample means is almost three points above the average for the population (69.5-66.7 = +2.8), and the lowest sample mean is about two points below the population average 64.6-66.7 = -2.1). Sample Number 1 with an average (65.98) which is .76 below the population mean is not an unusual one at all. Three out of the other 19 samples have lower means than Sample 1.

TABLE V

Numbers of Pupils, Sums of Pupil Scores, and Mean Scores for 20 Samples Each Consisting of 10 Schools

Sample	No. of Schools	No. of Pupils	Sum of Pupil Scores	Mean Score
1	10	697	45,990	65.98
2	10	439	29,595	67.41
3	10	886	57,215	64.58
4	10	809	53,853	66.57
5	10	1,162	79,422	68.35
6	10	580	38,173	65.82
7	10	1,004	66,773	66.51
8	10	783	53,293	68.06
9	10	412	28,099	68.20
10	10	603	40,315	66.86
11	10	615	40,745	66.25
12	10	914	63,557	69.54
13	10	675	44,342	65.69
14	10	1,362	91,849	67.44
15	10	766	51,263	66.92
16	10	1,009	68,902	68.29
17	10	675	45,465	67.36
18	10	1,077	73,973	68.68
19	10	1,055	70,990	67.29
20	10	698	47,839	68.54
Mean		811.05		67.22

The standard deviation of the 20 values representing sample means in Table V was computed by the formula[6]

$$\sigma_{\bar{x}'} = \frac{1}{N}\sqrt{N \sum \bar{x}'^2 - (\sum \bar{x}')^2}$$

where $\bar{x}$ = the mean of one sample

N = the number of samples (20)

---

[6]Since the true mean of the population is known, the formula $\sigma_{\bar{x}'} = \sqrt{\frac{\sum(\bar{x}' - \bar{x})^2}{N}}$ could have been used here instead of the one given in the text above. The latter formula was employed in order to confine the analysis to data secured from the samples only.

and $\sigma_{\bar{x}}'$ = the standard deviation of the 20 sample means which, in this case, is an estimate of the "standard error of the mean".

The computation is as follows:

Sum of the means = 1,344.33
Sum of the means squared = 90,389.42
Substituting in the formula, with N = 20

$$\sigma_{\bar{x}}' = \frac{1}{20}\sqrt{565.25}$$

$= 1.19$ an empirically determined estimate of the standard error of the mean for samples of 10 schools.

An unbiased estimate of $\tilde{\sigma}_{\bar{x}}'$ for a very large number of samples drawn in the same manner as these 20 is found by multiplying the obtained $\sigma_{\bar{x}}'$ (1.19) by $\sqrt{\frac{N}{N-1}}$, where N equals the number of cases, i.e., the number of samples.[7] The result gives $\sigma_{\bar{x}}' = 1.22$.

It is pertinent at this point to compare the empirically determined standard error of the mean for a sample of 10 schools with the estimate of $\sigma_{\bar{x}}'$ that may be derived from a single sample. First, for the purpose of studying the effects of selecting clusters (i.e., schools) rather than individual pupils it will be assumed that the group of 697 pupils constituting Sample Number 1 is equivalent to a randomly selected group of individuals. On the basis of this assumption the estimated standard error of the mean for Sample 1 is determined by computing the standard

[7] Charles C. Peters and Walter R. Van Voorhis, _Statistical Procedures and Their Mathematical Bases_, p. 132. New York: McGraw-Hill Book Company, 1940.

deviation of the 697 scores and dividing this by the square root of 696. The data required are as follows:

$$\sigma_{x_1} = 9.37$$

$$\sqrt{696} = 26.38$$

$$\text{estimated } \sigma_{\bar{x}}' = \frac{9.37}{26.38}$$

$$= .36$$

This estimated value of the standard error derived from the data of a sample is less than one-third the magnitude of the empirically determined standard error (.36 as compared with 1.22).

The standard deviation of scores and the estimated standard deviation of the mean (assuming random selection of individuals) were computed for the other 19 samples. The results are given in Table VI. The average of the 20 estimates of $\sigma_{\bar{x}}'$ shown in the column on the extreme right of the table is .34.[8] By comparison with the empirically determined estimate (1.22), a marked bias in the direction of underestimation is clearly shown in this array of standard errors. Only two of the twenty are even as large as one-third of the empirically determined value. The empirically determined value is more than five times as large as the smallest of the twenty estimated values.

---

[8] This average was computed by the formula $\sqrt{\frac{\sum \sigma_{\bar{x}}^{2}{}'}{N}}$ with N = the number of samples

TABLE VI

Number of Pupils, Standard Deviation of Pupil Scores and Estimated Standard Error of Mean for 20 Random Samples Consisting of 10 Schools Each

Sample	No. of Schools	No. of Pupils	S. D. of Scores	Est'd.* $\sigma_{\bar{x}}$'
1	10	697	9.37	.36
2	10	439	9.18	.44
3	10	886	9.88	.33
4	10	809	8.29	.29
5	10	1,162	9.90	.29
6	10	580	9.64	.40
7	10	1,004	9.42	.30
8	10	783	9.13	.33
9	10	412	8.77	.43
10	10	603	8.64	.35
11	10	615	9.34	.38
12	10	914	9.19	.30
13	10	675	7.91	.30
14	10	1,362	8.66	.23
15	10	766	8.67	.31
16	10	1,009	9.56	.30
17	10	675	8.97	.35
18	10	1,007	9.09	.28
19	10	1,055	8.66	.27
20	10	698	8.58	.32
Mean				.34

*Estimated standard errors of sample means based on the assumption of random selection of individual pupils.

These results are consistent both with experimental findings reported by Lindquist[9] and by Marks[10] and with the conclusions drawn from theoretical analyses presented by Hansen and Hurwitz[11] and by Walsh[12]. A sample drawn by taking

[9] E. F. Lindquist, op. cit.

[10] Eli S. Marks, op. cit.

[11] M. H. Hansen and W. H. Hurwitz, op. cit.

[12] J. E. Walsh, op. cit.

clusters of pupils (school groups) is not equivalent to a sample chosen by unrestricted selection of individuals even though the clusters are selected at random. The summaries of data in the preceding paragraphs demonstrate that sampling error is greatly increased by using the school rather than the pupil as the primary sampling unit. Samples consisting of 10 schools with an average of 811 pupils per sample show a sampling standard error approximately four times as large as would be found for samples of the same size drawn by selecting <u>pupils</u> at random. It is obviously inappropriate to use the "classical" formula ($\sigma_{\bar{x}}' = \frac{\sigma_x}{\sqrt{N - 1}}$) for estimating the standard error of a mean when individual measures are selected in clusters. The results obtained from this formula are almost certain to be seriously misleading.

The fact that cluster methods increase sampling error is not a valid argument against their use. It is possible to bring the precision up to any level desired by increasing the number of clusters to be included in the sample. For example, it may require a sample of as many as 50 per cent of the schools in a given city to secure a mean test score as reliable as the mean that would be obtained from a sample of 10 per cent of the individual pupils. In practice the first of these two alternatives would actually be preferable to the second one for reasons that have been given earlier. In either case, adequate methods of analyzing the sample results would need to be used for getting a dependable and objective estimate of the degree of precision attained. In the next section the error formula appropriate for use with samples of schools is applied to the reading test data.

Application of Error Formula for Cluster Sampling. The formula presented in Chapter II gives the standard error of the mean pupil score where there is random selection of groups of pupils but not unrestricted selection of individual pupils. The formula is applicable in any situation where the population is organized in definable groups (except where the number of groups is very small).[13] For instance, in a city-wide testing survey the "clusters" might be defined as class groups rather than school groups of pupils. Or, in a state-wide survey the groups to be sampled might be all eighth grade pupils in the different communities in the state. The average number of pupils per cluster would of course be greatest in the case of sampling by communities and smallest in the case of sampling by classes. Hansen and Hurwitz define the term, "cluster," and give illustrations of other types of groups used as the sampling unit for agricultural and marketing surveys.

> The sampling of clusters of elements refers to the sampling of units that contain more than one element. Examples of cluster sampling include the use of the city block or the county as the sampling unit where the purpose of the survey is

---

[13] Because of the fact that the formula is an approximation, appreciable error may be introduced in situations where the number of groups in the total population is quite small.

> to determine the properties of the population made up of the individual persons or individual households. In these instances the city block or county is referred to as the cluster of elements, and the individual person or household is referred to as the element.[14]

In Chapter II along with the presentation of the formula it was pointed out that its use, in the form given, called for complete data on all the clusters in the population. When completely summarized census data are at hand, there is usually no need to draw a sample to determine the properties of the population. They are already known. However, the formula is especially appropriate for this study in which data for all of the 97 schools are available. The specific problem here is not to estimate a population value from sample results but rather to compare the respective efficiencies of several alternative methods of drawing samples from that population. The use of the formula makes it possible, for example, to determine the standard error of the mean for random samples consisting of any desired number of schools.

The formula will first be applied to determine the standard error of the mean for samples of 10 schools. This is the same problem that was worked out empirically in the preceding section of this chapter. The formula as shown

---

[14] M. H. Hansen and W. H. Hurwitz, op. cit., p. 333.

earlier (Chapter II) is

$$\sigma^2_{\bar{x}}{}' = \frac{M - m}{(M - 1)m} \frac{\sum_{i=1}^{M} \left[N_i^2 \ (\bar{x}_i - \bar{x})^2\right]}{\bar{N}^2 M}$$

For a sample of 10 schools, m = 10 and M = 97 are substituted in the expression $\frac{M - m}{(M - 1)m}$. This gives $\frac{97 - 10}{(96)(10)}$ = .090625. The average number of pupils tested per school ($\bar{N}$) is 7724 ÷ 97 = 79.62887, and $\bar{N}^2$ = 6,340.757.

Determination of $N_i^2 (\bar{x}_i - \bar{x})^2$ for each of the 97 schools required a considerable amount amount of computation. The manner in which this was done will be illustrated using schools 01, 02, 03 and 98.[15] The data needed (N_i, Σx_{ij} and $\bar{x}$) in making the computations for each of these schools is shown in Columns 2 and 3 of Table VII. The mean of the pupil scores ($\bar{x}$) for the entire population is given at the bottom of Column 2 in the Table.

The heading of Column 5 in Table VII shows the product of N_i^2 and $(\bar{x}_i - \bar{x})^2$ to be expressed by $(\Sigma x_{ij} - N_i\bar{x})^2$. Since, by definition, $\Sigma x_{ij} = N_i\bar{x}_i$, the two expressions are identical except for their respective forms.[16] The latter

---

[15] See note at the foot of Table VII.

[16] By performing the indicated operations and substituting Σx_{ij} for $N_i\bar{x}_i$ it is seen that $N_i^2(\bar{x}_i - \bar{x})^2 =$

$$\left[N_i \ (\bar{x}_i - \bar{x})\right]^2 = (N_i\bar{x}_i - N_i\bar{x})^2$$
$$= (\ \Sigma x_{ij} - N_i\bar{x})^2$$

form was used in order to simplify the actual computation. The sum of scores for each school ($\sum x_{ij}$) was available as original data. Therefore, in getting the squared, weighted deviation of each school mean from the population mean it was unnecessary to get the square of the number of pupils (N_i^2) and the square of the deviation from the population mean $\left[\sum(\bar{x}_i - \bar{x})^2\right]$ separately.

TABLE VII

Squared, Weighted Deviations of School Means from The Population Mean: Illustration of Computation for Four Schools and Sum for 97 Schools

1	2	3	4	5
School	No. Pupils Tested (N_i)	Sum of Pupil Scores (Σx_{ij})	Weighted Deviation ($\Sigma x_{ij} - N_i\bar{x}$)	Squared, Weighted Deviation ($\Sigma x_{ij} - N_i\bar{x})^2$
01	37	2,381	- 88.20388	7,779.924
02	39	2,746	+143.32564	20,542.239
03	27	1,619	-182.85148	33,434.664
..	..	.....	....	....
..	..	.....	....	....
..	..	.....	....	....
98*	234	15,673	+ 56.95384	3,243.740
Sum	7724	515,463	+ .00624	18,574,360.556
Mean	79.62887	66.73524	+ .00006	

*The code number, 98, represents the 97th school. The codes skip from 79 to 81 leaving 80 as a blank. See page 33.

For school 01, the sum of scores (2,381), minus the product of the number of pupils (37) and the population mean (66.73524), gives - 88.20388 which is ($\sum x_{ij} - N_i\bar{x}$) as shown in column 4 of Table VII.[17] The weighted deviation for School 02 is +143.32564, for School 03 it is -182.85148, and so on for the remainder of the schools. The algebraic sum of the weighted deviations provides a check on the computations. Theoretically, except for the approximation in the formula, this sum is zero. In actual practice, of course, rounding off the population mean to a given number of significant digits and rounding off the result for each school also introduce small discrepancies.[18]

The squares of the weighted deviations were next computed. Results for the four illustrative schools are shown in column 5 of Table VII. Since only the sum of the squares of the deviations for all of the 97 schools is required for use in the standard error formula, there was no need to transcribe the results for individual schools. These results are in the table only for purposes of illustration. The sum for 97 schools is 18,574,360.556. With this value determined,

---

[17]This series of operations for each school was done without any intermediate transcription of numbers, using a Friden calculating machine equipped with automatic negative multiplication.

[18]At the foot of column 4 in Table VII it will be seen that the average of the 97 computations is accurate to four decimal places.

all of the quantities required for substitution in the standard error formula are known. The substitutions which give the standard error of the mean score for a sample of 10 schools are as follows:

$$\sigma^2_{\bar{x}'} = \left[\frac{97 - 10}{(96)(10)}\right]\left[\frac{18,574,360.556}{(6,340.757)(97)}\right]$$

$$= (.090625)(30.1996)$$

$$= 2.7368$$

$$\text{and } \sigma_{\bar{x}'} = 1.654$$

This value of 1.65 is the standard error of the means of all possible combinations of 10 schools that could be drawn from the 97. This value is somewhat larger than the empirically determined standard error (1.22) of the 20 sample means shown in Table V.[19] It is about five times as large as the average of the 20 estimates that were based on the assumption of unrestricted selection of individual pupils (Table VI).

The mean for a sample of only 31 individual pupils drawn at random from the 7724 pupils would be as reliable as the mean for a sample consisting of 10 schools and including 500 to 1000 pupils. By the conventional formula $\left(\sigma_{\bar{x}'} = \frac{\sigma_x}{\sqrt{N-1}}\right)$

---

[19]The empirical results are subject to random fluctuation since only 20 of the almost limitless number of possible samples were considered. Although the variance of these 20 samples is smaller than would be found for a distribution of means of all samples, this is not an extremely unusual group of samples. The standard error of the standard error of the empirically determined distribution of means is $(1.22) \div \sqrt{40} = .19$. Application of the t-test shows that when the true variance of sample means is 1.65 it would be expected that a difference as large as 1.22 - 1.65 would be found a little less frequently than five times out of a hundred.

the standard error for a sample of 31 pupils would be $9.05 \div \sqrt{30} = 1.65$. A group of 10 schools thus does not provide a very reliable estimate of the population mean. The results from a sample of 10 per cent of the schools are approximately equal in precision to the results from an unrestricted sample of one-half of one per cent of the individual pupils.

The precision of the results for sampling by school groups can be increased up to any given level by taking a larger number of schools for the sample. In order to determine how much the gains in accuracy would be for larger samples, plans were made for applying the formula to samples of 20 schools, 30 schools, and 50 schools. It was also decided to use repeated sampling to get an empirical estimate of the standard error for two of these three sample sizes (20 schools and 50 schools).

Standard Errors for Samples of 20, 30 and 50 Schools. The values substituted in the formula on page 47 give the standard error of the mean for samples of 10 schools. To find the standard error for 20 schools instead of 10, it is necessary to make only one change in these values. This is a change in m, the number of schools. Using $m = 20$ in place of $m = 10$ the formula becomes

$$\sigma^2_{\overline{x}'} = \left[\frac{97 - 20}{(96)(20)}\right]\left[\frac{18,574,360.556}{(6,340.757)(97)}\right]$$

$$= (.040104)(30.1996)$$

$$= 1.2111$$

and $\sigma_{\overline{x}}' = 1.101$

It will be recalled from page 47 that the standard error for 10 schools was 1.65. By increasing the number of schools from 10 to 20 the standard error has been reduced from 1.65 to 1.10. Doubling the number of schools in the sample reduced the standard error by one-third.

The standard error of means of samples of 20 schools was determined empirically by getting a distribution of means from successively drawn samples. Twenty samples were drawn using the table of random numbers in the same manner as described for selecting samples of 10 schools. The first sample selected in this way consists of the following 20 schools: 17, 19, 27, 28, 29, 38, 39, 40, 41, 43, 45, 46, 52, 53, 62, 68, 72, 85, 92 and 94. The total number of pupils tested in these schools is 1550 and the sum of the pupil scores is 105,637. The mean score is 68.15. Summaries of similar data for the 20 different samples, each including 20 schools are presented in Table VIII.

The average of the means of the 20 samples is 66.82. The means for individual samples vary from 63.93 (for Sample No. 2) to 68.34 (for Sample No. 8), a range of 4.41 points. The means for samples 1, 8, 10 and 20 are one point or more above the average for the 20 samples. The means for samples 2, 6 and 17 are one point or more below the average. The standard deviation of the 20 means is 1.14. An unbiased estimate of the standard deviation of a distribution of means for all possible samples of 20 schools

TABLE VIII

Numbers of Pupils, Sums of Pupil Scores and Mean Scores for 20 Samples Consisting of 20 Schools Each

Sample	No. of Schools	No. of Pupils	Sum of Pupil Scores	Mean Score
1	20	1550	105,637	68.15
2	20	1311	83,812	63.93
3	20	1572	104,922	66.74
4	20	2606	173,442	66.55
5	20	1451	95,876	66.08
6	20	1289	84,467	65.53
7	20	1566	103,343	65.99
8	20	1366	93,348	68.34
9	20	1763	119,389	67.72
10	20	1767	120,589	68.24
11	20	1219	82,162	67.40
12	20	1581	106,218	67.18
13	20	1301	87,363	67.15
14	20	1487	100,000	67.25
15	20	1622	108,948	67.17
16	20	1311	86,750	66.17
17	20	1256	81,512	64.90
18	20	1971	133,275	67.62
19	20	1584	104,946	66.25
20	20	1195	81,392	68.11
Average of Sample Means				66.82
Standard Deviation of Sample Means				1.14

is found by multiplying 1.14 by $\sqrt{\frac{N}{N-1}}$, where N is the number of "cases", i.e., the number of samples in Table VIII. The product of 1.14 and $\sqrt{\frac{20}{20-1}}$ is 1.17. This empirically derived value (1.17) is somewhat larger than the one obtained from the cluster formula (1.10). If another group of samples each consisting of 20 schools were drawn, the standard deviation of the means might turn out to be either more or <u>less</u> than 1.10. A group of means secured in this way is a <u>sample</u> of the entire hypothetical population of means for groups of 20 schools drawn from the 97 schools. The true variance of this population is given by data from all schools substituted in the cluster formula. A sample of means can provide an estimate of the variance of the total population of means just as a sample of individual pupil scores may be used to get an estimate of the variance of the total population of pupil scores.

A sample of 30 schools represents roughly one-third of the 97 schools. The standard error for samples of this size is found by substituting m = 30 in the cluster formula, as follows:

$$\sigma^2_{\bar{x}}{}' = \left[\frac{97 - 30}{(96)(30)}\right]\left[\frac{18,574,360.556}{(6,340.757)(97)}\right]$$

$$= (.023264)(30.1996)$$

$$= .7026$$

$$\text{and } \sigma_{\bar{x}}{}' = .838$$

In the paragraph above, it is seen that the comparable value for samples of 20 schools is 1.10. An increase of 10 in the size of the sample thus decreased the standard error from 1.10 to .84.

For samples of 50 schools, roughly one-half the total number of sampling units (97), the standard error of the mean is

$$\sigma^2_{\bar{x}}{}' = \left[\frac{97 - 50}{(96)(50)}\right](30.1996)$$

$$= .2957$$

$$\text{and } \sigma_{\bar{x}}{}' = .544$$

An estimate of $\sigma_{\bar{x}}{}'$ for samples of 50 schools was also determined empirically by actually drawing successive samples as was done for the 10-school and 20-school samples. Groups consisting of 50 schools each were selected using the table of random numbers as described earlier. The first 50 unlike two-digit numbers from 01 through 79 and 81 through 98, as read from the table, constituted the first sample. The next 50 unlike two-digit numbers in the table constituted the second sample, and so on. Since each group of 50 schools represents slightly more than half of the total number of schools, there is a considerable amount of overlapping among the different samples. For example, school 03 was selected in samples 1, 2, 4, 6, 8, 9 and 10; school 49 was selected in samples 1, 5, 6, 7, and 10; school 70 was selected in samples 3, 7 and 9, etc. Each group of 50 schools is a <u>random</u> sample despite the fact that different groups included a number of common elements. Summaries of data for ten 50-school samples are given in Table IX.

Although the number of schools is the same (50) for each sample in Table IX, the numbers of pupils in the different samples varies from 3328 to 4546. The average number of pupils per sample is 4044 or about 52 per cent of the total pupil population. The sample means vary from 66.07 to 68.04, a range of slightly less than two points, and the standard deviation of the 10 means is .59. This value (.59) multiplied by $\sqrt{\frac{10}{9}}$ gives .62 as an unbiased estimate of $\sigma_{\bar{x}}'$. It was shown above that the "true" value of the standard error for samples of this size as calculated from the formula is .54. The estimate derived from repeated sampling is thus larger by .08 than the calculated value. Another set of 10 sample means might, of course, give an estimated standard error which is lower than .54.

TABLE IX

Numbers of Pupils, Sums of Pupil Scores and Mean Scores for 10 Samples Consisting of 50 Schools Each

Sample	No. of Schools	No. of Pupils	Sum of Pupil Scores	Mean Score
1	50	4137	281,498	68.04
2	50	4145	275,673	66.51
3	50	4419	297,859	67.40
4	50	3328	219,894	66.07
5	50	4546	300,901	66.19
6	50	4172	279,748	67.05
7	50	4087	271,224	66.36
8	50	3534	236,549	66.94
9	50	3995	267,247	66.90
10	50	4075	270,102	66.28
Average of Sample Means				66.77
Standard Deviation of Sample Means				.59

It will be interesting to determine the size of a sample of individual pupils that would give results having the same precision as a sample of 50 schools. A similar calculation was made earlier in connection with the data for samples of 10 schools. Here, as in the case of the 10-school sample, the standard deviation of individual scores for the 7724 pupils is known to be 9.05. Using $\sigma_{\overline{X}}' = .54$ as the degree of precision desired, N becomes the only unknown in the equation $\sigma_{\overline{X}}' = \frac{\sigma_X}{\sqrt{N-1}}$, or $N = \left(\frac{\sigma_X}{\sigma_{\overline{X}}'}\right)^2 + 1$. By substitution, N is found to be 282, approximately.[20] The mean score for a random sample of 282 pupils, therefore, will have the same reliability as the mean for a random sample of 50 schools which includes about 4000 pupils.

Summary of Results for Simple Random Sampling of Schools. The means for small samples of schools showed relatively large standard errors even though the numbers of pupils were quite large (500 or more). For example, the standard error for samples of 10 schools was found to be 1.65 which is more than one-sixth the magnitude of the standard deviation of individual pupil scores. The results for 20, 30 and 50-school samples, respectively, were progressively more reliable with the 50-school sample having a standard error of .54. This is about one-third the size of the error for 10 school samples and approximately one-seventeenth the magnitude of the standard deviation of the individual pupils scores. A summary of the results for each of the four sample sizes, 10, 20, 30 and 50 schools, is given in Table X. Estimated standard errors derived from repeated sampling are shown along with calculated standard errors derived from the cluster formula.

Two out of the three estimated errors are larger than their corresponding calculated values and one is smaller. It was shown earlier that both the calculated and the estimated

---

[20] Strictly speaking, the correction for a finite population should be applied to sampling by individual pupils since a comparable correction for sampling by schools is included in the cluster formula. In this instance, however, the effect of the correction would be very small.

values are four to five times as large as the standard errors that would be found by sampling individual pupils in the same numbers as were included in the respective school samples. The larger error which accompanies sampling by schools is due to the fact that there is some degree of homogeneity within each school group. If there were no group homogeneity at all, the mean scores for individual schools would differ only by chance and the 60 pupils, for example, enroled in a single school would be equivalent to a sample of 60 pupils drawn at random from the entire pupil population in the 97 schools.

TABLE X

Standard Errors of Mean for Samples of 10, 20, 30 and 50 Schools

Sample Size (No. Schools)	Average No. of Pupils Per Sample	Standard Error of Mean	
		Calculated	Estimated
10	811	1.65	1.22
20	1538	1.10	1.17
30	2400*	.84	**
50	4044	.54	.62

*Estimated average for samples of 30 schools.
**There is no estimation of $\sigma_{\overline{x}}'$ for samples of 30 schools since samples of this size were not actually drawn.

The effect of group homogeneity on the precision of sample results is shown in an "inverse" manner in Table XI. For instance, a sample of 10 schools including a total of 800 pupils has a standard error of 1.65. This degree of precision can be equaled with a random sample of 31 individual pupils. Similarly a sample of 69 pupils will have the same standard error as a sample of 20 schools including about 1600 pupils. A 30-school sample consisting of 2400 pupils and a 50-school sample including 4000 pupils have the standard errors, respectively, of samples of 117 and 282 individuals.

TABLE XI

Number of Pupils Required to Attain Given Degrees of Precision When the Sampling Unit is a School and When the Sampling Unit is an Individual Pupil

Precision of Sample Results $\sigma'_{\bar{x}}$	Number of Pupils Required: Sampling By Schools	Number of Pupils Required: Sampling By Pupils
1.65	800*	31**
1.10	1600	68
.84	2400	117
.54	4000	282

*The numbers in this column are rounded off from the averages that would be found in repeated sampling.
**Numbers in this column were computed by the method described on page 53.

The data in Table XI show that although a sample of 10 schools (800 pupils) is about as good as sample of 3 x <u>10</u> individuals (30 pupils), a sample of 50 schools (4000 pupils) is a great deal better than a sample of 3 x <u>50</u> individuals. This is due, in this instance, to the fact that that 50 schools include more than half of the total population.

For relatively large populations the reliability of sample results is primarily a function of the absolute number of units rather than the proportion of units in the sample. For example, a 20 per cent sample of 1000 schools would almost certainly give more reliable results than a 50 per cent sample of 100 schools. This would be true even though the variance of individual school means were somewhat larger for the 1000 schools than for the 100. The reason is that in the first instance a 20 per cent sample consists of 200 schools while in the second the sample size is 50 schools. The mean for 200 "cases" is far more stable than the mean for "50" cases.

In a preceding section it was stated that any given degree of precision of results could be attained using the school as the sampling unit. For instance, a sample of 60 schools out of the 97 would give a mean with a standard error approximately one-twentieth the size of the standard deviation of individual pupil scores. This would perhaps be considered a reasonably satisfactory level of reliability for a test norm.[21] But some doubts may be raised as to whether sampling is worthwhile when considerably more than half the total population must be included to get sufficiently reliable results. The main purpose of sampling is to get a reasonably accurate picture of a total population using data secured from a fraction of its members. As the required fraction becomes larger than one-half, the advantages of sampling over complete enumeration decrease rapidly. Under ordinary conditions in a school system it would seem that 50 per cent of a population is about the upper limit of the proportion to be drawn into a sample if sampling is to be a worthwhile method of securing

---

[21] E. F. Lindquist, "Factors Determining the Reliability of Test Norms," Journal of Educational Psychology, Vol. 31, (1930), p. 516.

the information desired. For this reason, the maximum size of samples dealt with in this study is limited to approximately one-half of the total population.

It may be possible to increase the accuracy of results without increasing the size of the sample by taking advantage of certain similarities, either known or suspected to exist, among the members of various sub-groups of schools. To achieve such increased accuracy it is necessary to classify or to arrange the schools into sub-groups according to some control factor prior to drawing the sample. Individual schools are then selected at random from each sub-group, in turn, rather than from the entire population of schools. This method, known as stratified sampling will next be applied in drawing samples from the 97 schools.

III. RESULTS FOR STRATIFIED SAMPLING OF SCHOOLS

Stratification by Size of Enrolment

The Purpose of Stratification. The reason for stratifying the population of schools as a preliminary step to drawing a given number of schools from it as a sample is to try to obtain results that have a smaller error than the results secured from unrestricted selection.

Stratification makes it possible to take advantage of information about the schools already available which may be related to average test scores. For example, a reasonable hypothesis is that the variance of mean scores for very large schools is somewhat less than the variance of means for very small schools. There is some evidence to support this hypothesis even though the differences in this respect between large and small schools have actually been found to be much smaller than would be supposed.[22] Since the number of pupils enroled in each school is a matter of official record, it is possible to arrange schools into relatively homogeneous sub-groups with respect to enrolment prior to drawing a sample. If, in choosing a sample, schools are selected proportionately from each sub-group there should be some gain in accuracy.

Another possibility of increased accuracy for enrolment stratification is related to the fact that a large school, by virtue of the mere number of pupils enroled, contributes much more heavily to the average for the entire pupil population than does a small school. In a given field situation it may turn out, for example, that the average test score for schools is correlated with size of enrolment (perhaps by chance in a particular instance). In such a situation the precision of sample results may be significantly increased by enrolment stratification. There would probably be a measurable increase in accuracy even though the correlation between these two variables were very small.

The mechanics of stratifying schools in a city system by size of enrolment are quite simple. This method seems to be

---

[22]E. F. Lindquist, op. cit., p. 5.

worth applying in practice merely for the purpose of guaranteeing "representation" in the sample of schools of all sizes whether or not there is any appreciable increase in accuracy of results. The results will at least "look more plausible" to those not familiar with the logic of statistical inference. It is extremely unlikely that this type of stratification would yield a larger error than unrestricted selection. With respect to this last issue of possible loss in accuracy, Snedecor points out, encouragingly, that "The penalities for the failure of stratification to be effective are usually not serious."[23]

Method Used in Stratifying Schools by Size of Enrolment. The 97 code numbers, each representing a school, were arranged in order of number of grade 8A pupils enroled. The arrangement was made by number enroled rather than by number tested in view of the fact that under normal conditions this method would be employed in practice to select a single sample of schools to be tested. There would be no "number-tested" control factor before the test was given.

With the code numbers arranged in this order School 79 is at the bottom with three grade 8A pupils enroled and School 85 is at the top with 459. It was decided more or less arbitrarily to group the schools into 10 strata with each stratum containing approximately the same number of schools. This would give 7 strata containing 10 schools each and three containing 9 each for a total of 97 schools.

[23]George W. Snedecor, op. cit., p. 850.

The 10 smallest schools were designated as Stratum I. The codes for these schools arranged in ascending order of enrolment are: 79, 20, 78, 08, 54, 34, 36, 52, 73 and 71. School 71, the largest of the 10, has an enrolment of 21 grade 8A pupils. The next 10 schools in order of enrolment constitute Stratum II. The smallest enrolment in Stratum II is 23 and the largest is 30. The remaining eight strata, i.e., III through X, contain 10, 9, 9, 9, 10, 10, 10 and 10 schools, respectively. The smallest school in Stratum X enrols 227 pupils and the largest enrols 459. The complete stratification plan is shown in Table XII where the schools and their respective enrolments are listed for each of the 10 strata.

In comparing the enrolments shown for the various strata in Table XII it will be noted that the largest school in Stratum III, i.e., School 47, has the same enrolment as the smallest school in Stratum IV. In fact, there are two schools in Stratum III and one in Stratum IV with exactly the same enrolment, namely, 34. An exception to strict stratification by enrolment was made here in order to equalize the numbers of schools in the several strata. To keep the number of schools in Stratum III limited to 10 it was necessary to put either school 04, 14 or 47 into Stratum IV. School 04 was arbitrarily selected to be placed in Stratum IV and the other two remained in Stratum III. It was necessary to use this method also in equalizing the number of schools in Strata V and VI.

TABLE XII

List of Schools Arranged in 10 Strata According to Number of Grade 8A Pupils Enroled

Stratum																				
	I		II		III		IV		V		VI		VII		VIII		IX		X	
	Sch.	Enrol.	Sch.	Enrol.	Sch.	Enrol.	Sch.	Enrol.	Sch.	Enrol.	Sch.	Enrol.	Sch.	Enrol.	Sch.	Enrol.	Sch.	Enrol.	Sch.	Enrol
	79	(3)	55	(23)	03	(31)	04	(34)	01	(39)	24	(48)	30	(54)	32	(67)	29	(112)	83	(227)
	20	(6)	50	(24)	07	(32)	75	(35)	02	(41)	39	(48)	41	(54)	64	(68)	96	(146)	89	(233)
	78	(8)	68	(24)	19	(32)	33	(36)	11	(42)	74	(49)	51	(54)	77	(70)	88	(152)	94	(240)
	08	(14)	72	(24)	56	(32)	43	(36)	40	(42)	05	(50)	15	(55)	76	(75)	21	(159)	98	(246)
	54	(16)	59	(27)	61	(32)	70	(36)	53	(43)	16	(50)	23	(55)	25	(76)	84	(159)	81	(298)
	34	(17)	44	(29)	61	(32)	38	(37)	66	(44)	17	(50)	28	(55)	35	(78)	87	(164)	86	(301)
	36	(20)	12	(30)	69	(32)	48	(38)	31	(47)	22	(52)	63	(59)	26	(81)	92	(189)	91	(309)
	52	(20)	45	(30)	57	(33)	10	(39)	42	(47)	37	(52)	62	(60)	09	(89)	95	(204)	97	(401)
	73	(20)	46	(30)	14	(34)	27	(39)	06	(48)	65	(52)	60	(65)	13	(92)	82	(211)	90	(423)
	71	(21)	49	(30)	47	(34)	..	..	..	..	..	..	18	(66)	58	(105)	93	(214)	85	(459)
Mean		14.5		27.1		32.4		36.7		43.7		50.1		57.7		80.1		171.0		313.7

The mean enrolments for the different strata range from 14.5 for Stratum I to 313.7 for Stratum X. The average school in Stratum X, therefore, contributes more than 20 times as much weight to the determination of the mean pupil score for the entire population as does the average school in Stratum I. This stratification plan is certain to reduce chance variation in the <u>weights</u> of the deviations of individual school means from the mean of a sample (or from the mean of the population). For instance, in unrestricted selection of a sample of 10 schools the chances of choosing the 10 largest schools, or the 10 smallest schools, are exactly the same as the chances of choosing any other combination of 10 schools. When the schools are stratified according to an enrolment control, there is <u>no</u> chance of drawing a disproportionate number of very small or very large schools. Whether or not this control will also reduce the size of the standard error of a sample mean remains to be seen.

The next step taken to facilitate the mechanics of drawing a sample from the stratified arrangement of schools was the assignment of special one-digit code numbers, ranging from 0 to 9, to the schools <u>within</u> each stratum. The way in which the code sheet was actually set up and used is illustrated for Strata I through V in Table XIII. The smallest school in each stratum carries the new code of "0". It makes no difference, of course, how the codes are distributed among the schools within a stratum since selections from them are to be made at random. It will be noted that Strata IV and V each contain nine schools instead of ten and consequently the codes run from 0 through 8 rather than 0 through 9.

TABLE XIII

Stratification Codes for the Schools in Five of the Ten Enrolment Strata

Stratum I		Stratum II		Stratum III		Stratum IV		Stratum V	
Strat. Code	Sch.	Strat. Code	Sch.	Strat. Code	Sch.	Strat. Code	Sch.	Strat. Code	Sch.
0	79	0	55	0	03	0	04	0	01
1	20	1	50	1	07	1	75	1	02
2	78	2	68	2	19	2	33	2	11
3	08	3	72	3	56	3	43	3	40
4	54	4	59	4	61	4	70	4	53
5	34	5	44	5	67	5	38	5	66
6	36	6	12	6	69	6	48	6	31
7	52	7	45	7	57	7	10	7	42
8	73	8	46	8	14	8	27	8	06
9	71	9	49	9	47	..	..	..	..

With this simplified coding plan it is now possible to select a school from any stratum by merely choosing a number from 0 to 9. The complete identification of a selected school requires designation of both the stratum and the new code. The usefulness of the one-digit code lies in the fact that it permits more efficient use of a table of random numbers in drawing successive samples.

<u>Method Used in Selecting a Stratified Sample of Schools</u>. In choosing a sample of 10 schools, one school was taken from each stratum. The plan of selection was to take the "starting-point number" in the table of random numbers as the one school to be chosen from Stratum I. Going down the column from the starting point, the next number would represent the one school to be chosen from Stratum II, etc. The first 10 digits encountered in the table would thus specify the sample of 10 schools, one from each stratum. Since the highest code in Strata IV, V, VI is "8", it was necessary to skip all 9's that came up in the table in the fourth, fifth or sixth positions in a series of 10. Using this method, the first ten numbers in the table were found to be 1, 2, 2, 5, 2, 5, 6, 6, 6 and 8. This specified the sample as School 1 from Stratum I, School 2 from Stratum II, School 2 from Stratum III and so on through School 8 from Stratum X. It will be seen from the stratification codes illustrated in Table XIII that "School 1, Stratum I" is the designation for School 20. Similarly, "School 2, Stratum II" is School 68. The original codes for the ten schools in this sample are 20, 68, 19, 38, 11, 17, 63, 26, 92 and 90. The respective enrolments of these ten schools are 6, 24, 32, 37, 42, 50, 59, 81, 189 and 423.

Samples consisting of more than 10 schools were selected in essentially the same way. For a 50-school sample, five schools were drawn from each of the 10 strata. The first five unlike digits in the random numbers table represented the five schools to be taken from Stratum I, the next five digits represented the schools from Stratum II and so on for the remaining eight strata.

The sampling "rate", i.e., the proportion of schools drawn, is not exactly the same for each of the ten strata. For a 10-school sample, the proportion for Strata I to III and VII to X is one school out of 10, since each of these strata consist of 10 schools. The proportion is one school out of nine for strata IV, V and VI which consist of 9 schools each. A given school in any of these latter three strata has a slightly greater chance of being selected than a given school in any of the other six. This difference in the chances of being selected is quite small, however, and in order to simplify the handling of results the data will be treated as though the sampling rate were the same for all strata.

<u>Mean Scores for Stratified Samples of 10 Schools</u>. The results for one sample of 10 schools drawn by the method described above are given in Table XIV. It will be noted that each of the 10 strata is represented by one school. The control resulting from stratification is shown by the progressive increase in size of enrolment of the individual schools from Stratum I to Stratum X. The number of pupils tested in each school is also given in the table. The enrolment factor

is about equally effective in this case in stratifying by number tested even though the proportion of pupils tested in each of the ten schools is not exactly the same.

TABLE XIV

Enrolments, Numbers Tested, and Sums of Pupil Scores for a Sample of 10 Schools Stratified by Size of Grade 8A Enrolment

Stratum	School	No. Pupils Enroled	No. Pupils Tested	Sum of Scores
I	20	6	5	362
II	68	24	24	1,531
III	19	32	32	2,305
IV	38	37	34	2,161
V	11	42	39	2,864
VI	17	50	47	3,420
VII	63	59	56	3,872
VIII	26	81	80	5,616
IX	92	189	181	10,992
X	90	433	396	27,745
Total		943	894	60,868
Mean Pupil Score				68.08

The combined enrolments of the 10 schools in Table XIV is 943 and the total number tested is 894. The sum of the 894 scores is 60,868 and the average score is 68.08. This mean (68.08) for a stratified sample of 10 schools is higher than the mean of the total pupil population (66.74) by 1.34 points.

A population of <u>results</u>, i.e., means, for 10-school stratified samples was produced by drawing 20 such samples and computing the mean for each one. Every one of the 97 schools had an equal chance to come into each sample.[24] The stratification plan guaranteed that one school and only one could be taken into a sample from any one stratum. The chances that any given school would be selected on a particular drawing were therefore one in 10 (the number of schools in a stratum) rather than one in 97 as in the case of unrestricted selection. Summaries of results for each of the 20 stratified samples are given in Table XV. The data for Sample 1 in this table were shown above in full detail in Table XIV.

The smallest of the samples in Table XV, in terms of number of pupils, is Sample 9 with 673 tested. The largest is Sample 14 with 943 tested. As would be expected, the variation in number of pupils per sample is much less here than it was for the 10-school samples drawn by unrestricted selection. In the latter case, the smallest and the largest of 20 samples were found to include 412 and 1362 pupils, respectively.[25]

Among the 20 stratified samples, Sample 18 shows the lowest average score (65.53) which is about one and one-fourth points below the population mean of 66.74. Sample 14 has the highest average (68.53), approximately one and three-fourths

---

[24] This statement is not strictly accurate because of the exception concerning the three 9-school strata as pointed out on page 66.

[25] Table V, page 37.

points above the population mean. The difference between the highest and lowest sample means, i.e., the width of the "error band", is therefore about three points. The comparable error range for samples of 10 schools chosen by unrestricted selection was earlier shown to be approximately four and one-half points.[26]

TABLE XV

Numbers of Pupils, Sums of Scores and Mean Scores for Samples of Schools Stratified by Size of Grade 8A Enrolment

Sample	No. of Schools	No. Pupils Enroled	No. Pupils Tested	Per Cent Tested	Sum of Scores	Mean Score
1	10	943	894	94.8	60,868	68.08
2	10	734	700	95.4	45,917	65.60
3	10	871	811	93.1	54,560	67.28
4	10	779	733	94.1	49,235	67.17
5	10	993	942	94.9	62,978	66.86
6	10	957	908	94.9	60,880	67.05
7	10	815	754	92.5	49,829	66.09
8	10	765	728	95.2	48,282	66.32
9	10	705	673	95.5	44,609	66.28
10	10	775	733	94.6	48,956	66.79
11	10	858	811	94.5	54,351	67.02
12	10	986	939	95.2	63,810	67.96
13	10	730	698	95.6	46,820	67.02
14	10	983	943	95.9	64,628	68.53
15	10	966	902	93.4	61,274	67.93
16	10	789	741	93.9	49,682	67.05
17	10	779	753	96.7	50,610	67.21
18	10	769	723	94.0	47,379	65.53
19	10	749	708	94.5	46,457	65.62
20	10	851	793	94.3	52,393	66.07
Average of Sample Means						66.88
Standard Deviation of Sample Means						.83

The average of the 20 means in Table XV (66.88) is higher than the population mean by .08. The standard

26 Ibid.

deviation of the 20 means is .83, and the unbiased estimate of $\sigma_{\bar{x}}'$ is .86.[27] The corresponding value (computed by formula) for unrestricted samples of the same size was found to be 1.65. Stratification of schools by enrolment, therefore, appears to have brought about a substantial reduction in size of the standard error of the mean.

Mean Scores for Stratified Samples of 50 Schools. A sample consisting of 50 schools was selected from the stratification code sheet, Table XIII, by the method described above. Five schools were drawn from Stratum I, five from Stratum II and so on through Stratum X. The sample included one-half of the schools in each of the 10-school strata and five-ninths of the schools in the three strata which contain only nine schools each. The combined sums of pupil scores for the 50 schools was found to be 253,399 and the mean for the sample is 66.86. This is .12 above the population average, 66.74. Table XVI gives the results for 10 such samples of schools, stratified by size of grade 8A enrolment, each sample consisting of 50 schools.

A striking feature of the data in Table XVI is the small variation in size of the ten sample means. Sample 9 has the highest mean, 67.21, and Sample 8 has the lowest, 66.28. The difference between the highest and the lowest is less than one point. Five out of the ten means differ from the population average, 66.74, by less than .20. The average of the

---

[27] Computed as follows: $.83 \sqrt{\frac{20}{19}} = .86$

sample means is 66.75 and the standard deviation is .27. The estimate of $\sigma_{\bar{x}}$, is .28. The computed error for unrestricted samples of the same size (50 schools) derived from the formula was earlier shown to be .54. Enrolment stratification seems to have reduced the sampling error very markedly. It should be borne in mind, however, that the estimated error for this plan of stratification is based on a small number of "cases"--only ten means.

TABLE XVI

Numbers of Pupils, Sums of Scores and Mean Scores for Samples of 50 Schools Stratified by Size of Grade 8A Enrolment

Sample	No. of Schools	No. Pupils Tested	Sum of Scores	Mean Score
1	50	3790	253,399	66.86
2	50	3845	257,691	67.02
3	50	3893	259,071	66.55
4	50	3972	265,144	66.75
5	50	3935	263,203	66.89
6	50	3886	258,902	66.62
7	50	3761	249,886	66.44
8	50	3853	255,366	66.28
9	50	3837	257,899	67.21
10	50	3907	261,281	66.88
Average of Sample Means				66.75
Standard Deviation of Sample Means				.27

There is an efficient alternative to further repeated sampling as a means of determining with a high degree of reliability the standard error of the mean for stratified samples. The cluster formula may be applied to the data for <u>all</u> of the schools in each of the ten strata separately just as it was applied earlier to all 97 schools. Appropriate procedures for combining the variances of means for the ten sub-samples from the respective strata will give the variance of the mean for the entire sample.

Application of the Cluster Formula to Stratified Samples of Schools. Certain restrictions are imposed on the process of selection used in choosing a stratified sample. Nevertheless, the theory of uniform probability still applies so long as the conditions of random sampling are met in the drawing of each individual school within a given stratum. For example, Enrolment Stratum I may be thought of as representing a defined "population" of schools from which a sample is to be drawn. The variance of the means for samples taken from this "stratum population" may then be determined by applying the cluster formula to the data for the 10 schools in Stratum I in exactly the same way it was applied to the 97 schools. The standard error for samples of schools chosen from Stratum I is thus given by

$$\sigma^{\prime 2}_{\bar{x}_I} = \frac{M_I - m_I}{(M_I - 1)\, m_I} \cdot \frac{\sum_{j=1}^{M_I} \left[N_{Ij} (\bar{x}_{Ij} - \bar{x}_I) \right]^2}{\bar{N}_I^2 M_I}$$

where $\sigma^{\prime}_{\bar{x}_I}$ = the standard error of the mean pupil score for samples of schools drawn from Stratum I

M_I = the number of schools in Stratum I

m_I = the number of schools in the sample chosen from Stratum I

N_{Ij} = the number of pupils tested in a given school in Stratum I

$\bar{x}_{Ij}$ = the mean score for a given school in Stratum I

$\bar{N}_I$ = the average number of pupils tested per school in Stratum I

This formula is identical with the one used to determine the error for random selection from the 97 schools except for the subscript notation. Here the subscripts all refer to the schools which constitute Stratum I only. The number of schools in this stratum (M_I) is 10. The mean pupil score for the 10 schools ($\bar{x}_I$) is 64.80 and the average number of pupils tested per school ($\bar{N}_I$) is 13.7. The value represented by the expression $N_{Ij}(\bar{x}_{Ij} - \bar{x}_I)$ was computed for each of the 10 schools using the method described on page 44. The sum of these weighted deviations squared is 39,328.252. $\bar{N}_I^2$ is 187.69, and the equation for the standard error of the mean pupil score for samples of schools drawn from Stratum I may now be written with only one unknown in the right-hand member.

$$\sigma^2_{\bar{x}_I}{}' = \frac{10 - m_I}{(10 - 1)m_I} \; \frac{39,328.252}{(187.69)(10)}$$

$$\text{or } \sigma^2_{\bar{x}_I}{}' = \frac{10 - m_I}{9\ m_I} \quad (20.954)$$

Similar computations were made for the other nine strata. The right hand members of the equations for Strata I through X are as follows:

$$\text{Stratum I} \ldots \frac{10 - m_I}{9\ m_I} \quad (20.954)$$

$$\text{II} \ldots \frac{10 - m_{II}}{9\ m_{II}} \quad (11.195)$$

$$\text{III} \ldots \frac{10 - m_{III}}{9\, m_{III}} \qquad (19.680)$$

$$\text{IV} \ldots \frac{9 - m_{IV}}{8\, m_{IV}} \qquad (10.760)$$

$$\text{V} \ldots \frac{9 - m_{V}}{8\, m_{V}} \qquad (15.489)$$

$$\text{VI} \ldots \frac{9 - m_{VI}}{8\, m_{VI}} \qquad (28.450)$$

$$\text{VII} \ldots \frac{10 - m_{VII}}{9\, m_{VII}} \qquad (4.206)$$

$$\text{VIII} \ldots \frac{10 - m_{VIII}}{9\, m_{VIII}} \qquad (8.511)$$

$$\text{IX} \ldots \frac{10 - m_{IX}}{9\, m_{IX}} \qquad (14.892)$$

$$\text{X} \ldots \frac{10 - m_{X}}{9\, m_{X}} \qquad (6.525)$$

Assuming the same number of schools to be drawn from each stratum, the magnitude of $\frac{M - m}{(M - 1)m}$ is identical for seven of the strata. For Strata IV, V and VI containing 9 schools each, the value represented by this expression is only slightly smaller than for the other seven strata. Consequently, the relative sizes of the errors for the different strata may be determined directly by comparing the sizes of the quantities shown in parentheses. A sample drawn from Stratum I, for instance, which shows the quantity, "20.954", in parentheses will have a larger error than a sample from Stratum X with the quantity, "6.525", in parentheses. The error will be smallest for samples drawn from Stratum VII and largest for samples from Stratum VI, assuming the same sample size for each stratum.

To derive the error for an entire sample consisting of 10 sub-samples, one sub-sample from each stratum, it is necessary to combine the separate errors for the different strata. It is evident without any detailed analysis that a given number of schools taken from Stratum I, where the average number of pupils tested per school is 13.7, will not influence the error for the entire sample as much as the same number of schools drawn from Stratum X where the average number tested per school is 294.2. In the case of the latter stratum, a deviation of an individual school mean from the mean of a sample carries more than 20 times as much weight as a similar deviation for a school drawn from Stratum I. Therefore, in combining the results for the entire sample the error for a given stratum must be given a weight which is based on the number of <u>pupils</u> rather than the number of schools in that stratum.

The total number of pupils tested in the 10 schools in Stratum I is 137 and the total number tested in all strata is 7724. The weight to be assigned to the variance of the means for samples drawn from Stratum I is $\frac{137}{7724}$. This is represented by the expression[28]

$$\left(\frac{N_i}{\sum_{I}^{X} N_i}\right)^2$$

where N_i = the total number of pupils tested in a given stratum

$\sum_{I}^{X} N_i$ = the sum of the numbers of pupils tested in all 10 strata

That fraction of the error for the entire sample contributed by a sub-sample of schools from Stratum I is thus given by

$$\left(\frac{137}{7724}\right)^2 \left(\frac{10 - m_I}{9\ m_I}\right) (20.954)$$

$$= .000313 \left(\frac{10 - m_I}{9\ m_I}\right) (20.954)$$

$$= .0065 \left(\frac{10 - m_I}{9\ m_I}\right)$$

The number of pupils tested in the 10 schools in Stratum X is 2942. The weighted fraction of error contributed by a sub-sample from this stratum is therefore

$$\left(\frac{2942}{7724}\right)^2 \left(\frac{10 - m_X}{9\ m_X}\right) (6.525)$$

$$= .9466 \left(\frac{10 - m_X}{9\ m_X}\right)$$

---

[28] George W. Snedecor, Statistical Methods, pp. 461-466. Ames, Iowa: The Collegiate Press, Inc., 1946.

which is seen to be more than 100 times as great as the similar value for Stratum I, even though the error of the sub-sample itself is considerably smaller for Stratum X than for Stratum I (as shown by comparing 6.525 with 20.954).

The <u>sum</u> of comparably weighted squared errors for all strata represents the square of the standard error of the mean pupils score for the <u>entire sample</u>. Computations for each stratum give the following results which may be used with any desired value substituted for m, the number of schools drawn from a given stratum:

$$\text{Stratum I} \ldots \quad .0065 \left(\frac{10 - m_I}{9\, m_I}\right)$$

$$\text{II} \ldots \quad .0114 \left(\frac{10 - m_{II}}{9\, m_{II}}\right)$$

$$\text{III} \ldots \quad .0311 \left(\frac{10 - m_{III}}{9\, m_{III}}\right)$$

$$\text{IV} \ldots \quad .0172 \left(\frac{9 - m_{IV}}{8\, m_{IV}}\right)$$

$$\text{V} \ldots \quad .0367 \left(\frac{9 - m_V}{8\, m_V}\right)$$

$$\text{VI} \ldots \quad .0893 \left(\frac{9 - m_{VI}}{8\, m_{VI}}\right)$$

$$\text{VII} \ldots \quad .0219 \left(\frac{10 - m_{VII}}{9\ m_{VII}}\right)$$

$$\text{VIII} \ldots \quad .0887 \left(\frac{10 - m_{VIII}}{9\ m_{VIII}}\right)$$

$$\text{IX} \ldots \quad .6608 \left(\frac{10 - m_{IX}}{9\ m_{IX}}\right)$$

$$\text{X} \ldots \quad .9466 \left(\frac{10 - m_{X}}{9\ m_{X}}\right)$$

When a sample of 10 schools is chosen from the 97, consisting of one from each stratum, m is 1 and the expressions $\frac{10 - m}{9\ m}$ and $\frac{9 - m}{8\ m}$ each become 1, i.e., $\frac{10 - 1}{(9)(1)} = 1$. The standard error of the mean for the entire sample is, therefore,

$$\sigma^2_{\overline{x}'} = .0066 + .0114 + .0311 + .0172 + .0367 + .0893 + .0219 + .0887 + .6608 + .9466$$

$$= 1.9103$$

$$\sigma_{\overline{x}'} = 1.382, \text{ the standard error of the mean pupil score.}$$

This computed value (1.38) for $\sigma_{\overline{x}}'$ for a 10-school sample, stratified by grade 8A enrolment, is considerably larger than the estimated value[29] (.86) based on repeated drawing of

---

[29] The 20 sample means used in deriving the estimated $\sigma_{\overline{x}}'$ are given in Table XV, page 67.

stratified samples of this size. The test employed to determine the level of significance of the difference between the two variances is described in the succeeding paragraph.

It is known that the distribution of sample values of σ^2 may be put in the form of[30]

$$\chi^2 = \frac{N\sigma^2}{\sigma_P^2}$$

where the distribution of the population is assumed to be normal. Using this expression in conjunction with a table of χ^2 and using d.f. = N - 1, the sampling variation of σ^2 may be determined--provided the population value of σ^2 (i.e. σ_P^2) is known. Taking data from the paragraph above, N = 20, $\sigma^2 = (.86)^2$, and $\sigma_P^2 = (1.38)^2$. By substitution, $\chi^2 = 7.8$. A χ^2 of this magnitude, with d.f. = 19, is significant at the five per cent level but not at the one per cent level. The results of the significance test thus indicate that the group of 20 means secured by repeated sampling is somewhat unusual. A very large number of means secured in the same way would probably have a variance much closer to the computed value for stratified samples of 10 schools than this particular sample of 20.

As would be expected, the computed error for a 10-school stratified sample is smaller than the similarly computed error

---

[30] F. E. Croxton and D. J. Cowden, Applied General Statistics, p. 340. New York: Prentice Hall, Inc., 1939.

given in a preceding section of this chapter for a 10-school unrestricted sample--1.38 as compared with 1.65. The difference between the two errors is not large but it is certainly of sufficient magnitude to make stratification worthwhile.

The standard errors for stratified samples of 30 schools and 50 schools, respectively, were determined from the formula. In the case of a 30-school sample with three schools to be taken from every stratum, 3 is substituted for m_I, for m_{II} ...m_X in the formulae applying to the different strata as presented on pages 71-72. For Stratum I, the substitution is $\left[\frac{10 - 3}{(9)(3)}\right]$ (20.954). This quantity multiplied by the weighting factor for Stratum I gives

$$\left(\frac{137}{7724}\right)^2 \left[\frac{10 - 3}{(9)(3)}\right] (20.954) = .0017$$

as the fraction of the error squared for the entire sample contributed by three schools drawn from Stratum I. By summing the results of similar computations for all ten strata, the standard error for a 30-school stratified sample is found to be

$$\sigma^2_{\bar{x}}{}' = .0017 + .0030 + .0081 + .0043 + .0092$$
$$+ .0223 + .0057 + .0230 + .1714 + .2454$$
$$= .4941$$

and $\sigma_{\bar{x}}{}' = .703$

The comparable value for an unrestricted sample of 30 schools was shown on page 50 to be .84. Stratification has thus reduced the error for samples of this size by .14.

Computations were made for a 50-school stratified sample, using the formula in the same way it was applied above to the 30-school sample. Since for this sample size five schools are to be drawn from each stratum, the number "5" is substituted for m in the formula for each stratum. The respective weighting factors, of course, remain the same for the different strata regardless of the size of the sample drawn. With five schools taken from Stratum I, the computation for this stratum is

$$\left(\frac{137}{7724}\right)^2 \left[\frac{10 - 5}{(9)(5)}\right] (20.954) = .00073$$

The sum of the results of similar computations over all strata is as follows:

$$\sigma^2_{\bar{x}} = .0007 + .0013 + .0035 + .0017 + .0037 + .0089 + .0024 + .0099 + .0734 + .1052 = .2107$$

$$\text{and } \sigma_{\bar{x}}' = .459$$

For an <u>unrestricted sample</u> of 50 schools, the standard error of the mean computed from the cluster formula was found to be .54 as compared with .46 shown here for a stratified sample. In the case of a sample of this size, therefore, stratification has reduced the error by about .08. The proportionate reduction in error brought about by enrolment stratification is the same for all three sample sizes--10 schools, 30 schools and 50 schools. In each instance the error for the stratified sample is about eight-tenths the size of the error for the unrestricted sample.

The determination of an estimated error for a stratified sample of 50 schools was described in a preceding section of this chapter where data were presented showing the means of 10 successively drawn samples. The estimated error derived from this actual distribution of 10 means was found to be .28--considerably smaller than .46, the value obtained above from the formula. The significance test (X^2) described on page 79 was applied to these two variances. By substitution, X^2 is found to be 3.7. With d.f. = 9, a X^2 of this size is not significant at the five per cent level.

Estimation of Error by the Use of Data from a Single Sample of Schools. Up to this point, all of the generalizations made regarding various statistics that may be secured from samples of schools have been established by a line of reasoning that led from the population to the sample. The problem, essentially, has been one of determining the degree of similarity between the known characteristics of the total population and the observed characteristics found in sampled portions of that population. The method of repeated sampling, using various sample sizes and different sampling designs, has been employed to produce distributions of sample means that could be compared with the mean of the population. Also, in order to make inter-comparisons between the results for alternative sampling plans, data for the total population have been used in appropriate formulae. Generalizations made concerning the limits of accuracy within which a sample of a given size and selected by a given method will represent this population have been partially verified experimentally.

Both for repeated sampling and for application of the cluster formula the desired information concerning the entire "universe of schools" under consideration has been at hand to be used in experimentation and to be manipulated mathematically. Such a situation, of course, does not exist under practical conditions where the problem would be to select a single sample of schools and from it draw inferences concerning the larger population whose characteristics are unknown. If determination of the sampling error is to be undertaken at all in most practical situations, it must be estimated from the data secured from the sample itself. The formula which gives an estimated standard error of the mean pupil score for a single sample of schools differs only slightly from the one used above which required complete data from all schools. The essential difference between the two will be seen to consist of a modification in the denominator of the estimate formula which has the effect of correcting for bias due to the fact that variances of samples are systematically smaller than the variance of the population from which they are drawn. The formula[31] for estimating $\sigma_{\bar{X}}'$ is given below, with the

---

[31] The complete derivation of this formula is given by Eli S. Marks in "Sampling in the Revision of the Stanford-Binet Scale", Psychological Bulletin, Vol. 44 (1947), pp. 429-434.

notation having exactly the same meaning as in the one used earlier, except for the $\overline{x}'$ and $\overline{N}'$ which now represent, respectively, the average score for the sample, and the average number of pupils per school in the sample.

$$s_{\overline{x}'}^2 = \frac{M - m}{Mm} \; \frac{\sum_{i=1}^{m} \left[N_i (\overline{x}_i - \overline{x}') \right]^2}{(m - 1)(\overline{N}')^2}$$

The M here still represents the total number of clusters (schools) in the entire population. If M is not known exactly, but is very large in comparison with m, an approximation may be substituted without any appreciable loss in accuracy of results.[32]

In obtaining an estimate of the standard error for a sample that is stratified, the weighting factor to be used for each stratum is derived in the same way as shown earlier. This factor to be applied to $s_{\overline{x}'}^2$, for each stratum is represented by[33]

$$\left(\frac{N_i}{\sum_{i=1}^{R} N_i} \right)^2$$

[32] Eli S. Marks, op. cit., p. 421.

[33] To be rigorously accurate, it is necessary to use an additional component in the formula because the weighting factor, itself, represents an estimate. Where this estimate is reasonably good, the effect of the correction is trivial.

where N_i = the number of pupils in a given stratum in the sample

$\sum_{i=1}^{R} N_i$ = the total number of pupils in the entire sample, i.e., the sum of pupils drawn into the sample from all (R) strata

The generalized formula for deriving an estimate of the standard error of the mean pupil score from the data of a single stratified sample of schools drawn from R strata is therefore

$$s_{\bar{x}}^{2\prime} = \sum_{i=1}^{R} \left[\left(\frac{N_i}{\sum_{i=1}^{R} N_i} \right)^2 \left(\frac{M_i - m_i}{M_i \, m_i} \right) \frac{\sum_{i=1}^{m_i} \left[N_{ij} (\bar{x}_{ij} - \bar{x}_i') \right]^2}{(m - 1)(\bar{N}'_i)^2} \right]$$

where the i subscript refers to a particular stratum and the j subscript refers to a particular school.

The use of the formula will be illustrated by substituting in it the actual data secured by drawing a stratified sample of 50 schools. The sample to be used is the one summarized as "Sample 1" in Table XVI on page 71. As pointed out in connection with Table XVI, five schools selected from each of the ten strata constituted the 50 schools of "Sample 1". Schools 79, 54, 34, 36 and 73 were the five chosen from Stratum I. The data for this "stratum sample" may be arranged as follows:

	School	Number Tested	Sum of Scores
Sample from Stratum I	79	3	208
	54	16	1102
	34	16	980
	36	19	1021
	73	19	1275
		71	4586

The mean score, $\overline{x}'$, for this stratum (4586 ÷ 71) is 64.5915. For School 79, the weighted deviation of the school mean from the mean of the stratum, i.e., $(\Sigma x_{ijk} - N_{ij}\overline{x}_i')$ or $[208-(3)(64.5915)]$ is 14.2255. The comparable values for the remaining four sample schools from this stratum are, respectively, 68.5360, - 53.4640, - 77.0555 and 47.7615. The sum of the squares of these five values,

$$\sum_{j=1}^{m_i} \left[N_{ij}(\overline{x}_{ij} - \overline{x}_i) \right]^2, \text{ is } 15{,}976.658.$$

The average number of pupils tested per school, $\overline{N}_i'$, (i.e., 71 ÷ 5) is 14.2000 and $(\overline{N}')^2$ is 201.640. N_i, the total number tested in the five schools is, of course, 71. The total number tested in the entire sample of 50 schools is 3790 and therefore

$$\frac{N_i}{\sum_{i=1}^{R} N_i} = \frac{71}{3790}$$

Since M_i, the number of schools in Stratum I, for the entire population is known to be 10, there is no need for approximation in this instance. All the quantities required for substitution in the formula have now been obtained <u>for Stratum I</u>. The substitution is as follows:

$$\left(\frac{71}{3790}\right)^2\left[\frac{10 - 5}{(10)(5)}\right]\left[\frac{15,976.658}{(4)(201.640)}\right]$$

$$= (.000351)(.10000)(19.808)$$

$$= .00070$$

This value (.00070) is to be added to the nine comparable values which will be obtained by the same method from the other strata to give the estimated variance of the mean for the entire sample.

The five schools chosen from Stratum II in the sample of 50 were schools 50, 68, 72, 59 and 12. Data from these schools substituted in the formula in the same way as shown above for Stratum I give the following results for Stratum II:

$$\left(\frac{111}{3790}\right)^2\left[\frac{10 - 5}{(10)(5)}\right]\left[\frac{9,365.888}{(4)(201.640)}\right] = .00041$$

Similar results for the respective groups of five schools selected from the remaining eight strata are:

Stratum III $\left(\frac{156}{3790}\right)^2\left[\frac{10 - 5}{(10)(5)}\right]\left[\frac{60,047.872}{(4)(973.440)}\right] = .00261$

IV $\left(\frac{168}{3790}\right)^2\left[\frac{9 - 5}{(9)(5)}\right]\left[\frac{76,501.661}{(4)(1,128.960)}\right] = .00296$

V $\left(\frac{197}{3790}\right)^2\left[\frac{9 - 5}{(9)(5)}\right]\left[\frac{201,957.705}{(4)(1,552.360)}\right] = .00781$

VI $\left(\frac{239}{3790}\right)^2\left[\frac{9 - 5}{(9)(5)}\right]\left[\frac{429,119.535}{(4)(2,284.840)}\right] = .01662$

$$\text{VII} \quad \left(\frac{284}{3790}\right)^2 \left[\frac{10-5}{(10)(5)}\right] \left[\frac{46,826.502}{(4)(3,226.240)}\right] = .00204$$

$$\text{VIII} \quad \left(\frac{399}{3790}\right)^2 \left[\frac{10-5}{(10)(5)}\right] \left[\frac{349,420.195}{(4)(6,368.040}\right] = .01520$$

$$\text{IX} \quad \left(\frac{755}{3790}\right)^2 \left[\frac{10-5}{(10)(5)}\right] \left[\frac{2,872,327.612}{(4)(22,801.000)}\right] = .12496$$

$$\text{X} \quad \left(\frac{1410}{3790}\right)^2 \left[\frac{10-5}{(10)(5)}\right] \left[\frac{1,003,725.343}{(4)(79,524.000)}\right] = .04367$$

The sum of these results over all strata is

$$s^{2}_{\bar{x}}{}' = .00070 + .00041 + .00261 + .00296 + .00781 + .01662 + .00204 + .01520 + .12496 + .04367 = .21698$$

and $s_{\bar{x}}' = .4658$

The standard error of the mean pupil score for a stratified sample of 50 schools as <u>estimated</u> from the data of this <u>one</u> sample is therefore .466. The "true" standard error for stratified samples of this size, computed from the data for <u>all</u> schools was shown on page 81 to be .459. The estimate derived here from the data of the sample is thus .007 higher than the "true" value.

A second estimated error was computed using the data from a different sample of 50 schools--namely, "Sample 2" which is summarized in Table XVI. Data for the five schools in each stratum sub-sample were substituted in the formula as illustrated in the computations shown for "Sample 1" above with the following results:

$$s^{2}_{\bar{x}}{}' = .00115 + .00188 + .00113 + .00157 + .00375 + .01678 + .00320 + .01991 + .05325 + .09927$$

$$= .20189$$

and $s_{\bar{x}}' = .4493$

This second estimate of the error (.449) is somewhat lower than the first one (.466). Neither differs from the computed "true" value by more than .010.

A problem commonly encountered in survey testing in a city school system is to compare the mean score obtained in a given year with the mean secured from a similar survey conducted in some previous year. Let us assume for purposes of illustration that two such surveys to test reading skills of grade 8A pupils were conducted in Detroit--the first in 1940 and the second in 1945--and that each of the two survey plans involved the sampling of <u>schools</u> that enrol grade 8A pupils. To give further details of the illustration, in 1940 a stratified sample of 50 schools was selected and the pupils in these schools were tested. The results of the "1940 Survey" are represented by the data for Sample 1, above. The survey was repeated in 1945 when the same test was again administered to grade 8A pupils in a stratified sample of 50 schools. Results for the "1945 Survey" are represented by the data for Sample 2, above. The average in 1945 (Sample 2) was 67.0198.[34]

---

[34] Table XVI, p. 71.

The average is thus slightly higher for 1945 than for 1940. This difference does not indicate necessarily that the 1945 grade 8A pupils in Detroit read better than did the 1940 grade 8A pupils. The estimated standard error of the 1940 mean as determined by the formula for stratified samples of schools was found to be .466, and the estimated error for the 1945 mean is .449. Using the values of $s^2_{\bar{x}}{}'$ already computed above, the standard error of the difference[35] between the two means is

$$\sigma_{(\bar{x}_1' - \bar{x}_2')} = \sqrt{.20189 + .21689}$$
$$= .6471$$

The difference between the two means divided by the standard error of the difference is

$$\frac{67.0198 - 66.8599}{.6471} = \frac{.1599}{.6471}$$
$$= .247$$

This ratio (.247) clearly shows that there is practically no support for the conclusion that the average reading achievement for all grade 8A pupils in Detroit is higher in 1945 than it was in 1940. In order to be significant at the five per cent level, for example, the difference between the means of the two samples would have to be approximately eight times as large as the obtained difference.

---

[35] $\sigma_{(\bar{x}_1' - \bar{x}_2')} = \sqrt{\sigma^2_{\bar{x}_1'} + \sigma^2_{\bar{x}_2'}}$

The solution of this hypothetical problem illustrates the need for getting objective estimates of error from the data of the sample itself. It might also be emphasized that it was possible to make valid estimates for each of the samples used here because of the controls exercised in actually drawing the two samples.

Summary of Results for Sampling by Enrolment Strata. A sample of schools stratified by enrolment gives a better estimate of the mean pupil score for the 97 schools than does a sample of schools chosen by unrestricted selection. The errors for stratified samples of 10, 30 and 50 schools, are, respectively, 1.38, .70 and .46. For unrestricted samples of the same sizes the errors are 1.65, .84 and .54. In the case of a 10-school sample, stratification reduces the error by .27. For samples of 30 and 50 schools the reductions in error are .14 and .08. Although the absolute magnitude of the decrease is seen to be greater for a small sample than for a large sample, the proportionate decrease is the same for all sample sizes. The standard error of the mean for a stratified sample of a given size is approximately eight-tenths as large as the error for an unrestricted sample containing the same number of schools.

Stratification by size of enrolment represents, of course, only one of the many possible ways of arranging a population of schools into sub-groups in preparation for drawing a sample. The next section of this chapter deals with the results secured by using a different type of stratification control.

IV. RESULTS FOR STRATIFIED SAMPLING OF SCHOOLS

Stratification by Geographic Location

It is a well known fact that the average level of pupil achievement on subject matter tests is to some extent related to the level of economic status of the neighborhood in which the pupils live. This relationship is especially reflected in the schools of an industrial city where one school may be located in a slum area and another in a residential area of the economically privileged. Pupils who come from families that are moderately well to do in an economic sense show, on the average, systematically greater skill in conventional school work, i.e., learning from books, than do pupils from poor economic backgrounds. It is not pertinent to the present investigation to inquire into the complex relationships between the economic status of a family and the child's motivations, interests, purposes, aptitudes, and achievements in the tasks required of him in school. It is enough to know, or even to suspect, that the average test scores for schools located in certain under-privileged residential areas of the city will tend to be lower than the average scores for schools in the privileged areas. With this idea in mind, plans were made to stratify the 97 schools by geographic location, having the different strata represent, at least roughly, differences in economic levels of school neighborhoods. If there is some degree of homogeneity of school means within the various strata, the error of the mean of a sample drawn proportionately from each stratum will be less than could be obtained by unrestricted selection.

Method Used in Stratifying Schools by Geographic Location. Each of the 97 schools was spot-located on an outline map of the City of Detroit. A search was then made for sources of detailed information relating to economic status of residential areas surrounding each school. It was anticipated that reports of the federal census of 1940 might provide pertinent information in usable form, or that some of the recent reports published by a local governmental research agency[36] might be helpful in differentiating between various neighborhoods with respect to economic status. Neither of these two potential sources turned out to be usable. To be strictly objective in the matter, the economic level of a residential area might be defined either as average income per person residing in the area, or as average valuation per person of occupied dwellings in the area. The information needed in order to classify school neighborhoods according to either of these definitions is not readily available.

Next, an attempt was made, using the map with the 97 schools spotted, to draw lines that would correspond with certain "official" geographic boundaries within the city and at the same time enclose sub-groups of schools roughly representing various residential areas that are known to differ in

---

[36] Detroit Bureau of Governmental Research, Detroit, Michigan.

general economic level. For example, internal boundary lines used in the federal census of 1940 divide the city of Detroit into 15 sub-areas.[37] These lines were drawn on the school map with the idea of using the schools located in one sub-area, or perhaps two adjacent sub-areas as a geographic stratum. This plan was not feasible because of the wide differences in numbers of schools found in the respective census areas. For example, 12 schools are located in census area "Number 14" while there is only one school in area "Number 8". It was impossible to combine census areas in such a way as even to approximate a reasonable distribution of schools among different tentatively defined strata.

A second attempt to define geographic strata by using "official" boundary lines was also unsuccessful. The 208 elementary schools in Detroit are divided into eight geographic groups for administrative purposes. Each of the eight "districts" includes approximately the same total number of schools but not the same number of schools enroling grade 8A pupils. It was finally decided, in view of the very small number of schools that would be contained in each of several strata, to divide the 97 schools arbitrarily into 10 geographic subgroups of approximately equal size as was done in the case of the enrolment stratification. Proportionate sampling of the different strata could then be carried out by taking the same number of schools from each stratum.

---

[37]Bureau of the Census. Housing: Analytical Maps, Detroit, Michigan, Block Statistics, 16th Census of the United States, 1940, p. 3. Washington, D.C.: United States Department of Commerce, Bureau of the Census.

The first geographic stratum to be thus defined is a group of 10 schools in the northwest corner of the city. Stratum II is a group of 10 schools, directly east of Stratum I and bordering the city limits on the north. Stratum III is another group of 10 schools located in the northeast corner of the city. Seven additional strata were blocked out on the map by merely drawing a line around a group of schools adjacent to a groups already designated as constituting a stratum. An outline map appears on page 96 showing the ten strata and the location of the schools in each one.

Special one-digit codes were assigned to the respective schools within each stratum as was done for the enrolment strata. Table XVII gives the schools with their "geographic codes", arranged in groups according to the ten geographic areas shown on the map. By referring from the table (page 97) to the map it will be seen that the schools in Stratum I, namely, Schools 79, 56, 39, 37, 20, 17, 15, 13, 10 and 08 are located in the northwest corner of the city. The other groups of schools may be located geographically by making similar references from Table XVII to the map.

It is obvious that a purely arbitrary method has been used here in making distinctions between group of schools where the original purpose was to group together those schools whose surrounding neighborhoods show some degree of similarity. Nevertheless, the purpose was probably achieved in some measure by virtue of the fact that adjacent neighborhoods would

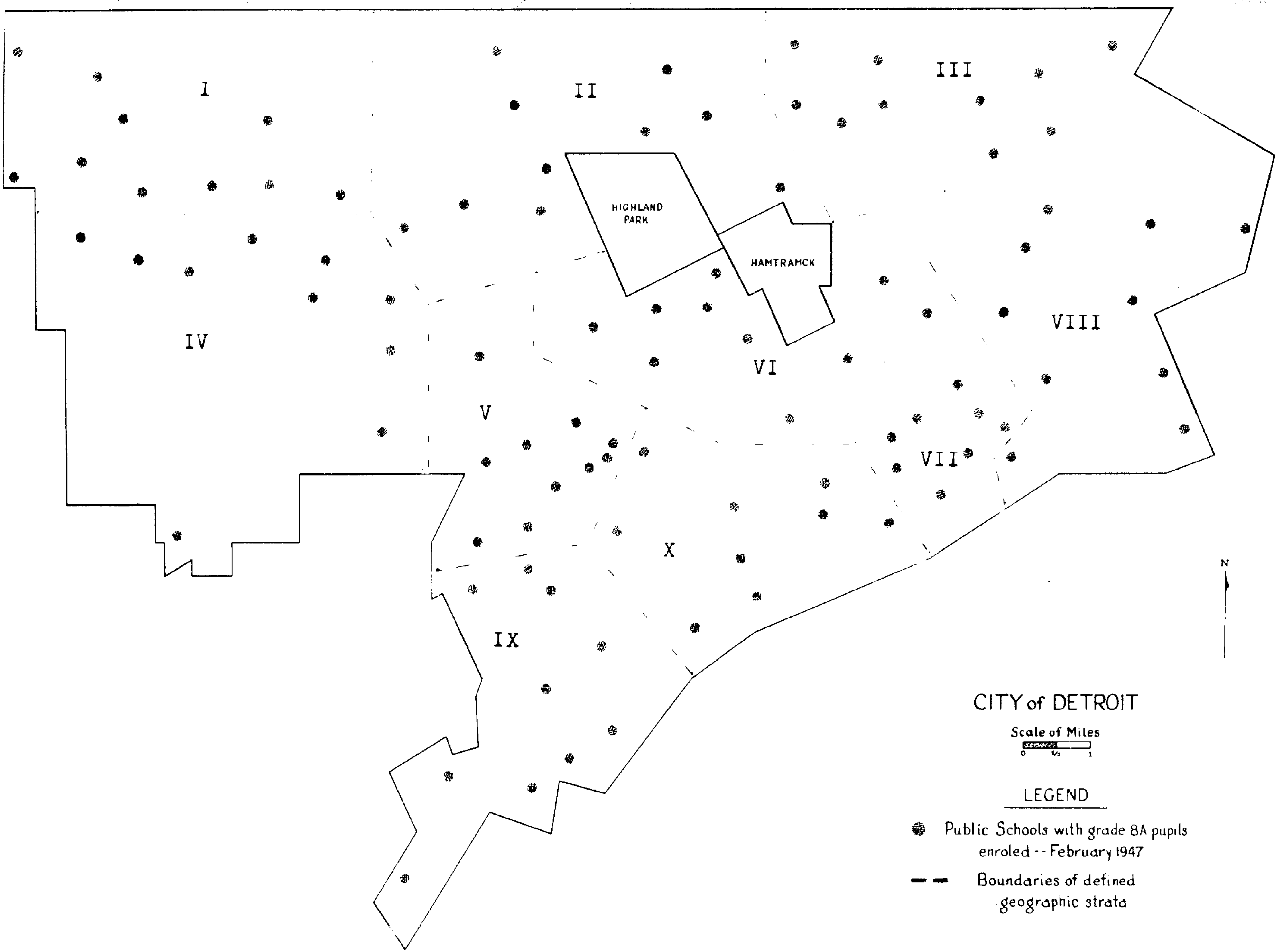

I
II
III
IV
V
VI
VII
VIII
IX
X
HIGHLAND PARK
HAMTRAMCK
N
CITY of DETROIT
Scale of Miles
0 ½ 1
LEGEND
Public Schools with grade 8A pupils enroled - - February 1947
Boundaries of defined geographic strata

TABLE XVII

List of Schools Arranged in 10 Strata According to Geographic Location

Stratum																			
I		II		III		IV		V		VI		VII		VIII		IX		X	
Strat. Code	Sch.	Strat. Code	Sch.	Strat. Code	Sch.	Strat. Code	Sch.	Strat. Code	Sch.	Strat. Code	Sch.	Strat. Code	Sch.	Strat. Code	Sch.	Strat. Code	Sch.	Strat. Code	Sch.
9	79	9	95	9	77	9	74	9	97	9	96	9	81	9	90	9	98	9	91
8	56	8	94	8	75	8	58	8	92	8	89	8	72	8	86	8	93	8	84
7	39	7	83	7	73	7	53	7	78	7	88	7	71	7	70	7	55	7	76
6	37	6	59	6	64	6	51	6	68	6	87	6	69	6	54	6	52	6	65
5	20	5	32	5	63	5	46	5	66	5	85	5	67	5	48	5	47	5	40
4	17	4	31	4	61	4	41	4	62	4	82	4	57	4	44	4	42	4	25
3	15	3	29	3	45	3	35	3	60	3	22	3	49	3	43	3	38	3	24
2	13	2	28	2	27	2	23	2	50	2	19	2	36	2	34	2	14	2	21
1	10	1	18	1	07	1	11	1	33	1	01	1	04	1	30	1	12	1	05
0	08	0	16	0	03	0	09	0	02	..	..	..	..	0	26	0	06	..	..

be expected to be more alike than widely separated neighborhoods. Even a general knowledge of differences in types of neighborhoods in various parts of the city, based on first hand experience, may be enough of a guide in setting up a stratification plan of this type to bring about a measurable improvement in sample accuracy. It may be added also that an admittedly crude method of stratification such as this one will guarantee, in a sample, proportionate representation of schools from all general areas in the city. It is hardly possible that the results from a geographically representative sample could be less accurate, in general, than the results from an unrestricted sample. A partial answer to the question of actual efficiency of the design will be secured by drawing successive samples.

Means of Samples of Schools Stratified by Geographic Location. Ten samples of 50 schools each were drawn successively, using the method of selection that was described in detail[38] in connection with the drawing of 50-school samples stratified by enrolment. In this instance as in the former case, a 50-school sample contains five schools from each of ten strata. Summaries of the results for the ten samples are shown in Table XVIII.

The sample means vary from a low of 66.24 to a high of 67.72 as compared with the population mean of 66.74. The means for six of the samples are above the population mean and four sample means are below. The average of the ten means

---

[38] Pages 65-66.

TABLE XVIII

Numbers of Pupils Tested, Sums of Scores and Mean Scores for Samples of 50 Schools Stratified by Geographic Location

Sample	No. of Schools	No. Pupils Tested	Sum of Scores	Mean Score
1	50	4296	288,317	67.11
2	50	4892	328,427	67.14
3	50	4125	275,195	66.71
4	50	3760	254,136	67.59
5	50	3185	210,960	66.24
6	50	4399	292,016	66.38
7	50	4428	297,819	67.26
8	50	3720	247,450	66.56
9	50	4406	294,261	66.79
10	50	3759	254,545	67.72
Average of Sample Means				66.94
Standard Deviation of Sample Means				.47

is 66.94 which is higher than the "true" mean by .20. The standard deviation of the ten sample average is .47. An unbiased estimate of the standard error of the mean for an unlimited number of samples selected by this method is therefore $.47 \times \sqrt{\frac{10}{9}}$ or .50. This error is smaller than the estimated error for unrestricted samples (.62) derived from repeated sampling, and it is smaller than the error for unrestricted samples of 50 schools as computed from the formula (.54). The estimate of error for the geographically representative sample is, however, considerably larger than the experimentally derived estimate for samples of this size stratified by enrolment (.28). It is larger also than the computed error for enrolment stratification (.46). The evidence thus far suggests that, in the case of this population, stratification by geographic location gives more accurate results than unrestricted selection and less accurate results than enrolment stratification. Additional evidence concerning the accuracy of a

sample drawn from the geographic design may be secured by using the formula to get an estimate of error from the data of a single sample.

<u>Estimate of Error from the Data of a Single Sample.</u> Data for the 50 schools represented as "Sample 1" in Table XVIII were listed and summarized by strata. The five schools drawn into the sample from Stratum I are Schools 79, 37, 17, 15 and 10. The summary of data for this "stratum sample" is as follows:

	School	Number Tested	Sum of Scores
Sample from Stratum I	79	3	208
	37	52	3749
	17	47	3420
	15	50	3453
	10	35	2350
		187	13,180

Data for the respective groups of five schools drawn from each of the other nine strata were arranged in this same form in preparation for using the cluster formula[39] to get an estimate of the standard error of the mean for the entire sample. The formula was then applied to the data for "Sample 1" in the manner described in full detail on pages 82-88. The estimated error for the entire sample is found to be

$$s_{\bar{x}}^{2\prime} = .00126 + .04479 + .01377 + .00305 + .01327 + .20327 + .02100 + .02040 + .00198 + .01587 = .33866$$

and $s_{\bar{x}}' = .5819$

[39] The formula is given on page 85.

This obtained estimate of .58 as the error of the mean for geographically stratified samples of 50 schools supports the tentative conclusion stated above that geographic stratification gives less accurate results than enrolment stratification. At the same time, however, it casts doubt on the tentative conclusion that geographic stratification gives more accurate results than unrestricted selection, the computed error for the latter being .54.

The estimate of error that would be obtained by the formula from a second geographically stratified sample might, of course, be either greater or less than the value (.58) obtained from "Sample 1". In order to get further evidence concerning the error for this design the formula was applied to "Sample 2", which is also represented in Table XVIII. The results are as follows:

$$s_{\bar{x}}^{2}{}' = .00090 + .02825 + .00087 + .00223 + .00537 + .11977 + .00057 + .01440 + .00288 + .03783 = .21307$$

and $s_{\bar{x}}' = .4616$

This estimate of error (.46) derived from "Sample 2" supports both of the tentative conclusions reached earlier--namely, that geographic stratification gives somewhat better results than unrestricted selection and somewhat poorer results than enrolment stratification. Since data for the entire population are available, it is possible to validate the conclusions reached up to this point concerning the relative accuracy achieved by geographic stratification. The "true" error for this design may be determined by substituting in the appropriate error formula the data for all of the schools in each of the ten geographic strata as was done in the case of the enrolment strata.

Sampling Error for Geographic Stratification of Schools Computed from Data for Entire Population. The numbers of pupils tested and the sums of scores for all 97 schools were arranged by geographic strata as illustrated here for Stratum I.

	School	Number Tested	Sum of Scores
Stratum I (all schools)	79	3	208
	56	29	1930
	39	47	3112
	37	52	3749
	20	5	362
	17	47	3420
	15	50	3453
	13	91	6372
	10	35	2350
	8	13	952
		372	25,908

Mean pupil score, $\bar{x}_I$, is 69.6452

Average number pupils per school, $\bar{N}_I$, is 37.2

The quantity $\sum_{j=1}^{M_i} \left[N_{ij}(\bar{x}_{ij} - \bar{x}_i) \right]^2$ or $\sum_{j=1}^{M_i} \left(\Sigma x_{ij} - N_{ij}\bar{x}_i \right)^2$,

from the error formula, is 83,897.709 for Stratum I. By substitution in the entire formula for stratified samples of schools,[40] the data for this stratum give

$$\left(\frac{372}{7724}\right)^2 \left(\frac{10 - m_i}{9\, m_i}\right) \left[\frac{83{,}897.709}{(1383.842)(10)}\right]$$

$= .0141 \left(\frac{10 - m_i}{9\, m_i}\right)$ as representing that part of the error for the total sample contributed by m_i schools drawn from Stratum I.

In order that the error for different sample sizes might be readily determined, similar computations were made for each of the ten strata. The results show that when m_i schools are drawn at random from each stratum the standard error of the mean pupil score for the entire sample is represented by

$$\sigma^2_{\bar{x}} = .0141 \left(\frac{10 - m_i}{9\, m_i}\right) + .0305 \left(\frac{10 - m_i}{9\, m_i}\right) + .0136 \left(\frac{10 - m_i}{9\, m_i}\right)$$

$$+ .0478 \left(\frac{10 - m_i}{9\, m_i}\right) + .3802 \left(\frac{10 - m_i}{9\, m_i}\right) + 1.3479 \left(\frac{9 - m_i}{8\, m_i}\right)$$

$$+ .1044 \left(\frac{9 - m_i}{8\, m_i}\right) + .0993 \left(\frac{10 - m_i}{9\, m_i}\right) + .0328 \left(\frac{10 - m_i}{9\, m_i}\right)$$

$$+ .2426 \left(\frac{9 - m_i}{8\, m_i}\right)$$

---

[40] The formula which calls for data from all schools in the population, viz.,

$$\sigma^2_{\bar{x}'} = \sum_{i=1}^{R} \left[\left(\frac{N_i}{\sum_{i=1}^{R} N_i}\right)^2 \left(\frac{M_i - m_i}{(M_i - 1)(m_i)}\right) \frac{\sum_{j=1}^{M_i} \left[N_{ij}(\bar{x}_{ij} - \bar{x}_i)\right]^2}{\bar{N}_i^{\,2}\, M_i} \right]$$

The weighted errors squared shown here for the ten strata vary in size from .0136 (for Stratum III) to 1.3479 (for Stratum VI), the latter being approximately 100 times as large as the former. The error for a sub-sample of schools from Stratum VI alone will account for more than half the total error for the entire sample.

To determine $\sigma^2_{\bar{x}}{}'$ for a sample of 50 schools, five from each stratum, the substitution $m_i = 5$ was made in the equation above. The result is

$$\sigma^2_{\bar{x}}{}' = .2381$$

$$\text{and} \quad \sigma_{\bar{x}}{}' = .488$$

This computed value of the standard error for geographically stratified samples (.49) is thus slightly smaller than the estimate based on repeated sampling (.50). It is found to be smaller than the first of the two estimates based on the data of a single sample (.58) and larger than the second estimate (.46) derived by that method.

By substituting $m_i = 1$, in place of $m_i = 5$, in the above equation, the standard error for a sample of 10 schools (one from each stratum) is found to be 1.52, or about three times the size of the error for a 50-school sample. The comparable value for a 30-school sample ($m_i = 3$) is .76.

Summary of Results for Sampling by Geographic Strata. In the case of each of three sample sizes, 10, 30 and 50 schools, the results for geographic stratification were found to be more accurate than for unrestricted selection and less accurate than for enrolment stratification. In interpreting these results it should be borne in mind that the particular definitions of the ten strata used here represent only one of an almost unlimited number of possible geographic arrangements of the 97 schools in sub-groups of nine or ten schools each. An inspection of the map on page 96 will make it evident that any number of different patterns of lines might have been drawn around groups of schools to get ten strata, each containing approximately the same number of schools. By contrast, the ten enrolment strata used earlier were specified by an objective index which could be used to array the 97 schools in only one way. The results secured here do not warrant broad generalizations concerning the relative efficiency of geographic stratification in general. Other possible patterns of arrangement of the schools by geographic sub-groups prior to drawing samples might yield better (or worse) results than those obtained.

In the case of geographic Stratum VI, for example, the weighted variance for a sample of schools turned out to be extremely large. This stratum happened to include a large school whose mean score is second highest out of the 97 and also a second large school whose mean is sixth from the lowest out of the 97. A general knowledge of the respective neighborhoods surrounding the two schools in question indicated in advance that grouping them together in the same stratum would probably add significantly to the variance for that stratum. It was not expected, however, that the effect would be as great as it actually turned out to be. It would have been quite appropriate in setting up the plan of grouping to have separated these two schools by placing them in different strata, even at the expense of distorting the boundary lines of adjacent strata. In fact it would be advisable under practical conditions to make such adjustments in strata based on more or less subjective judgments. Any increase in precision which may be brought about by such procedures cannot be a spurious increase because the estimate to be made of the precision actually attained will be based on the data of the sample randomly drawn. It has been shown earlier that the estimate of error for a stratified sample is quite independent of any "purposive" methods used at the outset in setting up the strata.

The fact that both the enrolment control and the geographic control yielded results that were less variable than those obtained by unrestricted selection suggested the hypothesis that a design involving a double control, i.e., geographic location *and* enrolment, might be more precise than either of the former sampling plans. This is the fourth and last of the designs that will be applied in which the school is used as the sampling unit.

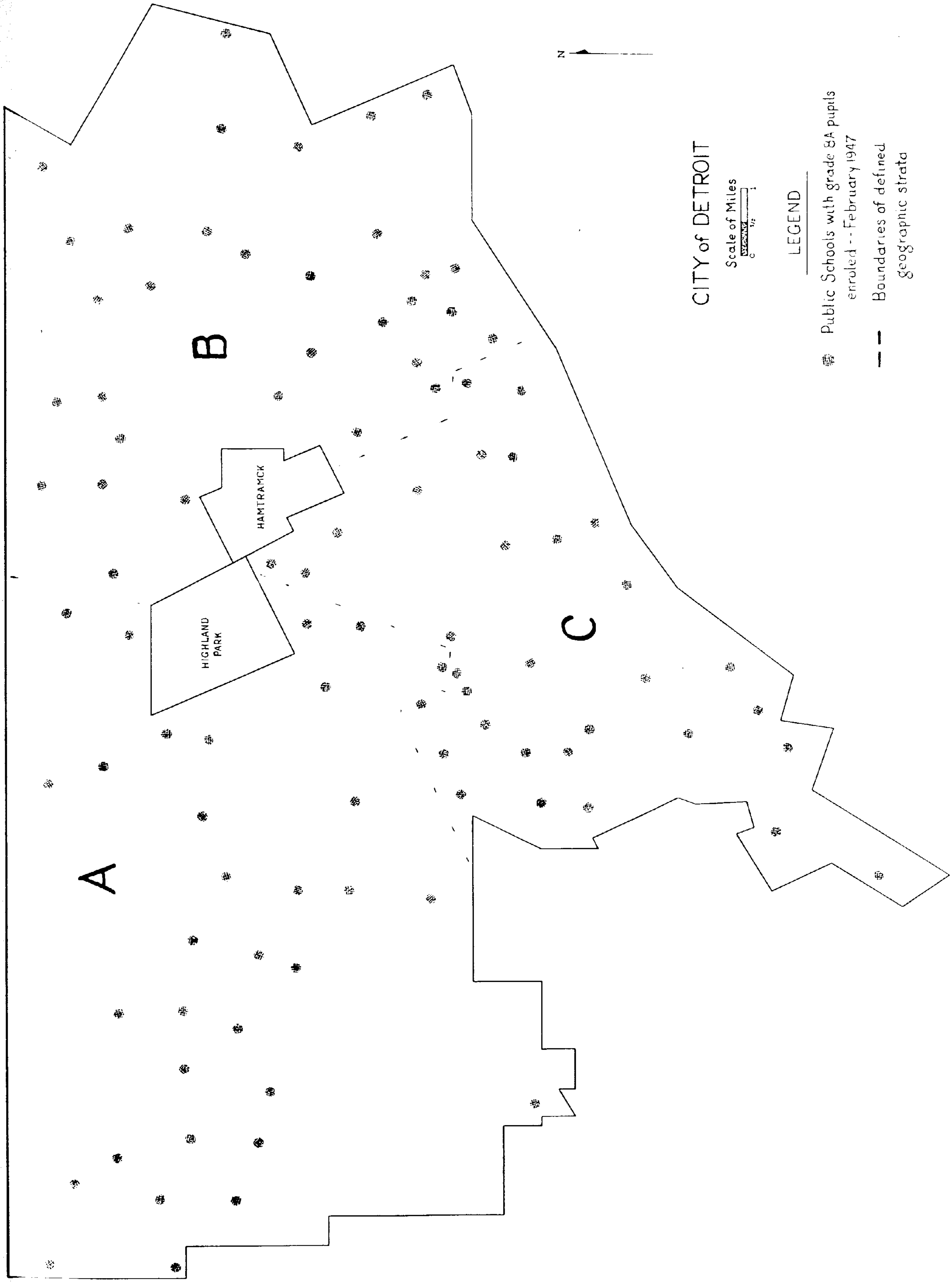
III. RESULTS FOR STRATIFIED SAMPLING OF SCHOOLS

Primary Stratification by Geographic Location
Secondary Stratification by Size of Enrolment

<u>Method Used in Stratifying the Schools</u>. Neither the geographic strata nor the enrolment strata used previously could be employed in this new design. An arrangement of schools in ten enrolment classes within each of the ten geographic strata would have resulted in a total of 100, i.e., (10 x 10) different cells or strata. The number of different strata would thus be larger than the total number of schools. Furthermore, the distributions of enrolments of the schools within the previously used geographic strata do not correspond to the distribution of enrolments for the 97 schools. A great many cells would therefore be empty. For these reasons it was necessary, first, to set up fewer geographic strata containing much larger numbers of schools, and second, to set up enrolment classes separately within each of the new geographic strata.

Using an outline map with the 97 schools spotted, lines were drawn dividing the city into three large areas, each containing approximately the same number of schools. These three geographic areas are designated on the map on page 108 by the symbols "A", "B" and "C". Area A includes 33 schools, Area B, 32 schools and Area C, 32 schools.

The next step was to list the schools <u>within each area</u> in order of size of enrolment for grade 8A. Among the 33 schools in Area A, School 85 is largest with an enrolment of 459 and School 79 is smallest with three pupils enroled. In


N
CITY of DETROIT
Scale of Miles
0 1/2 1
LEGEND
Public Schools with grade 8A pupils enroled - - February 1947
Boundaries of defined geographic strata
A
B
C
HAMTRAMCK
HIGHLAND PARK

Area B, School 90 is largest with an enrolment of 423 and School 54 is smallest with 16. The enrolments of the largest and smallest schools in Area C are 309 and 8, respectively. Although it was thought desirable at the outset to have each of the enrolment classes within a given geographic area include at least 10 schools, the markedly skewed distributions of school enrolments for the three areas made this plan appear inadvisable. If as many as 10 schools were to be included in each enrolment sub-group within Area A, for example, one such sub-group would have to include a range of enrolments from 88 to 459. It was decided, therefore, to set up smaller sub-groups for the very large and for the very small schools so as to get a reasonable degree of enrolment homogeneity within each stratum. The stratification plan as finally applied is illustrated below for Area A.

GEOGRAPHIC AREA A

Group I		Group II		Group III		Group IV	
School	No. Enrol.	School	No. Enrol.	School	No. Enrol.	School	No. Enrol.
79	3	74	49	58	105	97	401
59	27	56	32	35	78	95	204
46	30	53	43	32	67	94	240
20	6	51	54	29	112	92	189
19	32	41	54	13	92	89	233
08	14	39	48	09	89	85	459
		37	52				
		31	47				
		28	55				
		23	55				
		18	66				
		17	50				
		15	55				
		11	42				
		10	39				

The 33 schools in Area A are shown to be divided into four separate groups. Each of three groups includes six schools, and one group includes 15 schools. The respective enrolments of the six schools in Group I range in size from 3 to 32. In Group II which includes 15 schools the enrolments range from 32 to 66. Group IV which contains the largest schools shows enrolments varying from a low of 189 to a high of 459. It was not possible to attain, in these four groups, the degree of enrolment homogeneity achieved in the design used earlier where the 97 schools were divided into ten enrolment sub-groups without respect to the factor of geographic location.

Each of the four strata shown above for Area A may be sampled proportionately by taking two schools from Group I, five schools from Group II, two schools from Group III and two schools from Group IV. Such a sample of 11 schools would consist of one-third of the total number of schools in each group. Each of the 33 schools would have an equal chance of being drawn--one chance in three.

The 32 schools located in Area B, and the 32 schools located in Area C were arranged in four sub-groups according to size of enrolment by the same method used for Area A. The stratification design for all 97 schools is shown in Table XIX.

It will be noted in Table XIX that Stratum "A-II" contains 15 schools whereas Strata "B-II" and "C-II" each contain 14 schools. The odd number of schools (97) in the total population made it impossible to avoid this inequality in numbers

TABLE XIX

List of Schools, and Corresponding Stratification Codes, Arranged in 12 Groups According to Geographic Location and Size of Enrolment

Enrolment Sub-Group	Geographic Area A Strat. Code	A Sch.	B Strat. Code	B Sch.	C Strat. Code	C Sch.
	6	97	6	90	6	98
	5	95	5	88	5	93
IV	4	94	4	86	4	91
	3	92	3	83	3	87
	2	89	2	82	2	84
	1	85	1	81	1	21
	6	58	6	77	6	96
	5	35	5	64	5	76
III	4	32	4	63	4	65
	3	29	3	30	3	62
	2	13	2	26	2	60
	1	09	1	16	1	25
	15	74	..	..	..	..
	14	56	14	75	14	69
	13	53	13	70	13	66
	12	51	12	67	12	47
	11	41	11	61	11	42
II	10	39	10	57	10	40
	09	37	09	49	09	38
	08	31	08	48	08	33
	07	28	07	45	07	24
	06	23	06	44	06	22
	05	18	05	43	05	14
	04	17	04	27	04	06
	03	15	03	07	03	05
	02	11	02	04	02	02
	01	10	01	03	01	01
	6	79	6	73	6	78
	5	59	5	72	5	68
I	4	46	4	71	4	55
	3	20	3	54	3	52
	2	19	2	36	2	50
	1	08	1	14	1	12

of schools in corresponding strata for the three geographic areas. Since the inequality appears in the three largest strata, it will have a relatively small effect on the sampling rate when the same number of schools is drawn from each of them. For example, when five schools are selected from Stratum A-II the sampling rate is .333 (one out of three); when five schools are drawn either from Stratum B-II or from Stratum B-III the rate is .357 (five out of 14).

With twelve separate strata specified, four within each of the three geographic areas, and with each of the 97 schools identified by a special "stratum code number" as shown in Table XIX the design has been completely laid out and the next step is to draw a sample.

Mean Scores for Samples of 33 Schools. A 33-school sample was obtained by choosing two schools from each of the nine 6-school strata and five from each of the three 15-school (or 14-school) strata. The table of random numbers was used in making the selections within each stratum in exactly the same way as described in preceding sections of this chapter. For example, the first two unlike numbers encountered in the table were "4" and "1". These two numbers specified the two particular schools from Stratum A-IV to be included in the sample. It will be seen in Table XIX above that the two schools in Stratum A-IV bearing the respective codes "4" and "1" are School 94 and School 85. The next two unlike random numbers, "5" and "4", specified two schools in Stratum B-IV, School 88 and School 86. The two schools chosen from Stratum C-IV were

87 and 21. This procedure was followed for the nine strata containing six schools each. It was then necessary to take the first five unlike two-digit numbers between 01 and 15 in the random numbers table as specifying the five schools to be selected from Stratum A-II which contains 15 schools. The five numbers in the order appearing in the table were 14, 11, 18 and 53. By a similar procedure, five schools were chosen from Stratum B-II and five from Stratum C-II. Data for each of the 33 schools which constitute this sample are shown in Table XX.

TABLE XX

Numbers of Pupils Tested and Sums of Scores for a Sample of 33 Schools Stratified by Geographic Location and by Size of Enrolment

Enrolment Subgroup	Area A			Area B			Area C		
	Sch.	No. Tested	Sum of Scores	Sch.	No. Tested	Sum of Scores	Sch.	No. Tested	Sum of Scores
IV	94	228	15,754	88	146	9,243	87	155	9,351
	85	435	31,484	86	282	18,909	21	149	8,576
III	32	66	4,885	16	49	3,054	76	75	4,987
	09	88	6,441	30	53	3,586	62	57	3,865
II	56	29	1,930	43	36	2,291	24	46	2,985
	41	51	3,291	70	34	2,397	42	45	2,803
	11	39	2,864	61	30	2,021	22	47	2,769
	18	66	4,422	49	28	1,742	66	42	2,536
	53	42	2,794	04	31	1,832	01	37	2,381
I	20	5	362	34	16	980	55	22	1,435
	46	29	2,059	72	19	1,163	52	19	1,183

Sum of Number Tested Over All Strata	2,496
Sum of Sums of Scores Over All Strata	166,375
Mean Score for Sample	66.66

The dotted horizontal lines, along with the solid vertical lines in Table XX, divide the data into twelve sub-groups representing the twelve different strata from which the schools were drawn. The three sub-groups of data thus blocked in at the bottom of the table show the numbers of pupils tested and the sums of scores for the six small schools drawn into the sample--two from each geographic area. The three sub-groups of data extending across the table at the top give similar information for the six large schools. Comparisons among the columns headed "Number Tested" for Areas A, B and C show that the schools within each enrolment sub-group are roughly comparable in size from area to area. A total of 2,496 pupils is included in the entire sample. The mean score, shown at the foot of the table, is 66.66. This is .08 lower than the population average (66.74).

Nine additional samples of 33 schools each were drawn by the same method. The numbers tested, sums of scores and means are given in Table XXI for ten samples of this size, including the one that has been described in detail.

The ten means for 33-school samples vary in size from 64.88 to 67.12, a range of -1.86 to +.38 around the known population mean of 66.74. The average of the ten sample means is 66.25, approximately one-half a score point below the average for the population, and the standard deviation of the obtained means is .63. This empirically derived standard deviation multiplied by $\sqrt{\frac{10}{9}}$ gives .66 as an estimate of the standard error of the mean for <u>any</u> sample of 33 schools drawn by this method.

TABLE XXI

Numbers of Pupils Tested, Sums of Scores and Mean Scores for Samples of 33 Schools Stratified by Geographic Location and by Size of Enrolment

Sample	No. of Schools	No. Pupils Tested	Sum of Scores	Mean Score
1	33	2496	166,375	66.66
2	33	2418	162,083	67.03
3	33	2558	168,920	66.04
4	33	2391	156,955	65.64
5	33	2500	165,222	66.09
6	33	2909	195,247	67.12
7	33	2514	166,091	66.07
8	33	2503	166,577	66.55
9	33	2715	180,418	66.45
10	33	2257	146,427	64.88
Average of Sample Means				66.25
Standard Deviation of Sample Means				.63

The average of the ten means in Table XXI differs from the population average by .49. This difference is larger than any of the differences found previously between the average of an array of means secured by repeated sampling and the known average score for the 7724 pupils constituting the entire population. Such a result suggests at first glance that there may have been some element of bias in the method of selecting schools which constituted the samples. However, it was shown above in the detailed descriptions, both of the design and of the method of selection, that each of the 97 schools had an equal chance of being drawn into any sample. A minor violation of the principle of uniform probability of selection was pointed out in connection with strata B-II and C-II, but it seems very unlikely that the small difference between the proportion of schools chosen from these two strata and the proportion taken from the other 10 strata could produce a bias as large as the obtained difference between the average of the 10 sample means (66.25) and the population average (66.47). The t-test was applied in order to determine the level of significance of this difference.

It will simplify the interpretation of t-test results in this particular situation if (a) the 10 sample means under consideration are thought of as measurements made of 10 <u>individuals</u> chosen from a very large population of similar "individuals", and (b) the average score for the 7724 pupils, i.e., 66.74, is thought of as the true average for this new hypothetical population. Making these two assumptions, for the moment, the ten "individuals" were selected, presumably at random, from a "population" whose mean is known to be 66.74. The standard deviation for this hypothetical population, however, is not known. The sample of 10 cases was found to have a mean of 66.25 and a standard deviation of .63. The unknown standard deviation of the "population" is <u>estimated</u>, from the data of the sample, to be .66. The standard error of the <u>mean</u> of the 10 cases is therefore estimated to be, .66 divided by $\sqrt{9}$, or .22. The difference between the true mean and the mean of the sample, 66.74-66.25 or .49, is more than twice as large as the standard error of the sample mean. There is reason to suspect, therefore, that there may have been some element of bias in the process of selection used in drawing the sample after all, even though it was presumed to have been random. Because of the very small size of the sample--only 10 cases-- the t-test provides the most precise method for making further analyses so that the suspicion of bias in the selection process may either be confirmed or allayed.

As it applies to this problem, "t" is defined[41] as

$$t = \frac{\bar{x}' - \bar{x}}{s_{\bar{x}}'}$$

where $\bar{x}$ = the true mean

$\bar{x}'$ = the mean of the sample

$s_{\bar{x}}'$ = the estimated standard error of the mean of the sample

By substitution of the values given above for each of these three quantities

$$t = \frac{66.25 - 66.74}{.22}$$

$$= -2.23$$

From the t-table it is seen that a "t" as large as this one (2.23) would be found about as often as one time out of 20 when samples of 10 cases were actually drawn at random from the hypothetical population under consideration here. The t-test results thus indicate that the sampling procedure used in this instance in drawing 10 cases could have been strictly

---

[41] C. W. Odell, An Introduction to Educational Statistics, p. 246. New York: Prentice-Hall, Inc., 1946.

random, and that the somewhat unusual nature of this sample is due to chance variations that would be expected to occur in the case of one such sample out of 20. Although this conclusion does not greatly increase one's confidence that the sample was drawn at random, the obtained "t" is not so large as to raise crucial doubts concerning the sampling method.

Mean Scores for Samples of 48 Schools. If one-half of the schools in each of the 12 geographic-enrolment strata are drawn into a sample, the sample will consist of three from each of the nine strata containing six schools, for a total of 27, and seven from each of the three strata containing 14 (or 15) schools, for a total of 21, making a grand total of 48 schools. A sample of this size includes approximately one-half the total number of schools in the population (97), and is slightly smaller than the largest samples drawn according to the other sampling designs that have been described--i.e., 48 as compared with 50. It is obviously impossible to draw a sample of 50 schools from the geographic-enrolment design and have each of the 12 strata proportionately represented.

A group of means for 48-school samples was produced by repeated sampling, the individual schools being selected from the different strata according to the method described for drawing groups of 33-schools. The first sample thus drawn includes 3918 pupils. The mean pupil score for this sample is 66.41 which is lower than the population mean (66.74) by .11. Twenty such samples chosen successively and their respective means computed. After each drawing, the 48 schools constituting a given sample were, of course, put back into the population of schools so as to be eligible for selection in the succeeding sample.

It will be recalled that only ten 50-school samples were actually drawn from each of three previously described designs, (a) simple random, (b) enrolment stratification, and (c) geographic stratification. It was decided to secure data for 20 rather than 10 samples of 48 schools from this fourth design for two reasons: first, because of the desire to make a somewhat more reliable empirical test of the hypothesis that geographic-enrolment stratification would yield more accurate results than any of the other three plans and second, because of the desire to determine again for this design the significance of the difference between the average of the sample means and the true mean for the population, using twice as many sample means as were used when the comparable significance test was applied for 33-school samples. Data for the 20 samples each consisting of 48 schools are presented in Table XXII.

The actual standard deviation of the means in Table XXII is .52 and the estimate (based on the 20 cases) of the standard error for samples of this size is .53. The average of the 20 means is 66.73 which is almost exactly the same as the population mean (66.74). The difference between these two averages (.01) is so small that it is hardly worthwhile to apply the t-test to determine its degree of significance.[42] This result for 48-school samples tends definitely to allay the suspicion of bias in the method of selecting schools which was first raised in connection with results obtained by drawing 33-school samples.

TABLE XXII

Numbers of Pupils Tested, Sums of Scores and Mean Scores for 20 Samples of 48 Schools Stratified by Geographic Location and by Size of Enrolment

Sample	No. of Schools	No. Pupils Tested	Sum of Scores	Mean Score
1	48	3918	260,209	66.41
2	48	4094	275,370	67.26
3	48	4007	269,824	67.34
4	48	3650	240,730	65.95
5	48	4075	271,846	66.71
6	48	3530	234,423	66.41
7	48	4150	277,049	66.76
8	48	3839	255,321	66.51
9	48	3866	257,364	66.57
10	48	3702	247,753	66.92
11	48	3886	258,248	66.46
12	48	3926	260,620	66.38
13	48	3888	260,332	66.96
14	48	3695	245,710	66.50
15	48	3867	261,195	67.54
16	48	3999	270,722	67.70
17	48	3498	230,014	65.76
18	48	4086	274,645	67.22
19	48	3804	255,539	67.18
20	48	3820	252,034	65.98
Average of Sample Means				66.73
Standard Deviation of Sample Means				.52

42

$$t = \frac{66.74 - 66.73}{\frac{.53}{\sqrt{19}}} = .083.$$

With d.f. = 19, a "t" of .083 fails to be significant even at the 90 per cent level.

The data secured thus far from the design based on dual stratification controls--geographic location of the school and size of enrolment--have not supported the hypothesis that this design would give more accurate results than any of the other three methods of selecting samples of schools. Judging from experimental evidence, the dual control design appears to yield about the same accuracy as unrestricted selection and to give somewhat less accurate results than either enrolment stratification or geographic stratification when these latter two methods of grouping the schools are used independently of each other. The estimated standard error for 48-school samples selected from the 12 geographic-enrolment strata is .53, whereas the computed standard errors for 50-school samples from the other three designs, unrestricted selection, enrolment stratification and geographic stratification are, respectively, .54, .46 and .49. A more precise determination of error for the dual stratification design may be made by substituting in the cluster formula the data for all of the schools in each stratum just as was done for each of the other three plans.

Computed Standard Errors for 33-School Samples and 48-School Samples. Computations had to be made for 12 separate strata in deriving the error for this design. The detailed steps in the computation are precisely the same as those already described for enrolment stratification and for geographic stratification. The results for a 33-school sample using data from all of the schools in each stratum are as follows:

$$\sigma^2_{\bar{x}}{}' = .1780 + .0075 + .0086 + .0059 + .1179 + .0063 + .0048 + .0077 + .0920 + .0308 + .0112 + .0049 = .4756$$

and $\sigma_{\bar{x}}{}' = .690$

This computed value of the error for a 33-school sample (.69) differs by only .03 from the estimate (.66) derived from the 10 samples actually drawn. It is interesting to note in passing that the estimate based on the 10 "cases" was quite accurate in predicting the "true" <u>variance</u> of means even though the corresponding estimate made of the "true" <u>average</u> of a distribution of such means was shown above to be somewhat inaccurate.

The standard error for a 48-school sample, computed from the data for all schools in each stratum, gives the following results:

$$\sigma^2_{\bar{x}}{}' = .0890 + .0038 + .0049 + .0030 + .0590 + .0032 + .0027 + .0039 + .0460 + .0154 + .0062 + .0025 = .2396$$

and $\sigma_{\bar{x}}{}' = .489$

By comparing this computed error (.49) with the previously determined estimate (.53) based on the 20 means in Table XXII it is found that the difference between the two is .04. This difference is practically the same as that found between the estimated and computed errors for 33-school samples.

<u>Summary of Results for Sampling by Geographic-Enrolment Strata</u>. This stratification design was originally set up with the expectation that it would give more accurate results than any of the three plans tried out previously. The results expected were not achieved. Primary stratification of schools by geographic location with sub-stratification by size of enrolment proved to be better than unrestricted selection but not as good enrolment stratification used alone. Its accuracy is equal to that of geographic stratification used alone.

The failure of the dual control design to show any increase in accuracy over geographic stratification used alone is probably due to the much larger size of primary geographic strata used in the former design. It will be recalled that when the geographic control was used alone each stratum consisted of only 10 (or 9) schools whereas each of the three primary strata for the dual control plan consisted of 32 (or 33) schools. Apparently, the variability of school means within these larger strata is not much less than the variability of school means for the entire population. If this is true, there would, of course, be little gain from such stratification.

A similar interpretation may be made of the findings which show the geographic-enrolment plan to give less accurate results than enrolment stratification used alone. Maximum stratum homogeneity was achieved with respect to <u>size of enrolment</u> when the enrolment control alone was used to divide the 97 schools into 10 groups. Insofar as homogeneity in size of school enrolment is correlated with homogeneity of mean scores, any stratification plan which minimizes variability in enrolment within the strata will also minimize variability in mean scores within the strata.

It is a difficult task, at best, to set up a reasonable and promising looking plan for using more than one control factor in stratifying a population which consists of only 97 schools. This is especially true in a situation where there are wide differences among the schools with respect to one of the control factors such as size of enrolment. In the present instance it was necessary to sacrifice some measure of enrolment homogeneity and some measure of geographic homogeneity in order to apply both types of control simultaneously. This resulted in some loss of potential accuracy that could be achieved by using either of the two controls separately. However, the result was still better than could be attained by unrestricted selection.

VI. COMPARISON OF RESULTS FOR FOUR DIFFERENT METHODS OF SAMPLING BY SCHOOLS

This section concludes the interpretation of results secured from sampling a pupil population by using the school as the sampling unit. Four different methods of drawing samples of schools were applied for the purpose of predicting the average pupil score for the entire pupil population in 97 schools. The predictive efficiency, i.e., the variability of the sample mean, was determined for each of the four plans by two independent methods. The first of these two techniques for deriving the error of a sample mean consisted of (a) repeated drawing of samples in order to produce an actual distribution of sample means, and (b) deriving from this distribution an estimate of the standard deviation for an unlimited number of similar means that could be secured by the procedure of repeated sampling. The second method used in deriving the standard error of a sample mean was made possible by the fact that data were available for the entire population. This technique consisted of substituting in the error formula for cluster sampling the complete data for all schools in the population. Still a third procedure, namely substitution of data for a <u>single</u> sample in the appropriate form of the error formula, was used for purposes of illustration in the case of two of the four designs. Summary descriptions of the different methods of drawing samples will be given before comparing results obtained from their application.

<u>Results of the Four Designs</u>. The first method to be used consisted of drawing a given number of schools from the total group purely at random and without any restrictions whatever on the selection process. Under this plan each school has one chance out of 97 to be chosen on the first drawing. On the second drawing, each of the remaining 96 schools has one chance out of 96 to be taken into the sample, and so on. This design has been referred to as "unrestricted selection". When using this method of selection, it is possible to get a valid estimate of the error of the sample mean without having advance knowledge concerning any characteristic of the schools to be sampled.

The second method applied in choosing samples required advance knowledge of the number of grade 8A pupils enroled in each school. With this knowledge at hand the 97 schools were arranged in 10 sub-groups, the 10 smallest schools being brought together to constitute one of the groups, the next 10 schools in order of size constituting the second group, etc. Each of the 10 groups contained either 10 or 9 schools. With the 97 schools thus stratified by enrolment, samples were drawn from the different strata <u>separately</u>. When the first drawing was made from the 10 schools in Stratum I, for example, the remaining 87 schools in Strata II through X had <u>no</u> chance of being chosen. Even though each individual drawing was restricted to a given stratum in this way, every school in the population had an equal chance to be selected because of the fact that the same proportion of schools was taken from each stratum in turn. This design has been referred to as "stratification by enrolment".

The third method applied in choosing samples required advance knowledge of the geographic location of each school. The 97 schools were first spotted on an outline map, then boundary lines were drawn around groups of schools that are located relatively close together. These lines were drawn in such a way as to divide the city into 10 geographic areas or strata, each containing either 10 or 9 schools. These geographic areas were labeled "Stratum I", Stratum II", etc. With the 97 schools thus stratified by geographic location, samples were drawn from each of these strata separately just as was done in the case of the enrolment stratification plan. This arrangement of schools by geographic areas in preparation for drawing samples has been referred to as "geographic stratification".

The fourth and last method to be applied made use of advance knowledge concerning both geographic location of each school and size of enrolment of each school. With all the schools spotted on a map, boundary lines were drawn which divided the city into three large areas. One of these areas contained 33 schools and the other two contained 32 each. The schools within each area were then arranged in sub-groups according to size of enrolment. The smallest schools in Area A, for example, constituted "Stratum A-I", the smallest schools in Area B constituted "Stratum B-I", etc. Four enrolment subgroups were set up in this way within each of the three large geographic areas making a total of 12 separate strata. These strata did not all contain the same number of schools as in the case of the other two stratification plans. However, in drawing a sample the proportion of schools chosen was the same for all 12 strata. This design has been referred to as "geographic-enrolment stratification".

Standard Errors for the Four Designs. Comparisons of the respective errors for the four sampling plans will be made by using results obtained from the largest samples drawn. The largest sample drawn from the dual control (geographic-enrolment) design was 48 schools. Fifty-school samples were drawn from each of the other three. The two standard errors obtained for each plan are given in Table XXIII. One is an estimated value based on data secured by repeated sampling. The other is derived from substitution in the formula.

It will be seen in the table that the error is smallest for enrolment stratification and largest for unrestricted selection. Results from the former design are therefore most accurate and results from the latter are least accurate. Both the estimated errors and the computed errors support this conclusion.

TABLE XXIII

Standard Errors of the Mean (Estimated and Computed) for Four Different Methods of Sampling by Schools

Sampling Design	Size of Sample	Estimated Error*	Computed Error*
Unrestricted Selection	50 Schools	.62	.54
Enrolment Stratif.	50 Schools	.29	.46
Geographic Stratif.	50 Schools	.50	.49
Enrol. - Geog. Stratif.	48 Schools	.53	.49

*Standard error of the mean <u>pupil score</u>.

The estimated error for geographic stratification, given in Table XXIII, is slightly smaller than the corresponding error for the dual control (geographic-enrolment) design. On the basis of this evidence the former plan would be judged to be somewhat better than the latter. However, the standard errors computed from the formula show these two designs to give sample results that are equal in reliability.

From the point of view of accuracy of sample results, there is relatively little difference among the three stratified designs. The computed standard errors for samples of approximately 50 per cent of the schools drawn by these three methods are .46, .49 and .49. All three methods show a gain over unrestricted selection which gives a computed standard error of .54 for samples of the same size.

Conclusion. The analyses presented in this chapter show that when a pupil population is sampled by using a group of pupils, i.e., a school, as the sampling unit, it is possible to determine the magnitude of sampling error objectively and validly. If data for the entire population are available, the standard error for a sample of a given size drawn by a given method may be determined precisely. When the only data available are those secured from a single sample, it is possible to get an objective estimate of the standard error of the obtained mean pupil score by appropriate analysis of the data of the sample itself.

In the case of the population under investigation here, it has been shown that the mean score for a random sample of 10 schools including more than 800 pupils has about the same reliability as the mean score for a sample of 30 individual pupils drawn at random. It was necessary to increase the sample size up to approximately 50 per cent of the total number of schools in the population in order to get the precision desired in sample results. The desired level of precision was defined as a standard error of the sample mean equal to one-twentieth the standard deviation of the 7724 individual pupil scores ($9.05 \div 20 = .45$). This degree of precision could not be attained with a sample of 50 schools chosen by unrestricted selection. It was attained, approximately, with a sample of 50 schools stratified by size of enrolment. A 50-school sample and a 48-school sample drawn, respectively, from the other two stratification designs fall short of this standard by only a small margin.

The average number of pupils in a 50-school sample is about 4000. The desired level of precision, stated above, could have been achieved with a random sample of 380 <u>individual pupils</u>. From a purely statistical point of view, sampling by schools is therefore grossly "inefficient" as compared with sampling by pupils. But in view of practical considerations already discussed in detail, any one of the four school designs used here is actually more economical than sampling by individuals. The cost of additional test booklets needed for testing larger numbers of pupils in school groups would certainly be less than the cost of giving teachers special training in sampling to guarantee the exercise of rigorous control over the process of drawing individual pupils into a sample. And for reasons pointed out earlier, the administrative and supervisory cost of the survey itself would be increased if the individual rather than the school were used as the sampling unit.

It might have been possible to achieve some gain in practical efficiency for the three stratified designs described in this chapter by the use of "disproportionate" sampling among the different strata. Data for the enrolment design (page 78) showing much greater variability for Strata IX and X than for the other eight strata illustrate this possibility. In situations such as this one it is sometimes advantageous to draw relatively larger numbers of units from the strata known to have greater variability and relatively smaller numbers of units from the strata having less variability. If sufficient information is available in advance concerning the population to be sampled, it may be possible to apply the principle of "optimum allocation"[43] in designating the proportion of units to be drawn from each stratum.

In determining a test norm it is, of course, relatively unimportant that an arbitrarily defined level of precision be attained exactly. It is of crucial importance, however, to determine what the accuracy of obtained results actually turns out to be. There can be no objective evaluation of differences between "the norm" and other test results obtained from specific groups or from individual pupils without an estimate of the reliability of the norm itself. The sampling

[43] W. Edwards Deming and Willard Simmons, "On the Design of a Sample for Dealers' Inventories", Journal of the American Statistical Association, Vol. 41 (1946), pp. 21-23.

methods described here yielded valid, objective estimates of error even though the school rather than the individual pupil was used as the sampling unit. In the succeeding chapter similar methods will be used to draw samples of <u>classes</u> instead of schools.

CHAPTER IV

RESULTS FOR SAMPLING BY CLASS GROUPS

Sampling by classes calls for the selection of sub-groups of pupils within a school. For this reason the class is not quite as convenient to use as a sampling unit as is the school. In making plans for drawing a sample of classes it is necessary to secure up-to-date information on the class organization in every school in the designated population.[1]

Definition of Class Groups. Pupils are organized in class groups within a school for instructional purposes. In an elementary school where the basic curriculum is generally the same for all pupils, this grouping is usually made on the basis of grade status. Principals try to arrange the groups in such a way that all pupils in a given group are also in the same grade. It is sometimes necessary, however, to put two or more adjacent grades together in the same class in order to equalize the sizes of different classes. When this is done, the grade 8A pupils in such a class will constitute

---

[1] This would not be an essential requirement for more complex stratified designs for sampling very large populations of pupils in several thousand different schools. In such a situation the design might call for an initial selection of communities, then sub-sampling of classes within each chosen community. It would be necessary to get detailed information on classes only for those communities to be sub-sampled.

only a part of the total membership of the class. In a Detroit elementary school all class groups are designated as sections and each section is identified by a number. The pupils in a given section remain together for their various curricular experiences throughout the school day whether the section represents only one half-grade or two half-grades. Sections thus represent administrative units as well as an instructional groupings. Membership reports prepared by elementary schools each month show analyses of membership by grade for each section. From these reports it is possible to identify every elementary class (section) in the city that enrols pupils in a given half-grade.

About one-half of the pupils in grades 7 and 8 in Detroit attend junior high schools[2] rather than elementary schools. Pupils are not grouped by "sections" in the junior high schools. An alternative method of grouping is necessary because there is some measure of curriculum differentiation requiring greater flexibility in individual pupils' schedules of classes during the school day. For example, a class, consisting of a group of grade 8A pupils brought together the first period in the morning for instruction in English, may not remain an intact class group during the rest of the day. For this reason, the term "class" does not designate an administrative grouping of pupils in a junior high school as the term "section" does for

---

[2] "Intermediate" rather than junior high is the term used in Detroit to designate a school enroling pupils in grades 7, 8 and 9 only.

an elementary school. "Homerooms" rather than classes represent the administrative sub-groups. The average size of these homeroom groups is about 35 pupils, approximately the same as the average class. Each pupil is a member of such a group which meets for 30 minutes every day. Membership reports prepared by junior high schools each month show analyses of total membership by grade for each homeroom. From these reports it is possible to identify every junior high school homeroom group in the city that enrols pupils in a given half-grade. Therefore, in defining the population of grade 8A clusters (classes) from which samples were to be drawn, the cluster was designated as a "section" for elementary schools and as a "homeroom" for junior high schools. This dual designation specifies exactly the sub-clusters of grade 8A pupils within every school.

Membership reports for February 1947[3] show that grade 8A pupils were enroled in 116 different sections in 79 different elementary schools. Junior high schools membership reports for the same date show grade 8A pupils enroled in 121 different homerooms in 18 different schools. There was a total, therefore, of 237 classes with grade 8A pupils enroled.

The first of the several major divisions of this chapter presents summaries of data for all 237 classes. These data give a detailed description of the "population of classes" from which samples will next be drawn.

---

[3] Detroit Public Schools, Form 533, February 1947.

I. SUMMARY OF DATA BY CLASS GROUPS FOR TOTAL POPULATION

Number of Pupils Enroled and Number Tested in Each Class. In a number of elementary schools the membership in a given class consisted of both grade 8B and grade 8A pupils. In these instances the defined "class" is in reality only a portion of the actual class group. For example, a class (Section 19) in one elementary school has a membership of 38. Thirty of these pupils are in grade 8B. The remaining eight are in grade 8A. According to the definition used here this "class" consists of eight grade 8A pupils. The fact that two adjacent half-grades are often combined in the elementary school in organizing groups for instructional purposes accounts for the small numbers of pupils found in a number of grade 8A "classes". In the junior high schools it was found that with a single exception the membership in all grade 8A homerooms consisted of pupils in only this one half-grade.

Table XXIV shows distributions of sizes of membership and sizes of groups tested for the 237 classes. In each of eight classes the 8A membership is fewer than 10 pupils. In seven classes the membership is 45 or more. For 191 of the classes the grade 8A membership ranges from 30 to 44. In summary, approximately three per cent of the classes enrol 45 to 54 grade 8A pupils, 81 per cent enrol from 30 to 44, 13 per cent enrol from 11 to 29 and three per cent enrol from 1 to 9 pupils.[4]

---

[4] Memberships in the largest and smallest classes, are, respectively, 50 and 3.

The average number of pupils per class is 34.3.

TABLE XXIV

Sizes of Grade 8A Class Groups

Size of Grade 8A Group (No. of Pupils)	Number of Grade 8A Class Groups	
	Enroled	Tested
50-54	1	..
45-49	6	5
40-44	69	36
35-39	79	83
30-34	43	60
25-29	5	18
20-24	10	4
15-19	7	12
10-14	9	10
5-9	7	8
1-4	1	1
Total Number of Groups	237	237
Total Number of Pupils	8139	7724
Mean No. Pupils Per Group	34.3	32.6

The sizes of class groups tested are shown in the right-hand column of Table XXIV. This distribution is roughly comparable to the distribution of class memberships. For example, 45 or more were tested in each of five groups. Similarly, in each of nine groups, fewer than 10 pupils were tested. For approximately three-fourths of the classes the number tested varied from 30 to 44. The average number tested per class was 32.6 as compared with the average class membership of 34.3.

Mean Test Scores for the 237 Class Groups. Pupils in the class showing the highest achievement on the test had an average score of 80.67. Those in the class showing the lowest achievement made an average score of 53.53. This difference of 27 points between the means of the highest group and the lowest group is approximately three times as large as the standard deviation of individual scores (9.05).[5] The distribution of average scores for the 237 classes is given in Table XXV.

TABLE XXV

Average Scores on Reading Test for Class Groups

Average Score	No. of Classes
80-82	1
78-79	..
76-77	6
74-75	8
72-73	23
70-71	27
68-69	24
66-67	33
64-65	37
62-63	29
60-61	24
58-59	12
56-57	7
54-55	5
52-53	1
Total	237
Mean of Class Averages	66.40*
Standard Deviation of Class Averages	5.22*

*Computed from ungrouped data given in Table XLV in the Appendix.

[5] See Table I, p. 24.

The mean of the class averages is shown in the table to be 66.40. This is lower than the average of the 7724 individual scores (66.74) by .34. The standard deviation of class means, i.e., the standard deviation of the actual distribution of the 237 class averages shown in Table XXV, is 5.22 which is more than half as large as the standard deviation of individual pupil scores (9.05).

When the school was used as the sampling unit, the assumption was made that the <u>number tested</u> in a given school constituted the total population in that school. A similar assumption will be made in applying sampling procedures to the 237 class groups. The total population in each class is defined as the pupils actually tested. It has been pointed out earlier that the only pupils for whom no data were secured were those absent on the testing date.

With all data summarized separately for each class the same four sampling designs applied to schools were also applied to classes. These four designs are:

1. Simple random sampling of classes
2. Stratification of classes by size of enrolment--random sampling of classes within each stratum
3. Stratification of classes by geographic location--random sampling of classes within each stratum
4. Primary stratification of classes by geographic location; substratification by size of enrolment--random sampling of classes within each sub-stratum

The next section of this chapter describes the results secured from the first of the four designs.

II. RESULTS FOR SIMPLE RANDOM SAMPLING OF CLASSES

Method Used in Drawing a Sample. Samples of classes were chosen by the same method used in drawing random samples of schools. The first step was to assign each class a three-digit code number. The code numbers used were 001 through 237. This made it possible to use a table of random numbers for selecting specific classes.

Mean Scores for Samples of 80 Classes. A group of 80 classes (about one-third of the total population) was chosen by using the random numbers table in exactly the same way it was used for schools except that in this instance three-digit rather than two-digit numbers were taken from the table. The sample of 80 classes was specified by taking from the table the first 80 unlike three-digit numbers.

As in the case of sampling by schools the first step in finding the mean pupil score for the entire sample was to compute the sum of the "sums of scores" and the sum of the numbers tested for all sampling units drawn. The data in Table XXVI illustrate this computation for the first sample of 80 classes. The summary at the foot of the table shows the mean score for the sample to be 67.09. This is higher than the average score for the total pupil population by .35.

A distribution of means comparable to the one shown at the foot of Table XXVI was produced by drawing nine additional samples, each consisting of 80 classes. Although the classes in each separate sample were drawn without replacement, the 80 codes were all "replaced" prior to drawing the succeeding sample.

TABLE XXVI

Numbers of Pupils Tested and Sums of Scores for a Sample of Classes--Illustration of Computation of Mean Score for One Random Sample of 80 Classes

Class	No. Pupils Tested	Sum of Scores
002	32	2305
005	46	2985
008	39	2373
010	41	2739
...	...	....
...	...	....
...	...	....
...	...	....
231	34	2337
233	39	2658
236	31	1856
Total	2549	171,013
Mean Score for Entire Sample		67.09

With this relatively large number of cases (237) in the population the mechanics of selecting 80 cases would have been made much simpler by using systematic rather than strictly random selection. Since the class codes were arranged in numerical order running from 001 through 237, the first sample might have been designated as every third class beginning either with class 001, 002 or 003--the selection of one of these three starting points being determined from a table of random numbers. There are two reasons why such a plan of systematic selection was not used. In the first place, it was thought desirable to employ a method of selecting classes identical with that used in selecting schools. In the second place, the simplest plan for drawing systematic samples of approximately one-third of the total number of classes provides for getting only three such samples--the first sample might begin with class 001, and the second with 002 and the third with 003. A fourth systematic sample of classes beginning with 004 would be a duplicate of the first sample. Since 10 separate samples were desired, it seemed advisable to continue to use the device of random numbers for the selection of each separate class.

Table XXVII summarizes the results for 10 random samples of 80 classes. It will be noted that Sample 1 in this table has already been described in Table XXVI. The means for the

TABLE XXVII

Numbers Tested, Sums of Scores and Mean Scores for 10 Samples Consisting of 80 Classes Each

Sample	No. of Classes	No. Pupils Tested	Sum of Scores	Mean Score
1	80	2549	171,013	67.09
2	80	2527	167,857	66.43
3	80	2493	166,044	66.60
4	80	2547	171,030	67.15
5	80	2500	168,430	67.37
6	80	2472	163,836	66.28
7	80	2698	178,392	66.12
8	80	2493	167,779	67.30
9	80	2547	170,133	66.80
10	80	2657	176,389	66.39
Average of Sample Means				66.75
Standard Deviation of Sample Means				.43

10 samples range from 66.12 to 67.37. This is somewhat less than the range in mean scores found from repeated random sampling of 50 schools. The absolute range of means for ten 50-school samples was greater than that found for ten 80-class

samples even though the number of pupils per sample is considerably less for the class samples than for the school samples.

The standard deviation of the 10 averages in Table XXVII is .43. The estimated standard error for an unlimited number of means for samples of this size drawn by the same method is .45 (.43 multiplied by $\sqrt{\frac{10}{9}}$). This estimated error for a sample which includes about one-third of the pupils in the population is considerably smaller than the corresponding computed error for a sample consisting of approximately one-half the pupils in the population when the sample is secured by using the school rather than the class as the sampling unit.

Application of Error Formula to Data for the 237 Classes. The cluster formula[6] for computing the standard error of the mean may be used in the same way it was used with schools. Now, however, a class rather than a school constitutes the "cluster". By substituting data for the 237 classes in the formula, the error for any sample size may be computed. It will thus be possible to make direct comparison between the respective efficiencies for school samples and class samples, each of which contain approximately the same number of pupils.

The computations involved in substituting class data in the formula are illustrated in Table XXVIII. By comparing the data presented here with those in Table VII it will be seen that the indicated computations for classes 001, 106 and 007, shown in Table XXVIII are identical with the

---

[6] See page 44.

computations for schools 01, 02 and 03 in Table VII.

TABLE XXVIII

Squared, Weighted Deviations of Class Means from the Population Mean--Illustration of Computation for Five Classes and Sum for 237 Classes

Class	No. Pupils Tested (N_i)	Sum of Pupil Scores (Σx_i)	Weighted Deviation ($\Sigma x_i - N_i\bar{x}$)	Squared Weighted Deviation $(\Sigma x_i - N_i\bar{x}_i)^2$
001	37	2,381	- 88,20388	7,779.924
...	..	.....	....	....
...	..	.....	....	....
106	39	2,746	+143.32564	20,542.239
...	..	.....	....	....
...	..	.....	....	....
007	27	1,619	-182.85148	33,434.664
...	..	.....	....	....
...	..	.....	....	....
236	31	1,856	-212.79244	45,280.623
237	32	2,177	+ 41.47232	1,720.063
Sum	7724	515,463	+ .00624	7,186,060.810
Mean	32.59072*	66.73524**	+ .00003*	

*Computed by dividing the corresponding sum by 237.
**Computed by dividing the corresponding sum by 7724.

The reason for this identity is that schools 01, 02 and 03 contain only one class each. Classes 001, 106 and 007, therefore, are merely alternative designations of the same groups of pupils represented by the three school codes. New computations had to be made, however, in every case where there was more than one class group of 8A pupils in a school. The detailed steps in the computations shown in Table XXVIII are described in full on pages 44-47. The only difference in meaning of the notation of the formula lies in the fact that here the formula gives the standard error of the mean pupil score for a sample of m <u>classes</u> drawn from a finite population consisting of M <u>classes</u>.

By substituting the quantities at the foot of Table XXVIII, the error for a random sample of m classes may be expressed as

$$\sigma^2_{\bar{x}}{}' = \left[\frac{237 - m}{(236)(m)}\right] \left[\frac{7,186,060.810}{(32.59072)^2(237)}\right]$$

When the sample size is 80 classes, the standard error of the mean pupils score is therefore

$$\sigma^2_{\bar{x}}{}' = \left[\frac{237 - 80}{(236)(80)}\right] \left[\frac{7,186,060.810}{(1,062.155)(237)}\right]$$

$$= .2374$$

$$\text{and} \quad \sigma_{\bar{x}}{}' = .487$$

This computed error (.49) for a random sample of 80 classes is somewhat larger than the estimated value (.45) derived from repeated sampling.

The respective errors for a sample of 20 classes, which represents a little less than 10 per cent of all the classes, and for a sample of 120 classes, approximately 50 per cent of the population, will now be computed. Substituting m = 20 in the formula on page 144 gives

$$\sigma^2_{\bar{x}}{}' = \left[\frac{237 - 20}{(236)(20)}\right] \left[\frac{7,186,060.810}{(1,062.155)(237)}\right]$$

$$= 1.3124$$

and $\sigma_{\bar{x}}' = 1.146$

The error for a 20-class sample is thus a little more than twice as large as the error for an 80-class sample. By comparing this result with the corresponding result for a sample of approximately 10 per cent of the <u>schools</u> it is seen that sampling by classes gives more accurate results. The error for a sample of 10 out of 97 schools is 1.65 whereas the error for 20 out of 237 classes is 1.15.

The error for a 50 per cent sample of classes, i.e., 120 classes, is

$$\sigma^2_{\bar{x}}{}' = .1179$$

and $\sigma_{\bar{x}}' = .343$

A 50 per cent random sample of classes thus gives a standard error (.34) which is considerably less than one-twentieth of the standard deviation of individual pupil scores which was designated earlier as the arbitrary level of efficiency desired.

<u>Summary of Results for Random Sampling of Classes</u>. Repeated sampling of class groups, taking about one-third of the total number of classes in each sample, gave a distribution of means with a standard deviation of .43. The

estimated standard error based on this actual distribution of averages is .45. The computed standard error for samples of this size was shown by the formula to be .49. Both the estimated and computed errors for a sample of <u>one-third</u> of the classes are smaller than either the estimated or computed errors determined earlier for a sample of <u>one-half</u> of the schools. The computed errors for 20-and 120-class samples respectively are 1.15 and .34. These two values are considerably smaller than the values derived for school samples containing approximately the same numbers of pupils.

For a given sample size (number of pupils), sampling by classes is thus found to give much more accurate results than sampling by schools when the samples are chosen by unrestricted selection. The relative superiority of the class over the school depends of course on the fact that many of the schools contain more than one class. Wherever this occurs, the use of the class as the sampling unit has the effect of reducing the size of the clusters. If practically all of the schools in a designated population contained only one class each, there would of course be no gain from sampling by classes. The "school unit" and the "class unit" would be identical.

III. RESULTS FOR STRATIFIED SAMPLING OF CLASSES

Stratification by Size of Grade 8A Membership

It was not expected that stratification of classes by size of enrolment would give an increase in efficiency over random selection as large as the gains achieved by this stratification control in the case of sampling by schools. In an earlier section of this chapter it was shown that the size of enrolment in more than three-fourths of the classes was from 30 to 44 pupils. The range in enrolments among the entire 237 classes was from three to 50. By comparison, the range in sizes of grade 8A enrolment for the 97 schools was from three pupils to 459 pupils. The differences among the average class sizes for separate enrolment strata will therefore be much less than the differences among the average school sizes for the several enrolment strata described in Chapter III where the school was used as the sampling unit.

Method Used in Stratifying Classes by Size of Membership. The classes were first listed in order of size of membership. Class 048 was at the top of the list with 50 pupils enroled and Class 066 was at the bottom of the list with three pupils enroled. It did not seem appropriate to set up a given number of strata each containing the same number of classes as was done in the case of schools. This method was thought inadvisable because of the large concentration of classes within a relatively small portion of the distribution of memberships. For example, there are 168 classes whose respective memberships vary from 33 to 43, a range of only 10 pupils. On the other hand, by starting with the smallest class and going up the list, it is necessary to include only 14 classes in order to get a group having a difference in enrolment of 10 pupils between the largest and the smallest class. After trying out several alternative patterns of enrolment "dividing lines" for separating the classes into different strata, it was finally decided to use seven strata. Four of these strata contain 12 classes each, one contains 24 classes, one contains 82 classes, and one contains 83 classes. The enrolment range within each stratum is as follows: Stratum I, 3-10; Stratum II, 10-19; Stratum III, 20-26; Stratum IV, 27-33; Stratum V, 33-39; Stratum VI, 39-43; Stratum VII, 43-50.

The reason for organizing the strata so as to include 12 classes, 24 classes, 83 classes and 82 classes was to provide for drawing proportionate samples from each stratum. The 24-class stratum is exactly twice as large as the 12-class stratum. The 82-class and 83-class strata are approximately seven times as large as the 12-class stratum. It is thus possible to draw samples from this design representing roughly one-tenth, one-third and one-half of the total number of classes. These proportions are approximately the same as the proportions of schools drawn from the different designs described in the preceding chapter.

In order to achieve such a division of classes into groups it was necessary to make several minor compromises in applying the enrolment control factor. For example, the four largest classes in Stratum I each have a grade 8A membership of 10 pupils and the smallest class in Stratum II also has a membership of 10 pupils. A strict enrolment stratification would have required a break in size of enrolment between Stratum I and Stratum II. This break could not be made because of the prior requirement that these two strata contain the same number of classes. Similar compromises were made in designating the "dividing lines" between Strata IV and V, V and VI, and VII.

Method Used in Selecting a Stratified Sample of Classes. A class stratification code sheet was next set up in preparation for drawing samples. The 12 classes in Stratum I were assigned special two-digit codes running from 00 through 11. The 82 classes in Stratum VI were given two-digit codes running from 00 through 81, and so on for the other strata. The code sheet is illustrated in Table XXIX.

This code sheet was used in conjunction with the table of random numbers in selecting a sample of 80 classes. A sample of this size drawn proportionately from each of the seven strata consists of four classes from Stratum I, four classes from Stratum II, four classes from Stratum III, eight classes from Stratum IV, 28 classes from Stratum V, 28 classes from Stratum VI and four classes from Stratum VII. Using the table of random numbers the four codes selected from Stratum I were 01, 02, 03 and 11. It will be seen in Table XXIX that these codes represent respectively classes 049, 077, 217 and 009. The classes to be taken into the sample from each of the other six strata were chosen by the same method. The sum of the number of pupils tested for the 80-class sample is 2613. The sum of pupil scores is 174,729 and the mean pupil score is 66.87.

TABLE XXIX

Lists of Classes in the Seven Enrolment Strata and Special Stratum Code for Each Class

Stratum													
I		II		III		IV		V		VI		VII	
Strat. Code	Cl.	Strat. Code	Cl.	Strat. Code	Cl.	Strat. Code	Cl.	Strat. Code	Cl.	Strat. Code	Cl.	Strat. Code	Cl.
11	009	11	013	11	017	23	113	82	023	..	...	11	048
10	015	10	024	10	011	22	158	81	129	81	063	10	065
09	047	09	103	09	006	21	002	80	195	80	102	09	005
08	068	08	087	08	091	20	003	79	027	79	117	08	098
07	058	07	051	07	115	19	020	78	053	78	143	07	046
06	093	06	062	06	101	18	028	77	079	77	223	06	076
05	105	05	109	05	090	17	064	76	147	76	225	05	124
04	116	04	037	04	004	16	078	75	149	75	022	04	120
03	217	03	083	03	019	15	089	74	152	74	042	03	125
02	077	02	045	02	030	14	114	73	153	73	075	02	057
01	049	01	060	01	084	13	150	72	156	72	119	01	118
00	066	00	040	00	100	12	209	71	160	71	121	00	127
						11	219	70	182	70	122		
						10	007	69	196	69	123		
						09	161	68	199	68	126		
						08	026	67	200	67	155		
						07	086	66	205	66	170		
						06	094	65	234	65	186		
						05	107	64	014	64	213		
						04	112	63	035	63	221		
						03	210	62	052	62	222		
						02	085	61	054	61	226		
						01	163	60	056	60	010		
						00	018	59	059	59	016		
								58	074	58	055		
								*	*	*	*		
								..	...	..	...		
								..	...	..	...		
								02	208	02	177		
								01	232	01	190		
								00	235	00	233		

*Codes 03 through 57 are not included for Strata V and VI.

<u>Mean Scores for Samples of 80 Classes Stratified by Size of Grade 8A Membership</u>. A total of 10 stratified samples of 80 classes each was drawn by the method described. The data for these 10 samples, summarized in Table XXX, show that the largest mean is 67.11 and the smallest is 66.54. The standard deviation of the 10 averages (.21) and the estimated standard error (.22) are considerably smaller than the corresponding values found for 80-class samples drawn by unrestricted selection. The mean of the sample averages is 66.84 or .10 above the average for the pupil population. It is interesting to note in Table XXX that the number of pupils tested varies very little among the 10 samples. The smallest of the samples included 2600 pupils and the largest included 2642 pupils.

TABLE XXX

Numbers Tested, Sums of Scores and Means for Ten 80-Class Samples Stratified by Size of Enrolment

Sample	No. of Classes	No. Pupils Tested	Sum of Scores	Mean Score
1	80	2613	174,729	66.87
2	80	2612	175,280	67.11
3	80	2642	175,831	66.55
4	80	2628	176,107	67.01
5	80	2602	173,376	66.63
6	80	2629	174,940	66.54
7	80	2600	174,480	67.11
8	80	2617	175,201	66.95
9	80	2603	174,146	66.90
10	80	2617	174,532	66.69
Average of Sample Means				66.84
Standard Deviation of Sample Means				.21

<u>Application of Error Formula for Stratified Clusters</u>. The formula appropriate for use with stratified samples was described on pages 72-78 along with illustrations of the detailed

computational steps involved. This formula was applied to the data for all <u>classes</u> in each of the seven enrolment strata in the same way it was applied to all of the <u>schools</u> in each of the 10 enrolment strata. It will be recalled that computations need to be made separately for each stratum and the sum of the "stratum computations" represents the standard error squared for the entire sample. The computed error for an 80-class sample stratified by enrolment is

$$\sigma^2_{\bar{x}}{}' = .0013 + .0016 + .0014 + .0149 + .0747 + .1110 + .0114 = .2163$$

and $$\sigma_{\bar{x}}{}' = .466$$

This error (.47), derived from the formula, is considerably larger than the estimated error (.22), based on data secured from repeated sampling. The significance of the divergence of the estimated error from the computed error was determined by applying the X^2 test described on page 79. The obtained X^2 is significant at the two per cent level but not at the one per cent level. It appears, therefore, that the variance of the 10 means shown in Table XXX is not typical of the variances that would be expected for other groups of 10 sample means that might be secured by the same method.

A sample of <u>20</u> classes stratified by enrolment consists of one class each from Strata I, II, III and VII, two classes from Stratum IV and seven classes each from Strata V and VI.

The computed standard error for a sample of this size is found to be

$$\sigma^2_{\bar{x}}{}' = .0069 + .0089 + .0079 + .1033 + .4131 + .6165 + .0625 = 1.2191$$

$$\text{and} \quad \sigma_{\bar{x}}{}' = 1.104$$

It is interesting to note in passing that this standard error (1.10) for a stratified sample of 20 classes is exactly the same as the computed error for an unrestricted sample of 20 schools[7] even though there are more than twice as many pupils in a group of 20 schools as in a group of 20 classes.

A sample of approximately one-half the total number of classes in the population may be chosen by taking six classes from each of the first three enrolment strata, 12 classes from Stratum IV, 42 classes from Stratum V and from Stratum VI and six classes from Stratum VII. The computed error for a stratified sample of this size is

$$\sigma^2_{\bar{x}}{}' = .00063 + .00081 + .00072 + .00939 + .03713 + .05481 + .00568 = .10917$$

$$\text{and} \quad \sigma_{\bar{x}}{}' = .330$$

The computed error (.33) for a stratified sample of one-half the classes in the population is only slightly smaller than

---

[7] Page 48.

the computed error (.34) for an unrestricted sample of the same number of classes.

Estimation of Error by the Use of Data from a Single Sample of Classes Stratified by Size of Grade 8A Membership. If a sample were to be drawn in practice for the purpose of determining a test norm, a stratification design could be set up in advance exactly like the one described here. A single sample would then be drawn, the mean of the sample would be computed and the error of the sample mean would have to be estimated from the data of the sample. No other data from the population would be available. The procedure for estimating the error of the mean for a single sample of 80 classes was applied to the data for Sample 6 shown in Table XXX. This particular sample was chosen because the mean score is the smallest among the 10 samples drawn. The formula to be used is exactly the same as the one employed in deriving an estimate of error for a single sample of schools stratified by enrolment.[8] All formula notations which formerly referred to a school unit now of course designate a class unit.

The indicated computations were performed for the four classes in the sample which had been drawn from Stratum I, the four classes drawn from Stratum II, etc. The results are as follows:

---

[8] See page 85.

$$s^{2\prime}_{\bar{x}} = .00079 + .00263 + .00168 + .03531 + .07509 + .07238 + .00673 = .1946$$

and $$s_{\bar{x}}' = .441$$

This value (.44) is an estimate of the standard error of the mean (66.54) of Sample 6 based on analyses of the data from Sample 6 only. This estimated error is slightly smaller than the error (.47) for stratified samples of 80 classes as computed from the data from the entire population of classes from each stratum.

A second estimate of the error for a sample of 80 classes was made from the data of Sample 7 which is also summarized in Table XXX. This particular sample was chosen because its mean score is the <u>largest</u> among the 10 means secured by repeated sampling. The data from the classes in this sample substituted in the formula in the same way as the data for Sample 6 give an estimated error of .48 which is somewhat larger than the error (.46) computed from the data for all of the schools in each stratum. One of the two estimates of error derived from a single sample is thus slightly smaller than the "true" error and the other is slightly larger.

<u>Summary of Results for Sampling by Enrolment Strata</u>. The errors (computed) for samples of classes stratified by enrolment were found to be slightly smaller than the errors for unrestricted samples of the same size. The respective errors for enrolment stratification and unrestricted selection are

as follows: for 20-class samples, 1.10 and 1.15; for 80-class samples, .46 and .49; and for 120-class samples, .33 and .34. These findings show that relatively little increase in accuracy was achieved by stratifying the classes according to size of membership.

IV. RESULTS FOR STRATIFIED SAMPLING OF CLASSES

Stratification by Geographic Location

Method Used in Stratifying Classes by Geographic Location. This third design to be applied to classes called for dividing the 237 classes into a number of different sub-groups according to geographic location. The first step was to "spot" each class on an outline map of the city in the same way the schools were spotted. The location of a given school of course represented the location of all the classes within that school. Where a school contained only one grade 8A class, one class symbol was placed at the appropriate point on the map. For schools that contained two, five or 10 different classes, the corresponding numbers of symbols were placed on the map, grouped around the points representing the locations of the respective schools. The completed map contained 237 classes symbols distributed over the entire city.

It was decided to use 10 geographic strata for classes, the same as for schools. A study of the map showed that it would be possible to set up 10 strata, each containing approximately the same number of classes, whose boundary lines would correspond roughly to the boundary lines used in dividing the schools into 10 geographic strata. The actual locations of these lines dividing the several groups of classes from each other were decided on more or less arbitrarily so as to have approximately the same numbers of classes in all 10 strata.

The 24 classes in the west and northwest section of the city were designated as Stratum I. The 23 classes east of Stratum I and bordering the city limits on the north were designated as Stratum II, and so on. The map on page 159 shows the 237 classes divided into 10 geographic groupings.

It will be seen from the map that Strata IV and VI each contain 25 classes, Strata I, III and VII each contain 24 classes and Strata II, V, VIII, IX and X contain 23 classes each. The large <u>schools</u> within each stratum may be located easily by noting the tight "clusters" of class symbols. For example, Stratum V near the center of the city contains two large schools whereas Stratum I in the west section contains no school with more than 3 classes of grade 8A pupils.

The next step in preparation for drawing a sample was to assign special code numbers to the classes within each stratum. The 24 classes in Stratum I were coded 01 through 24. The 23 classes in Stratum II were coded 01 through 23 and so on for the other 8 strata. Table XXXI shows the complete coding plan for Strata I through V. The lists of classes given in the table show that Strata I and III contain 24 classes, Strata II and V contain 23 classes and Stratum IV contains 25 classes.

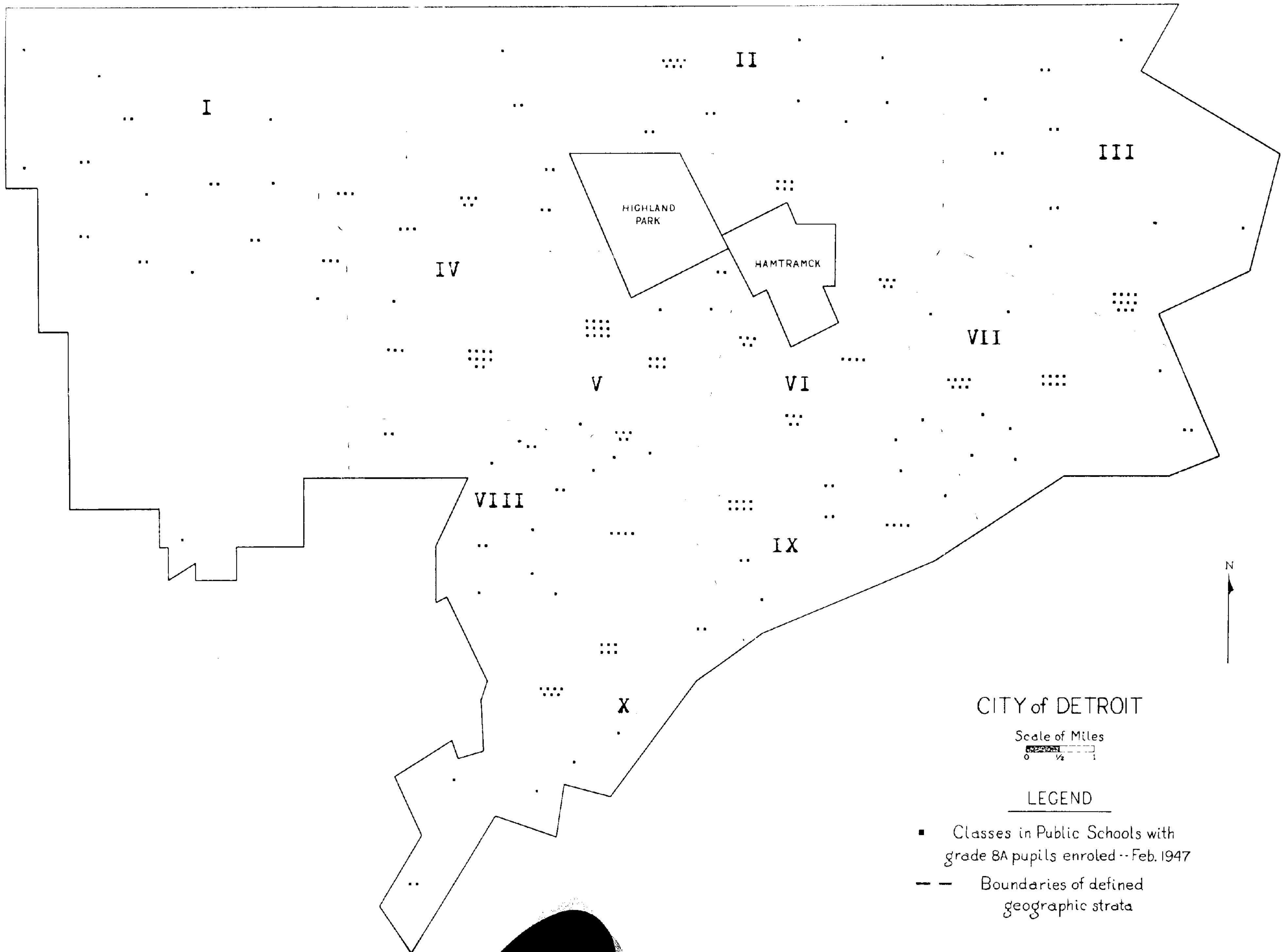

I
II
III
IV
V
VI
VII
VIII
IX
X
HIGHLAND PARK
HAMTRAMCK
N
CITY of DETROIT
Scale of Miles
0
½
1
LEGEND
Classes in Public Schools with grade 8A pupils enroled -- Feb. 1947
Boundaries of defined geographic strata

TABLE XXXI

List of Classes with Special Codes for Five of the 10 Geographic Strata*

Stratum I		Stratum II		Stratum III		Stratum IV		Stratum V	
Strat. Code	Class	Strat. Code	Class	Strat. Code	Class	Strat. Code	Class	Strat. Code	Class
..	..	..	..	..	..	25	230	..	..
24	066	..	..	24	184	24	229	..	..
23	065	23	210	23	183	23	228	23	173
22	064	22	209	22	182	22	227	22	172
21	107	21	208	21	181	21	226	21	171
20	061	20	207	20	180	20	225	20	170
19	062	19	206	19	179	19	224	19	169
18	059	18	205	18	178	18	223	18	168
17	060	17	204	17	177	17	222	17	150
16	057	16	019	16	176	16	221	16	149
15	058	15	028	15	175	15	215	15	148
14	055	14	018	14	174	14	214	14	147
13	056	13	026	13	036	13	213	13	146
12	050	12	017	12	035	12	212	12	145
11	051	11	016	11	034	11	211	11	144
10	049	10	015	10	033	10	112	10	143
09	048	09	014	09	032	09	111	09	142
08	046	08	013	08	031	08	110	08	141
07	047	07	012	07	030	07	063	07	140
06	042	06	023	06	029	06	054	06	139
05	041	05	011	05	027	05	053	05	004
04	038	04	010	04	024	04	052	04	003
03	039	03	009	03	022	03	045	03	002
02	040	02	008	02	021	02	044	02	106
01	037	01	007	01	020	01	043	01	001

*Lists of classes and codes for Strata VI through X are not included in the table.

The sampling rate will not be exactly the same for all strata when a given number of classes is drawn from each one. The differences in rate from stratum to stratum will be very small, however, since the largest and the smallest of the 10 strata differ in size by only two classes. In the analyses to be made for samples drawn from this design it is assumed that the sampling rate over all strata is identical.

Mean Scores for 80-Class Samples Stratified by Geographic Location. A sample of 80 classes was drawn by taking eight classes from each of the 10 strata in turn. The first eight two-digit numbers from the random numbers table were 21, 15, 12, 24, 13, 17, 16 and 14. The classes bearing these codes were the ones chosen from Stratum I. Eight classes drawn by the same method from each of the remaining nine strata provided the total sample. The number of pupils tested for this sample is 2551. The sum of scores is 171,833 and the mean score is 67.36.

Nine additional samples of 80 schools each were selected by the same method. The results given in Table XXXII show that the highest among the 10 sample means is the one for the first sample selected. Sample 5 has the lowest mean (66.51). The average of the 10 means (66.84) is higher than the population mean (66.74) by .10. The standard deviation of the 10 means is .31 which gives an estimated standard error of .33 for an unlimited number of 80-class samples drawn by the same method.

TABLE XXXII

Numbers Tested, Sums of Scores and Mean Scores for Ten 80-Class Samples Stratified by Geographic Location

Sample	No. of Classes	No. Pupils Tested	Sum of Scores	Mean Score
1	80	2551	171,833	67.36
2	80	2575	173,258	67.28
3	80	2629	175,417	66.72
4	80	2719	181,192	66.64
5	80	2529	168,200	66.51
6	80	2601	173,445	66.68
7	80	2637	175,418	66.52
8	80	2613	175,426	67.14
9	80	2730	182,964	67.02
10	80	2486	165,485	66.57
Average of Sample Means				66.84
Standard Deviation of Sample Means				.31

The estimated error for samples of this size stratified by geographic location is thus found to be somewhat larger than the estimated error (.22) for samples of the same size stratified by size of enrolment. The estimated error for the geographic design, however, is smaller than the computed error (.47) for enrolment stratification.

Estimate of Error from the Data of a Single Sample. Detailed data from the two samples designated in Table XXXII as Samples 6 and 7 were substituted in the formula in order to derive two independent estimates. The estimate for Sample 6 is:

$$s^2_{\bar{x}}{}' = .01277 + .02273 + .01544 + .02479 + .03840$$
$$+ .00465 + .01887 + .01079 + .01477 + .00636$$
$$= .16927$$

$$s_{\bar{x}}{}' = .411$$

Similar data from Sample 7 give

$$s^2_{\bar{x}}{}' = .00361 + .00785 + .01535 + .01905 + .04215$$
$$+ .01324 + .00108 + .01803 + .01066 + .01336$$
$$= .14438$$

and $s_{\bar{x}}{}' = .380$

Both of these errors (.41 and .38) which represent estimates derived from single samples are larger than the estimate (.33) based on the variance of the 10 sample means shown in Table XXXII.

<u>Application of Error Formula to the Entire Population.</u> The "true" error for samples of 80 classes stratified by geographic location was next computed by substituting data for all classes in each of the 10 strata in the formula which requires complete information from every element in the population. The results are as follows:

$$\sigma^2_{\bar{x}}{}' = .0101 + .0167 + .0124 + .0222 + .0301$$
$$+ .0180 + .0121 + .0153 + .0204 + .0095$$
$$= .1668$$

and $\sigma_{\bar{x}}{}' = .409$

This computed value (.41) based on complete population data is exactly the same as the estimate derived from one of the single samples. It is somewhat larger than the estimate from the second single sample and it is larger by .08 than the estimate (.33) derived from the 10 means of 80-class samples secured by repeated sampling.

A 20-class sample drawn from this design would consist of two classes from each of the 10 strata. The computed error for a sample of this size is .96. The computed error for a 120-class sample (12 classes from each stratum) is .29.

Summary of Results for Sampling by Geographic Strata. Three independent estimates of error were made for samples of 80 classes stratified by geographic location. These three estimates were .33, .41 and .38. Data for the entire population were then substituted in the formula to get .41 as the "true" error for samples of 80 classes. By similar substitution the errors for a 20-class sample and for a 120-class sample were found to be .96 and .29 respectively. These results show that for any sample size the error for geographic stratification is somewhat less than for either unrestricted selection or enrolment stratification.

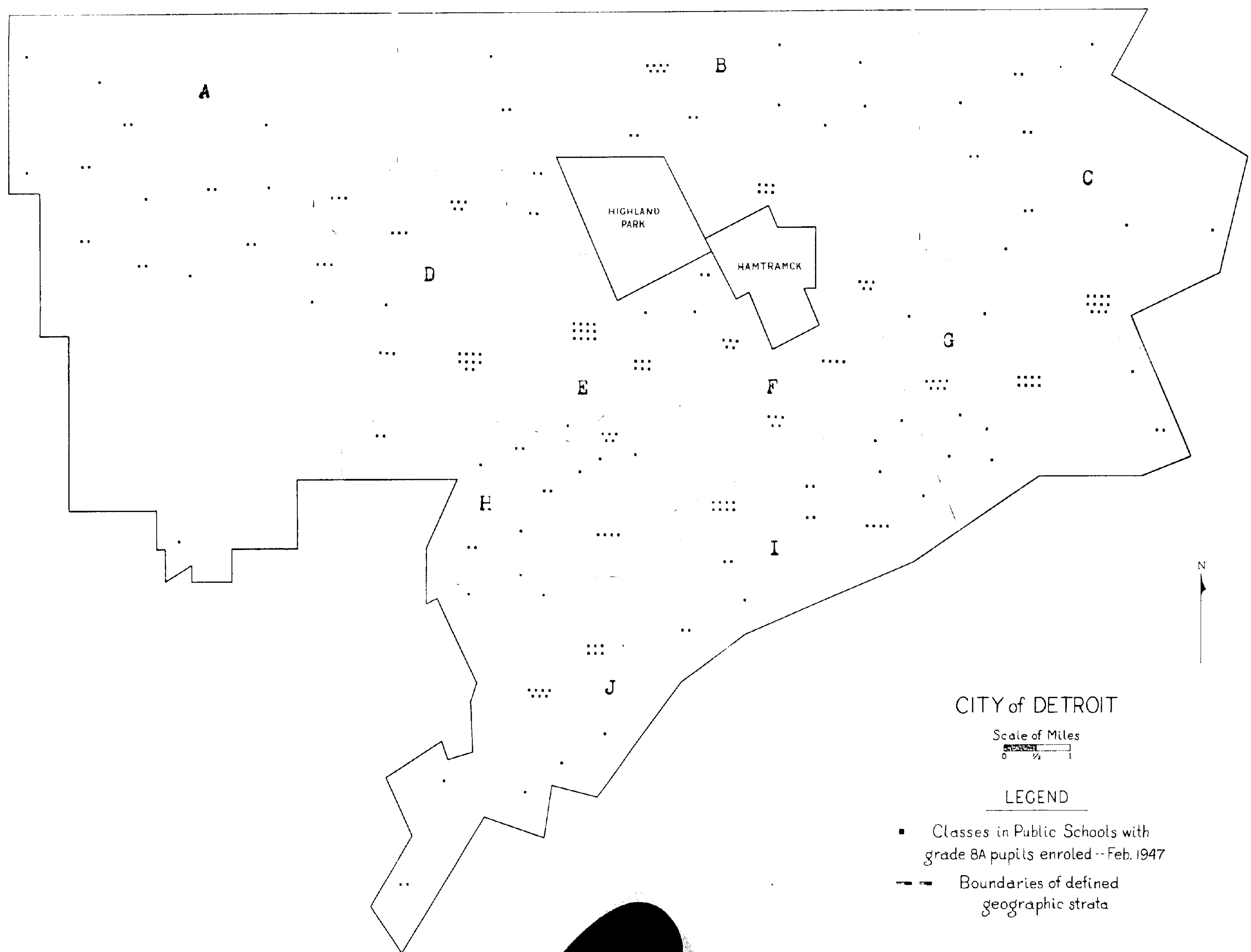
V. RESULTS FOR STRATIFIED SAMPLING OF CLASSES

Primary Stratification by Geographic Location Sub-Stratification by Size of Enrolment

In the preceding chapter it was found that the dual stratification control, geographic location and size of enrolment, when applied to schools, was no more effective in reducing sampling error than either one of the two controls used separately. In the case of schools it was concluded that this relative ineffectiveness was due to the fact that much larger primary groupings of schools were used when both controls were applied simultaneously than when either geographic or enrolment stratification was used alone. Much of the gain achieved by geographic control alone was apparently lost when the size of geographic groupings had to be increased. On the basis of this conclusion it was decided to apply the dual control to classes by using as primary strata the geographic groupings of classes already defined. This stratification pattern is known to be more efficient than unrestricted selection. Sub-stratification by size of enrolment could then be made within each defined geographic grouping. Such a plan seemed feasible to apply to classes for two reasons: first, each geographic group of classes contains more than 20 units whereas in the case of schools no geographic group contains more than 10 units; second, the variation in size among the classes is very much less than among the schools.

Method Used in Stratifying the 237 Classes by Geographic Location and by Size of Enrolment. The map on the following page shows boundary lines dividing the 237 classes into 10 groups. These "area" groupings, designated as A, B, C, etc., represent the primary geographic control. By comparing this map with the one on page 159 it will be seen that Area A (page 166) includes the same 24 classes designed as Stratum I on page 159. Similarly, Area B (page 166) includes the 23 schools designated as Stratum II on page 159, etc.

The 24 classes in "Area" were next divided into two sub-groups according to size of enrolment. The 12 smallest classes were labeled Sub-Group I and the 12 largest classes were labeled Sub-Group II. Similarly, the 23 classes in


A
B
C
D
E
F
G
H
I
J
HIGHLAND PARK
HAMTRAMCK
N
CITY of DETROIT
Scale of Miles
0 ½ 1
LEGEND
Classes in Public Schools with grade 8A pupils enroled -- Feb. 1947
Boundaries of defined geographic strata

"Area B" were divided into enrolment sub-groups, the 12 smallest classes being designated as Sub-Group I and the 11 largest classes as Sub-Group II. The classes in each of the eight other geographic areas were divided into two sub-groups by the same method to give a total of 20 different sub-groups or "strata". Each of these 20 strata could now be identified by a double symbol. For example, Stratum A-I consists of the 12 smallest classes in geographic Area A.

Special "stratum codes" were next assigned to the classes constituting each of the 20 strata. The complete coding plan for 10 of the strata, A-I through E-II is shown in Table XXXIII. It will be seen in the table that the number of class units is not the same for all strata. For example, Stratum A-II contains 12 classes; Stratum B-II contains 11 classes; and Stratum D-I contains 13 classes. Therefore when the same number of classes is drawn from each stratum the chances of being selected are slightly greater for a given class in Stratum B-II than for a given class in Stratum A-I or D-II. However the difference in sampling rate from stratum to stratum will be relatively small and it will be assumed that the same number of classes chosen from each stratum gives proportionate representation of each one.

Mean Scores for Samples of 20 Classes and 40 Classes. A sample of 20 classes was drawn from the design illustrated in Table XXXIII by taking one class from each of the 20 strata.

TABLE XXXIII

Lists of Classes and Special Codes for 10 of the 20 Geographic-Enrolment Strata*

Enrol-ment Sub-Group	Geographic Area									
	A		B		C		D		E	
	Strat. Code	Cl.	Strat. Code	Cl.	Strat. Code	Cl.	Strat. Code	Cl.	Strat. Code	Cl.
	12	048	..	..	12	022	12	225	..	..
	11	065	11	016	11	184	11	223	11	143
	10	046	10	010	10	180	10	063	10	170
	09	057	09	008	09	179	09	226	09	168
	08	042	08	023	08	178	08	222	08	146
II	07	055	07	205	07	177	07	221	07	145
	06	050	06	014	06	174	06	213	06	144
	05	039	05	206	05	029	05	224	05	106
	04	061	04	204	04	021	04	214	04	171
	03	041	03	012	03	182	03	212	03	140
	02	038	02	207	02	027	02	227	02	139
	01	056	01	208	01	176	01	211	01	001
	..	...	..	...	..	...	13	228	..	...
	12	059	12	209	12	181	12	215	12	168
	11	066	11	028	11	035	11	110	11	149
	10	107	10	007	10	183	10	043	10	147
	09	062	09	210	09	033	09	044	09	173
	08	051	08	026	08	175	08	053	08	148
I	07	037	07	018	07	034	07	054	07	172
	06	060	06	017	06	032	06	052	06	142
	05	047	05	011	05	031	05	111	05	141
	04	040	04	019	04	036	04	230	04	150
	03	058	03	013	03	020	03	229	03	003
	02	049	02	015	02	030	02	112	02	002
	01	064	01	009	01	024	01	045	01	004

*Strata F-I through J-II are not included in the table.

The table of random numbers was used in the manner described earlier. The first 20 two-digit numbers between 01 and 13 were 10, 10, 11, 01, and so on. These code numbers represent Stratum A-II: Class 046; Stratum B-II: Class 010; Stratum C-II: Class 184; etc. The number of pupils tested in the 20-class sample is 648, the sum of the scores is 42,802 and the mean score is 66.05. The results for this sample and for four additional samples drawn by the same method are given in Table XXXIV. The average of the five sample means is 66.70 which is lower than the population mean (66.74) by .04. The mean for Sample 5 is highest (67.14) and the mean for Sample 1 is lowest (66.05). The standard deviation of the five means is .39. With N = 5, the estimated standard error is .44. This estimate is of course relatively unreliable because of the small number of cases on which it is based. These findings do suggest, however, that the dual stratification design may give somewhat more accurate results than either enrolment stratification or geographic stratification used alone.

TABLE XXXIV

Numbers Tested, Sums of Scores and Mean Scores for Five 20-Class Samples Stratified by Geographic Location and by Size of Enrolment

Sample	No. of Classes	No. Pupils Tested	Sum of Scores	Mean Score
1	20	648	42,802	66.05
2	20	681	45,520	66.84
3	20	595	39,544	66.46
4	20	591	39,591	66.99
5	20	692	46,462	67.14
Average of Sample Means				66.70
Standard Deviation of Sample Means				.39

Five samples consisting of 40 classes each were drawn from the geographic-enrolment design by taking two classes instead of one from each of the 20 strata. The results are shown in Table XXXV. The average of the five means (66.48) is lower than the true mean of the population (66.74) by .26. It will be recalled that the average of the means for the five 20-class samples shown in the preceding table differed from the

TABLE XXXV

Numbers Tested, Sums of Scores and Mean Scores for Five 40-Class Samples Stratified by Geographic Location and by Size of Enrolment

Sample	No. of Classes	No. Pupils Tested	Sum of Scores	Mean Score
1	40	1284	84,666	65.94
2	40	1343	87,698	65.30
3	40	1270	84,742	66.73
4	40	1321	89,183	67.51
5	40	1320	88,306	66.90
Average of Sample Means				66.48
Standard Deviation of Sample Means				.77

population mean by .04. The variance of the means for the five 40-class samples is found to be greater than the corresponding variance for the five 20-class samples. Using the data from Table XXXV the estimated standard error for a 40-class sample is .85. The corresponding estimate of error for a 20-class sample was .44. The results secured from these two groups of samples offer an interesting illustration of the "unreliability" of the inferences that may be drawn from a small sample. It is almost self-evident that the mean of a 40-class sample is more reliable than the mean of a 20-class sample even though the data in Tables XXXIV and XXXV would suggest the opposite conclusion.

Mean Scores for Samples of 80 Classes. Ten 80-class samples were drawn successively from the geographic-enrolment design so that direct comparisons could be made with the results for samples of the same size secured from each of the other three "class" designs. Four classes chosen at random from each of the 20 strata constitute a single sample. Results given in Table XXXVI show the average of the 10 sample means to be 66.93. This is higher than the population mean

TABLE XXVI

Numbers Tested, Sums of Scores and Mean Scores for Ten 80-Class Samples Stratified by Geographic Location and by Size of Enrolment

Sample	No. of Classes	No. Pupils Tested	Sum of Scores	Mean Score
1	80	2518	170,862	67.86
2	80	2689	179,989	66.94
3	80	2661	178,506	67.08
4	80	2601	174,679	67.16
5	80	2588	173,214	66.93
6	80	2630	174,607	66.39
7	80	2664	176,003	66.07
8	80	2683	179,219	66.80
9	80	2700	180,352	66.80
10	80	2650	178,252	67.26
Average of Sample Means				66.93
Standard Deviation of Sample Means				.46

(66.74) by .19. The standard deviation of the 10 means is .46 and the estimated standard error for a single sample based on the data from these 10 cases is .48. This value (.48) for 80-class samples is larger than the estimate of error (.44) for 20-class samples drawn from this design. The group of means for the five 20-class samples shown in Table XXXIV thus appears to have a variance much smaller than would be found for most groups of 20-class samples that might be secured by the same method.

These results indicate that highly reliable estimates of error can be obtained from repeated sampling only when the number of samples actually drawn is very large. In order to detect small differences in accuracy for different designs used with a given population it might be necessary to draw several hundred different samples for each of the designs under consideration. Use of the appropriate formula will of course give more precise determinations of sampling error in situations such as this one where all data are already available in advance.

Error for Geographic-Enrolment Stratification of Classes Computed by Formula. Data for the 12 classes in Stratum B-I and so on through the 11 classes in Stratum J-II were substituted in turn in the error formula for stratified clusters.[9] The computations and substitutions for each of the 20 strata give the following results for a sample of 20

---

[9] The formula is given on page 103.

classes, i.e., one class from each stratum:

$$\sigma^{2}_{\bar{x}}{}' = .0085 + .0128 + .0366 + .0706 + .0850$$
$$+ .0194 + .0231 + .0622 + .0212 + .0297$$
$$+ .0426 + .0717 + .0273 + .0473 + .0679$$
$$+ .0411 + .0339 + .0287 + .0936 + .0251$$
$$= .8483$$

and $\sigma_{\bar{x}}{}' = .921$

This computed value (.92) for the standard error for a 20-class sample is more than twice as large as the estimate (.44) derived from the five sample means in Table XXXIV. The χ^2 test[10] shows that this divergence of the estimated variance from the computed variance is not significant at the ten per cent level. The group of five means obtained by repeated sampling is thus found to be less "unusual" than it was first thought to be.

In computing the error for a 40-class sample it was necessary to make only one change in the substitution which gave the error for a 20-class sample. This change consists of the use of m = 2 instead of m = 1 in the formula for each of the 20 strata. The result is:

$$\sigma^{2}_{\bar{x}}{}' = .0039 + .0058 + .0166 + .0324 + .0386$$
$$+ .0089 + .0105 + .0283 + .0096 + .0135$$
$$+ .0194 + .0323 + .0124 + .0215 + .0306$$
$$+ .0187 + .0154 + .0129 + .0421 + .0113$$
$$= .3847$$

and $\sigma_{\bar{x}}{}' = .620$

---

10 See page 79.

This computed value (.62) indicates that the corresponding error (.86) previously determined from the five 40-school samples in Table XXXV, was an over-estimate.

Since an 80-school sample consists of four classes from each stratum, the error for samples of this size was computed by substituting m = 4 in the formula. The sum of the 20 results secured from the respective strata is

$$\sigma^2_{\bar{x}}{}' = .1528$$
$$\sigma_{\bar{x}}{}' = .391$$

A comparison of this result (.39) with the estimate of $\sigma_{\bar{x}}{}'$ (.48) derived from the data in Table XXXVI shows the estimate to be somewhat high.

By substituting m = 6 in the formula for each of the 20 strata, the error for a 120-class sample was found to be

$$\sigma^2_{\bar{x}}{}' = .0755$$
$$\sigma_{\bar{x}}{}' = .275$$

Since samples of this size were not actually drawn, there is no estimated value to be compared with the computed error of .27.

It is interesting to note that the estimates of error determined by drawing repeated samples of 20 classes, 40 classes and 80 classes turned out to be misleading in two respects. First, the estimated standard error of the mean for a 20-class sample was actually smaller than the estimated error for an 80-class sample. Secondly, the estimated error for an 80-class sample, by comparison with both the computed values and the corresponding estimates for samples of this size drawn from the other three designs indicated that results secured from the dual control design were about equal in accuracy to those obtained from unrestricted selection and were somewhat less accurate than the results secured either from enrolment or from geographic stratification used alone. However, the <u>computed</u> errors for the dual control design show it to be more effective than any of the other three.

<u>Summary of Results for Sampling by Geographic-Enrolment Strata</u>. Repeated sampling was used to produce populations of means for three different sample sizes, 20 classes, 40 classes and 80 classes. The estimates of error derived from the three sets of data secured in this manner were .44, .86 and .48. There were obvious inconsistencies in these results since they show the error for a 20-class sample to be smaller than the error for a 40-class sample or an 80-class sample. It was concluded that the observed inconsistency was due to the relatively low degree of reliability that may be expected from a small number of cases. Only 5 samples of 20 classes, five samples of 40 classes and 10 samples of 80 classes were drawn.

Standard errors were determined precisely for <u>four</u> different sample sizes--20 classes, 40 classes, 80 classes and 120 classes--by substituting data for all classes in each of the 20 strata in the appropriate formula. The computed errors for these four sample sizes were found to be .92, .62, .39, and .27 respectively.

VI. COMPARISON OF RESULTS FOR FOUR DIFFERENT METHODS OF SAMPLING BY CLASSES

The four plans used in drawing samples of classes were the same as the four different plans used in drawing samples of schools. Three of the designs involved stratification of the classes by one or more control factors. Two different control factors were used. One was size of class membership and the other was geographic location of the school in which the class was found. Where a given school contained more than one class enroling grade 8A pupils, the "location" of all classes in that school was of course the same.

Review of the Four Designs. The first method to be applied was simple random sampling of the 237 classes. Samples were drawn without restriction, every one of the 237 classes having an equal chance to be taken into the sample on each drawing.

The essential feature of the second method to be applied was arrangement of the classes in seven different sub-groups, according to size of class membership, before any samples were drawn. The different sub-groups contained from 12 classes to 84 classes, each being relatively homogeneous with respect to size of membership of the classes contained in it. Samples were secured from this design by taking a proportionate number of classes from each of the sub-groups.

The third method to be applied consisted of arranging classes in 10 sub-groups according to geographic location prior to drawing samples. Each geographic sub-group contained approximately the same number of classes and a sample was secured by drawing a given number of classes at random from each sub-group in turn.

The fourth and last method to be applied involved the use of the geographic and the enrolment factors simultaneously in setting up sub-groups of classes from which samples were to be chosen. Within each of the 10 geographic groups used in the "third method", the classes were further divided into two sub-groups, approximately equal in size, on the basis of size of class membership. This gave a total of 20 different sub-groups, each containing about 12 classes. Samples were secured by selecting a given number of classes at random from each of the 20 sub-groups in turn.

Although certain restrictions were placed on the drawing of samples from each of the three designs employing stratification, every class had an equal chance of selection (with the minor exceptions noted earlier) in every sample chosen. Stratification thus eliminated the possibility of choosing certain combinations of classes in any one sample without affecting the equality of opportunity of selection for any given class.

Standard Errors for the Four Designs. The variability of sample means was determined for each of the four designs by two independent methods. The two methods were the same as those applied in determining the error of the mean for samples of schools. The first method consisted of securing an actual distribution of means by repeated sampling and using the data of this distribution to derive an estimate of the standard error. The second method of determining sampling error consisted of substituting data for the entire population in appropriate formulae.

Both the estimated errors derived from repeated sampling and the computed errors secured from the formulae are shown in Table XXXVII for the different designs. It has been pointed out in preceding sections that the estimates of error are relatively unreliable because of the small numbers of cases on which they are based. The computed errors given in

TABLE XXXVII

Estimated Errors and Computed Errors for Four Different Methods of Sampling by Classes

Sampling Design	Sample Size	Standard Error Estimated*	Standard Error Computed**
Unrestricted Selection	80 Classes	.45	.49
Enrolment Stratification	80 Classes	.22	.47
Geographic Stratification	80 Classes	.33	.41
Geo. - Enrol. Stratification	80 Classes	.48	.39

*Estimates derived from distributions of means secured by repeated sampling.
**Values computed from data for entire population.

the table, based in each instance on data from the entire population of classes, show that the design which involves stratification by both geographic location and size of enrolment has the smallest standard error (.39). The remaining three designs listed in descending order of accuracy of sample results are geographic stratification, enrolment stratification and unrestricted selection.

There is relatively little difference between the error for unrestricted selection (.49) and the error for enrolment stratification (.47). Similarly there is little difference between the error for geographic stratification used alone (.41) and the error for the design which employed both geographic and enrolment controls (.39). The latter two designs, however, are both shown to be superior to the former two designs.

In the preceding chapter (Chapter III) comparisons were made among the standard errors for four different designs each of which employed the school as the sampling unit. The results presented in this chapter (Chapter IV) were also derived from four designs. These four designs are basically the same as those used earlier in selecting schools, except that the class instead of the school was used as the sampling unit. This concludes the experimentation with alternative methods of drawing samples of schools and samples of classes. The next step in analyzing the sample data presented in Chapters III and IV will be to compare the school unit with the class unit on the basis of accuracy of results obtained. This will be done by bringing together the results for schools and the results for classes in such a way that the accuracy for a sample of schools drawn from a given design may be compared with the accuracy of a sample of classes drawn from the same design.

CHAPTER V

COMPARISON OF ACCURACIES FOR TWO DIFFERENT SAMPLING UNITS: THE SCHOOL AND THE CLASS

An individual becomes a public school pupil only when he is actually enroled in a given school. Hence the primary grouping of a pupil population is a school. The population under investigation in this study consists of 7724 pupils in 97 different schools with each school representing a cluster of grade 8A pupils. These school clusters, varying in size from three pupils to 435 pupils, were the first of the two types of sampling units to be used. A pupil could be drawn into a sample only if the school in which he was enroled were drawn into it. He could not be selected as individual.

Within each school, pupils are organized in classes for purposes of instruction and administration. Thus the secondary grouping of a pupil population is a class. The population under consideration here therefore consists also of the same 7724 pupils organized in 237 classes. When this alternative cluster was used as the sampling unit, an individual could still be drawn into a sample only as a member of a group. Whether the sampling unit was defined as a school or as a class, individuals were not selected at all. Each of the designs called for the selection of designated groups of pupils.

Comparison of The Four Designs Used in Sampling by Schools and by Classes. The first three sampling designs to be applied, (1) unrestricted selection, (2) enrolment stratification and (3) geographic stratification were essentially the same for both schools and classes. However the methods used in setting up the fourth design (stratification by both geographic location and enrolment) were not the same for the two sampling units. In the case of classes, the primary stratification by geographic location as set up for the preceding class design was retained and sub-stratification by enrolment was carried out within each area thus defined. In the case of schools, however, it was thought that the wide variation of enrolments within each of the originally designated geographic areas made these areas inappropriate to use as primary data when the dual control was applied. New geographic areas were therefore set up and the schools were sub-stratified by enrolment within each of these new areas. There were only three such primary geographic strata for schools as compared with ten for classes.

Since the first three designs were virtually identical for schools and for classes, it is possible to make three separate comparisons of the two different sampling units with respect to accuracy of obtained results. The analyses in Chapters III and IV compared the accuracies of results of different designs with the sampling unit held constant. In this chapter comparisons will be made between the two sampling units with the design held constant. The first of the several comparisons will bring together the results for a sample of schools and for a sample of classes, both chosen by unrestricted selection.

Unrestricted Selection: Results for Schools and for Classes. Since the average number of pupils per school was found to be more than twice the average number of pupils per class, it will be helpful to think of sample size as designating both the number of units and the number of pupils included in the sample. For instance, a group of 50 schools represents a 50-unit sample and a group of 120 classes represents a 120-unit sample. However 120 classes includes somewhat fewer pupils than 50 schools--about 3800 as compared with 4000. These two sample sizes were the largest considered in the analyses of results for schools and classes separately. Since 50 schools and 120 classes contain roughly comparable numbers of pupils, these two sample sizes will be used as the basis for all comparisons to be made of the respective accuracies obtained from the two different sampling units.

Table XXXVIII shows that the standard error for an unrestricted sample of 50 schools including a total of 4000 pupils is .54. The standard error for a comparable number of pupils selected by class groups is .34. When the sample size (number of pupils) is the same, the class unit gives considerably more accurate results than the school unit. The error for schools as shown in the table is more than one and one-half times as large as the error for classes. This relative superiority of class unit over the school unit will remain approximately the

TABLE XXXVIII

Standard Errors for Samples of 50-Schools and 120-Classes Drawn by Unrestricted Selection

Sampling Unit	Size of Sample: No. of Units	No. of Pupils	Standard Error*
School	50	4000	.54
Class	120	3800	.34

*Computed by formula using data from entire population.

same for any given number of pupils that may be selected.

Enrolment Stratification: Results for Schools and for Classes. Table XXXIX compares the errors of mean pupil scores obtained from a sample of schools stratified by size of enrolment and from a sample of classes stratified by size of enrolment. The error for a school sample is .46 as compared with an error of .33 for a class sample. Again the results for classes are shown to be considerably less variable than the results for schools. The relative superiority of the class over the school is not quite as great as in the case of unrestricted selection. This is accounted for by the fact that application of the enrolment control factor resulted in a definite increase in accuracy for school samples as compared with unrestricted selection. When this control factor was applied to classes, it produced almost no increase in accuracy of results as compared with unrestricted selection of classes.

TABLE XXXIX

Standard Errors for Samples of 50-Schools and 120-Classes Each Stratified by Enrolment

Sampling Unit	Size of Sample: No. of Units	Size of Sample: No. of Pupils	Standard Error*
School	50	4000	.46
Class	120	3800	.33

*Computed by formula using data from entire population.

<u>Geographic Stratification: Results for Schools and for Classes.</u> The superiority of the class unit over the school unit in the case of this design is greater than for either unrestricted selection or enrolment stratification. The standard error of the mean for a sample of classes is shown in Table XL to be .29 as compared with .49 for schools. The error for schools is thus almost one and three-fourths times as large as for classes.

The two maps (pages 96 and 159) which give the stratification plans for schools and classes respectively show that the areas designated as school strata are roughly the same as the areas designated as class strata. Each of the defined class strata contains a large proportion of the same pupils included in corresponding school strata. The greater accuracy obtained by selecting class units appears to be due, at least in part, to some factor or factors in addition to the difference in size of the respective clusters (school and class). If this were not the case, it would be expected that the relative superiority of the class unit over the school unit would be about the same for this design as it was for unrestricted selection.

TABLE XL

Standard Errors for Samples of 50-Schools and 120-Classes Each Stratified by Geographic Location

Sampling Unit	Size of Sample: No. of Units	Size of Sample: No. of Pupils	Standard Error*
School	50	4000	.49
Class	120	3800	.29

*Computed by formula using data from entire population.

<u>Geographic-Enrolment Stratification: Results for Schools and for Classes.</u> Certain questions have been raised earlier about the appropriateness of comparing the accuracies of the class unit and the school unit on the basis of results secured from the geographic-enrolment design. The maps on pages 108 and 166 show that the areas representing primary stratification of schools are not the same as the areas representing primary stratification of classes. The results secured in turn from school units and from class units are in a sense "non-comparable" when the basic purpose is to determine the relative accuracies of the two different sampling units with the design held constant. The obtained results should therefore be considered only as suggestive of the relative accuracies of the class and the school in a situation where this type of dual control could be applied to both types of units

in exactly the same way.

The results actually obtained from applying one type of dual control to schools and a different type to classes are shown in Table XLI. The errors are .49 for schools and .27

TABLE XLI

Standard Errors for Samples of 48-Schools and 120-Classes Each Stratified by Both Geographic Location and Enrolment

Sampling Unit	Size of Sample No. of Units	No. of Pupils	Standard Error*
School	48	3800	.49
Class	120	3800	.27

*Computed by formula using data from entire population.

for classes, showing the class to be markedly superior to the school. The error for classes is in fact only a little more than one-half the size of the error for schools. Had the two designs been identical the class would undoubtedly have proved superior to the school but the difference between the two would probably have been somewhat less than indicated in Table XLI.

Summary of Evidence on the Relative Accuracies of School Samples and Class Samples. For each of the four designs applied in this study, samples obtained by selecting classes gave definitely more accurate results than samples obtained by selecting schools. When the total number of pupils in a sample is held constant, the standard error of the mean pupil score for a sample of schools is, on the average, a little more than one and one-half times as large as the comparable error for a sample of classes. The greater accuracy of the class unit was most marked for (1) the geographic design and (2) the combination geographic-enrolment design. The margin of superiority of the class over the school was smallest in the case of enrolment stratification. The enrolment control turned out to be relatively ineffective for classes because of the close similarity in size of membership for a large majority of the class group. For schools, on the other hand, stratification by enrolment was the design which gave the smallest error.

Comparison of Accuracies Obtained from Sampling by Schools, by Classes and by Individual Pupils. Several systematic samples of individual pupils were drawn from the total population (7724) in order to produce a distribution of means for samples of pupils which would be comparable to the distributions of means secured by repeated sampling of schools and of classes. One such sample of individuals was drawn by taking the first, eleventh, twenty-first, etc., pupil from each of the 237 classes. The second sample consisted of pupils 02, 12, 22, etc., from each class, and so on to the tenth sample consisting of pupils 10, 20, 30 etc. This procedure gave 10 different samples with no overlap in membership.[1] Table XLII shows the number of pupils in each sample together with the

[1] A detailed description of the method by which these samples were drawn and the method used in computing the means of the samples is given in Appendix D.

mean, the standard deviation and the estimated standard error of the mean for each sample.

TABLE XLII

Means, Standard Deviation of Scores and Estimated Standard Errors of Means for Ten Samples of Individual Pupils Drawn by Systematic Selection

Sample	No. of Pupils	Mean Score	Stand. Dev. of Scores (σ_x)	Est'd.* Stand. Error of Mean ($s_{\bar{x}}'$)
1	862	66.69	9.30	.32
2	849	66.71	8.91	.31
3	830	66.30	9.21	.32
4	797	66.77	9.13	.32
5	775	66.62	8.96	.32
6	761	66.89	9.25	.34
7	753	66.65	8.90	.32
8	717	67.54	8.89	.33
9	704	66.60	9.07	.34
10	676	66.67	9.04	.35
Mean	772.4	66.74	9.07	.33
Standard Deviation of Sample Averages				.30

*Computed by the formula $s_{\bar{x}}' = \frac{\sigma_x}{\sqrt{N - 1}}$

It will be seen that the sizes of the different samples range from 676 pupils to 862 pupils, each containing roughly 10 per cent of the population (7724). The sample means vary from 66.30 for Sample 3 to 67.54 for Sample 8. This latter sample is observed to be an "unusual" one among the 10 since none of the other means is as high as 67.0. The average of the 10 means is 66.74 and the standard deviation is .30.

This value (.30) represents the "true" standard error of the mean for the 10 possible systematic samples of individuals that can be drawn by the method used.

An estimate of the standard error for each of the 10 means is shown in the extreme right-hand column of Table XLII. These estimates are based on the assumption of random selection of individuals.[2] The average[3] of the 10 estimated errors is .33 which is slightly larger than the empirically determined error (.30).

The computed error for a strictly random sample of 10 per cent of the 7724 pupils with the appropriate correction[4] based on the finite population is .31. There is only a slight variation in the three values of the standard error for a 10 per cent sample of individuals, each derived by a different method. The three obtained values are .30, .33 and .31.

The respective accuracies of the eight different methods of cluster sampling, i.e., four different designs, using two different sampling units for each design, will now be compared by employing an "independent criterion" as the basis of

---

[2]Since the design used was actually stratification by classes and systematic selection within each class, the error estimates which assume unrestricted random selection would be expected to be somewhat larger than the "true" error for a random sample.

[3]Computed by the method described on page 39.

[4]Charles C. Peters and Walter R. Van Voorhis, Statistical Procedures and Their Mathematical Bases, pp. 132-133, New York: McGraw-Hill Book Company, 1940.

comparison. This independent basis for comparison is the computed standard error of the mean (.31) for a random sample of 10 per cent of the individual pupils in the population.

Table XLIII presents the array of errors for the eight cluster methods arranged in order of magnitude along with the error for a 10 per cent random sample of individuals. Each of

TABLE XLIII

Sampling Standard Errors for Different Designs Using the School, the Class and the Individual Pupil as the Sampling Unit

Design	Sampling Unit	No. of Units	No. of Pupils	Per Cent of Pupils	Error of Mean* ($\sigma_{\bar{x}}'$)
Geog.- Enrol. Stratif.	Class	120	3800	49	.27
Geographic Stratif.	Class	120	3800	49	.29
<u>Unrestricted Selection</u>	<u>Individual</u>	<u>772</u>	<u>772</u>	<u>10</u>	.<u>31</u>
Enrolment Stratif.	Class	120	3800	49	.33
Unrestricted Selection	Class	120	3800	49	.34
Enrolment Stratif.	School	50	4000	52	.46
Geog.-Enrol. Stratif.	School	48	3800	49	.49
Geographic Stratif.	School	50	4000	52	.49
Unrestricted Selection	School	50	4000	52	.54

*Computed by formulae from data for entire population.

the cluster samples represented in the table includes roughly 50 per cent of the individual pupils.

It will be seen that only two of the cluster methods give results that are more reliable than the results that could be obtained from a 10 per cent sample of individuals. These two methods are (1) combined enrolment and geographic stratification of classes and (2) geographic stratification of classes used alone. A 50 per cent sample of classes stratified by enrolment and a 50 per cent sample of classes drawn at random are both slightly less reliable than a 10 per cent sample of individuals. None of the 50 per cent samples of schools approaches a 10 per cent sample of individuals in reliability.

Respective Advantages of the School and of the Class as Sampling Units. It has been demonstrated conclusively that a sample consisting of a given number of pupils secured by selecting classes gives a distinctly more accurate representation of the population than a sample of the same size obtained by selecting schools. Samples of classes which include about one-third of the pupil population have been shown to be consistently more reliable than samples of schools which include a little more than one-half the pupil population.

This statistical superiority of the class unit would undoubtedly hold for any type of stratified design which could be applied to both schools and classes. The obvious reason for the greater accuracy of the class unit is that the class cluster contains fewer pupils than the school cluster. The class unit would almost certainly prove superior to the school unit in sampling any pupil population where the average school contains more than one class group. It is hardly necessary to point out that the latter generalization applies to any attribute of the pupil population that might be studied--for example, scores on tests other than reading, ages of pupils, measurements of attitudes and opinions, height, weight, etc.

In some situations the statistical advantages of sampling by classes may be more than offset by the administrative advantages of using the school as the primary unit. Schools are always more readily identified than classes. The time and effort required for planning and supervising the testing of pupils in one-half the schools in a large city system would probably be less than that required for testing one-third of the classes. The school sample, however, would be more expensive than the class sample in terms of pupil time, teacher time and cost of testing materials. In choosing a state-wide or a nation-wide sample the school or the community would, of course, be the only accessible primary units.

Where both the school and the class are under consideration as possible sampling units, some thought should be given to the total number of clusters that would be drawn when one or the other of the two types of units is used. In the case of a design which does not call for sub-sampling of individuals within either the school or the class, the reliability of sample results is more closely related to the total number of <u>units</u>

included than to the total number of <u>individuals</u> included. This fact was illustrated repeatedly in Chapter IV where it was shown that 80-class samples gave more accurate results than 50-school samples even though the latter contained considerably larger numbers of pupils.

CHAPTER VI

SUMMARY AND CONCLUSION

When sampling procedures are used to determine a test norm, two inferences need to be made from the data of the sample. The first is an estimate of the average of the population from which the sample is drawn; the second is an estimate of the variability of this average. Derivations of estimates of an average and of its variability are not two separate problems but rather two different aspects of the same problem. In the brief analysis of certain issues in sampling theory presented in the introductory chapter, it was pointed out that there is at present no theory of estimations for samplings other than random.

It is almost impossible to secure a random sample of individual pupil test scores because of the conditions under which tests have to be administered in the typical school. The only practical means of getting test scores that are widely representative of a fairly large population is to sample groups of pupils rather than individuals. The fact that the units of sampling are not individuals but groups of these individuals does not necessarily involve a negation of the randomness of the sampling. However when a sample is secured by selecting groups, the classical formula for estimating the variability of the obtained average does not apply.

The mathematical theory of estimation of the error of results obtained from sampling by "clusters" is not new. In recent years this method of selecting samples from various types of populations has been applied in fields other than education and psychology.[1,2] In spite of the demonstrated value of this type of sample design it has found little application in the field of education where the nature of organization of pupil populations make it particularly appropriate. The central purpose of this study was therefore stated as follows: To apply appropriate theory to the dual task of (a) selecting samples which consist of existing groups of pupils, i.e., schools and classes and (b) deriving objective measures of the precision of results.

It was explained in Chapter II that the purpose of the study was to be achieved by applying alternative designs for drawing samples from a defined population, in each case using either the school or the class as the sampling unit. The specific problem was to determine which design "yields results most closely representing the population from which samples are drawn."

---

[1] M. H. Hansen, and W. N. Hurwitz, "Relative Efficiencies of Various Sampling Units in Population Inquiries", Journal of the American Statistical Association, Vol. 37 (1942), pp.89-94.

[2] W. G. Cochran, "The Use of Analysis of Variance in Enumeration by Sampling", Journal of American Statistical Association, Vol. 34 (1939), pp. 492-510.

<u>Summary of Results Obtained</u>. The eight different sampling designs applied in turn to the designated population of 7724 pupils may be outlined as follows:

Sample Design	Sampling Unit	Stratification
A	School	None
B	School	Enrolment
C	School	Geographic Location
D	School	Geographic Location and Enrolment
E	Class	None
F	Class	Enrolment
G	Class	Geographic Location
H	Class	Geographic Location and Enrolment

The standard error of the average pupil score for a sample was used as the index of accuracy with which a given sample "represented" the total population.

When the number of pupils drawn into a sample was held constant, it was found that Design H (geographic-enrolment stratification of classes) yielded results which represented the total population more accurately than any of the other designs. Using the same criterion, Design A (unrestricted selection of schools) gave the least accurate results.

The least accurate of the four "class" designs had a smaller error than the most accurate of the four "school" designs. Since the membership in a school group is much larger than the membership in a class group, this indicates that greater differences between sample accuracies were produced by varying the size of the sampling unit (number of pupils per unit) than by varying the type of stratification.

Analysis of the means obtained by drawing samples of individuals showed the latter method to give far more accurate results (for a given number of pupils) than sampling either by schools or by classes. It was shown for example that a 10 per cent random sample of individual pupils (about 800) gave a better representation of the population than a 50 per cent sample of schools (about 4000 pupils). A 50 per cent sample of classes (about 4000 pupils) drawn from the most accurate of the stratified designs gave a slightly better representation of the population than the 10 per cent sample of individuals.

The level of reliability to be considered satisfactory in an achievement test norm was arbitrarily set at one-twentieth of the standard deviation of individual pupil scores. This level of accuracy was achieved with a sample consisting of approximately one-half the total number of schools and with a sample consisting of somewhat less than one-third the total number of classes.

The Importance of "Measurableness" in a Sample Design. All mathematical formulae used in deriving estimates of sampling error are based on the theory of uniform (or known) probability of selection of any element in a designated population. Even in the case of a stratified design which calls for different sampling rates within different strata the theory still holds as far as each separate stratum is concerned.

Therefore the first step in securing valid estimates of sampling error consists of setting up a type of design which will yield valid estimates. After the sample results are in, no amount of statistical manipulation can overcome faults in the design itself. It is not only inappropriate but quite invalid to apply error formulae to a "non-measurable" design. Highly developed statistical techniques thus become useless without adequate control of the sampling method.

Objective estimates of error are of particular importance in the case of a test norm since the sample results are always used to represent a larger population. Scores made by individuals and by groups will be compared with this norm. But there is no way to evaluate an observed difference between an obtained measure and the norm unless there is an estimate of the error of the norm itself. The average score, which we call a norm, also represents a sample result just as does the average score made by the pupils in a few classes. The significance of the difference between two such averages can be determined only where valid estimates of error are available for both.

Implications with Respect to "National Norms." National norms for an achievement test purport to represent average performance of defined populations of pupils in schools throughout the nation. Such norms are usually given for grade groups, for age groups or for both. There are no reports to indicate that any test has ever been given to pupils in grade 8, for example, in all schools in the nation. Test publishers who provide such norms have secured scores from a relatively small proportion of all the schools in the country that enrol grade 8 pupils. It is not likely that any publisher has given serious thought to the task of securing a sample of grade 8 test scores by a process of selecting individuals. Either the school or the community is used as the primary sampling unit even though there may be sub-sampling of individuals.

It is probably impossible for a commercial establishment such as a test publishing concern to exercise complete control over the method of drawing a sample of schools or communities. Participation by a local school in a test standardization program usually involves some monetary cost to the school as well as the expenditure of time and effort. If participation of a given school is to be secured at all, it must be secured on a voluntary basis. Therefore the schools from which data are collected cannot be "drawn at random".

A group of a few hundred schools who select themselves, as it were, in a standardization program may be widely representative but the degree of accuracy with which the results obtained from them represent all schools in the nation is not measureable. Knowledge of the communities sampled may make it appear as though the results are highly reliable. And, as a matter of fact, the average score obtained might be very close to the true population average. The variability of the obtained average, however, cannot be derived from the data of the sample. An approximation of error might be obtained by using the cluster formula as it was applied in this study. Such an approximation even though it be of questionable validity is perhaps better than no "guess" at all as to the reliability of the norm. It is unfortunately true that shrewd guesses do sometimes have to serve as substitutes for objective estimates.

The author of this study has been unable to find a single published achievement test having national norms accompanied by objective estimates of sampling error derived from appropriate cluster formulae. At the present time, therefore, it is not possible to determine precisely the significance of a difference between a national norm and the average score achieved by pupils in a given school or in a given school system.

One hears occasional statements to the effect that, in general, pupils nowadays do not achieve on as high a level as did the pupils of two or three decades ago. At least one research worker feels that evidence from repeated national standardization programs tends to support this generalization.[3] Such a conclusion cannot be either supported or refuted by available data. No reported study of achievement test norms involving a national sample, available to the author of this

---

[3] Correspondence with the Director of the Research Division of World Book Company.

study, gives results whose degree of accuracy is measurable.

Implications for School Surveys. Achievement tests have been commonly used as one means of collecting evidence in connection with comprehensive surveys of city school systems, such as those conducted during the last decade in Boston,[4] St. Louis[5] and Pittsburgh.[6] Although tests were not administered during the course of the Boston Survey, test scores secured earlier in the school year were interpreted by the survey staff.[7]

The description of one phase of the testing program of the St. Louis Survey states that "... 34 elementary schools [out of 104] representing various areas in the city were selected, including schools for white pupils and those for Negro pupils in proportion to the total enrolment of all schools. Groups within each of these schools were selected on a random basis."[8] Later in this description the statement is made that

---

[4]George D. Strayer, A Report of a Survey of the Public Schools of Boston, Massachusetts. Boston: City of Boston Printing Department, 1944. Pp.xxxii + 1127.

[5]George D. Strayer, A Report of a Survey of the Public Schools of St. Louis, Missouri. New York: Bureau of Publications, Teachers College, Columbia University, 1939. Pp. xxiv + 468.

[6]George D. Strayer, The Report of a Survey of the Public Schools of Pittsburgh, Pennsylvania. New York: Bureau of Publications, Teachers College, Columbia University, 1940. Pp.xviii + 564

[7]George D. Strayer, Report of Survey of the Public Schools of Boston, Massachusetts, op. cit., pp. 423-426.

[8]George D. Strayer, A Report of a Survey of the Public Schools of St. Louis, Missouri, op. cit., p. 43.

"...tests were given to 250 pupils in the fifth grade and 250 pupils in the eighth grade chosen to represent a random sampling of pupils in selected schools."[9] Although the sampling designs used for various tests are not described in detail, the partial descriptions suggest that the basic plan was "purposive" selection of schools and random selection of classes within the chosen schools. No mention is made of sampling error although obtained scores are compared with established norms.

In the St. Louis survey, better estimates of the average scores for the total population could have been secured had the basic sampling design called for using the class rather than the school as the primary sampling unit. The classes could have been stratified by the same factors that were actually used in the purposive selection of schools.

In the Pittsburgh Survey, "Twenty elementary schools, four junior high schools, four senior high schools and the four three-year senior high schools were selected for testing...Only a relatively few pupils were tested in any one school and not all the types of tests were given in each school".[10] The primary sample of 20 elementary schools, for example, constituted approximately one-fifth of the elementary schools in the city. Only one of the seven different tests

---

[9] George D. Strayer, op. cit., p. 45.

[10] George D. Strayer, The Report of a Survey of the Public Schools of Pittsburgh, Pennsylvania, op. cit., p. 72-74.

used in the elementary grades was actually administered in as many as 20 different schools. Three of the tests were administered in fewer than 15 of the schools.[11] Since the number of pupils tested per school was approximately 30, it is unlikely that the results were sufficiently reliable to warrant the generalization that "...Pittsburgh children are somewhat retarded [on tests for which comparative results from other communities are available] though perhaps not alarmingly so."[12]

Samples of individuals in the numbers tested in the Pittsburgh Survey would have yielded results permitting fairly accurate generalizations concerning the population. It is extremely improbable that samples drawn, as reported, by using the school as the sampling unit could give reasonably reliable results regardless of the type of stratification used.

In conducting comprehensive city school surveys such as the ones referred to here it would be quite feasible to set up designs for drawing stratified random samples of classes. The sampling plan could be arranged in such a way that even where six or eight different tests were used no particular class would be called on to take more than half of the total battery. With complete control over the process of actually choosing the classes to be tested, the results would yield thoroughly objective estimates of error.

---

[11] Ibid.

[12] George D. Strayer, op. cit., p. 87.

<u>Problems for Further Research</u>. Test norms are commonly considered to refer to a hypothetical population as well as a real population. For instance, the norm derived from a sample of eighth grade pupils in a given year will be used to interpret scores made by eighth grade pupils in the succeeding year and perhaps for some years thereafter. The total population with which the norm is ultimately used does not exist at any given time. Inferences drawn from the sample data are thus applied to the hypothetical population as successive fractions of it come into existence. As we pass through time, the accumulation of generations of pupils in a given grade group or age group becomes larger and larger. The sample on which the norm is based therefore comes to represent a smaller and smaller proportion of the accumulated segments of the hypothetical population.

The expected variation in average achievement of the different "generations" of pupils who constitute the populations of a given grade from year to year is probably not known. It seems reasonable to assume that there is some chance variability from year to year, i.e., variability which is unrelated to instructional materials, teaching methods, promotion policies, etc. Norms which are to be used over a considerable period of time therefore might be found somewhat more reliable if instead of testing a fairly large sample in a single year, relatively smaller samples were chosen in each of two or three successive years. The results from the smaller samples could then be thrown together and considered to represent a single sample drawn from that part of the total hypothetical population which came into existence during a two or three year period rather than a one year period. The data required for such a study could be readily secured by sampling schools or classes.

Certain types of information are sometimes sought from pupil populations which do not require the actual collection of data by groups as in the case of giving a test. Age-grade status is an example. This information already exists as a matter of record. The procedure of sampling would consist of pulling records and transcribing the desired information. Although the pupil records are still accessible only by school or class groups, it is quite practical to sub-sample by individuals. An experimental study could be set up to test the relative efficiencies of sub-sampling by schools and by classes.

Parents' evaluations of a school program in relation to their own children as well as their attitudes toward education in the community at large are matters of first importance to administrators and teachers. Adults who have children enroled in schools are identifiable as clusters the same as pupils. Names and addresses of parents are available both as "school groups" and as "class groups". The most convenient primary sampling unit to use in such a study would probably be the school group. A relatively efficient design could be set up involving stratification of the schools by one or more control factors and sub-sampling of individual parents within the selected schools. A design of this type would yield objective estimates of error since there would be complete control of the method of selecting the individuals from whom information was to be secured by questionnaire or by interview.

Conclusion. The essential problem in determining a test norm for a relatively large population is to administer the test to a sample of pupils which is representative of that population. No attempt has been made in this study to define a "generally representative sample". The procedures used, however, have illustrated a representative method of sampling and a consistent method of estimation of error. A representative method may be defined in a sentence as one which "... makes possible an estimate of the accuracy of estimation irrespective of the unknown properties of the population studied."[13] Each of the designs used for sampling schools and classes satisfied the conditions of this definition. Although the particular attribute under investigation was the score achieved on a reading test, it might have been any other pupil characteristic such as score on a different type of achievement test, height, weight, age, etc. The relative efficiencies of the different designs used here might not remain the same if some other variable were studied. However each design would give a valid estimate of error for any definable pupil characteristic.

---

[13] Jerzy Neyman, "On Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection", Journal of the Royal Statistical Society, Vol. 97 (1934), p. 585.

A norm is an estimate of an average. It has been pointed out that an estimate is of no value whatever unless there is some knowledge of its degree of precision. In some types of investigations, practical considerations may make it impossible to exercise sufficient control over the selection of elements drawn into the sample to warrant actual computation of a standard error. Even in these instances it is essential to have at least a fair idea of the degree of accuracy obtained based on general knowledge and experience.

The accuracy of an achievement test norm for a city school system does not need to be estimated subjectively. It is anticipated that norms to be established in the future for Detroit Public Schools will be based on stratified random sampling of classes. All of the advantages sometimes thought to be associated with purposive selection can be achieved by appropriate stratification. Actual selections of classes can then be made strictly at random. The results secured will yield precise estimates of sampling error.

APPENDIX A

PROCEDURES USED IN ADMINISTERING AND SCORING THE TEST

PROCEDURES USED IN ADMINISTERING AND SCORING THE TEST

I. COPY OF MEMORANDUM, DATED FEBRUARY 4, 1947, ADDRESSED TO PRINCIPALS OF ALL PUBLIC SCHOOLS IN DETROIT ENROLING GRADE 8A PUPILS

1. Introduction

Over the past two decades the Detroit schools have used survey testing as one means of securing information for the study of instructional problems. City-wide surveys in reading were conducted in 1928 and 1937. A third reading survey was undertaken in a sample of elementary and intermediate schools in 1943.

In keeping with this practice of periodic surveys, the Departments of Language Education and Instructional Research recommended that a standardized reading test be given to all 8A pupils during the second semester of this school year. An advisory committee, representing the supervising principals, intermediate school principals, high school principals, language education, and research, has prepared detailed plans for the survey which will be conducted during the last week in February, 1947.

2. Specific Purposes of the Survey

a. To provide basic information on reading achievement to be used as one basis for study and appraisal of the reading program in the elementary and intermediate schools.

b. To provide highly reliable measures of reading achievement for every pupil in grade 8A to be used by 8A teachers in individual guidance.

c. To provide an individual measure of reading achievement to be forwarded to the secondary school, where the pupil will be enroled in grade 9B, for use in individual guidance and counseling.

3. Test to be Used

The Stanford Reading Test, Form DM, will be used in the survey. There are two parts to this test: Paragraph Meaning and Word Meaning. The total testing time for both parts of the Stanford Test is 30 minutes. An additional 30 minutes should be allowed for passing out materials, giving directions, and having the pupils score the test. **Therefore**, the total time required for administering and scoring is about 60 minutes.

4. <u>Administration of the Test</u>

Specially designated examiners have been selected by the supervising principal to give the test in each district. These examiners will be exchanged between schools so that no person gives the test in "his own" school. The supervising principal in each district has also designated a person in the district center school who will serve as <u>District Coordinator</u> for general supervision of the survey.

The District Coordinator will assist in arranging the details of assignment of examiners to individual schools, and will supervise the distribution of testing materials to the examiners.

Examiners assigned to do the testing will meet with representatives of the Department of Instructional Research during the week of February 17 to discuss specific procedures for giving and scoring the test. This will insure that testing conditions will be as nearly uniform as possible in every school. The date, time, and place of this meeting will be given in a special announcement from the District Coordinator.

In general, pupils should be tested in their regular homeroom groups. Some of the examiners will be administering the Stanford Test for the first time. It is important to keep the size of the group small enough for the examiner and the regular teacher to be sure that every pupil fully understands the directions.

One of the regular teachers of the group should remain in the room while the examiner is giving the test. He may assist the examiner in checking to see that all pupils have understood the directions. The presence of a regular teacher should tend to reduce any uneasiness that may be felt by some pupils when tested by a person whom they do not know.

Tests will be scored by the pupils under the supervision of the examiner. This will be done immediately after the test is given. Both administering and scoring can be completed in the 60-minute period scheduled for testing.

A specially prepared homeroom record sheet for recording several items of information about each 8A pupil in addition to his test score will be sent to the principal before the survey date. These record forms should be filled out by the homeroom teacher and returned to the district school after the test is given.

Testing supplies, including test booklets, answer sheets, and tabulation forms will be brought to the building by the examiner on the day scheduled for testing.

5. Date of Testing

The test will be given in each school either on Tuesday, February 25, or on Wednesday, February 26, 1947.

6. Summary of Plans for the Survey

a.	Grade	: All Grade 8A students.
b.	Date	: Tuesday, February 25, or Wednesday, February 26, 1947.
c.	Test	: Stanford Reading Test, Form DM.
d.	Testing Time	: Part 1 of the test requires 20 minutes and Part 2, 10 minutes. A total of about 60 minutes should be allowed for giving and scoring the test.
e.	Size of Groups	: Pupils should be tested in their regular homeroom groups.
f.	Examiners	: In each school, the test will be administered by a person from another school. A regular teacher of the group should be in the room during the testing.
g.	Scoring	: Tests will be scored by the pupils under the supervision of the examiner.
h.	Class Record Sheets	: Individual test scores and certain other items of information for each pupils are to be recorded by the homeroom teacher on a specially prepared form. Copies of this form will be sent to the principal before the date of the survey.
i.	Reporting Results	: Class record sheets are to be returned to the district school.

II. TRAINING OF EXAMINERS

The examiners (assistant principals and teachers) selected to give the test in the different schools were brought together in groups of approximately 15 for a half-day period of instruction in the details of administering and scoring the Stanford Reading Test. Each examiner actually "took" a part of the test himself and carried through the several steps in the scoring process. These steps include counting the numbers of correct answers for the two sub-tests, translating the "numbers right" into equated scores and averaging the equated scores to get a total score.

Since the general plan called for having the tests scored by pupils, the following specific scoring directions were prepared for use by the examiners.

<u>Directions for Scoring the Test</u>

MAJOR STEPS
IN SCORING

Step 1

The test should be scored by pupils under the supervision of the examiner immediately after time is called on Test 2. Have the pupils exchange answer sheets. Explain that any question having more than one answer space marked is to be counted wrong. Then say:

"I WILL READ THE CORRECT ANSWERS FOR TEST 1. IF THE ANSWER IS RIGHT ON YOUR PAPER MAKE A SMALL 'C' ON THE LEFT-HAND SIDE OF THE NUMBER OF THE QUESTION, LIKE THIS: (Put an illustration on the blackboard → C1). IF THE ANSWER IS WRONG, CROSS OUT THE NUMBER OF THE QUESTION LIKE THIS:" (Put an illustration on the blackboard → ~~1~~).

Step 2

"THE FIRST QUESTION TO BE MARKED IS UNDER <u>TEST 1</u> IN THE SECOND COLUMN ON THE ANSWER <u>SHEET</u>. EVERYONE LOOK AT QUESTION 1. HOW MANY FIND A MARK IN SPACE 12? (Wait for

Step 2 (Cont'd.)

pupils to raise their hands) Then say: IF (12' IS MARKED, THE ANSWER IS CORRECT. MARK IT 'C' LIKE THE SAMPLE ON THE BOARD. IF IT IS WRONG, MARK IT 'X' LIKE THE SAMPLE ON THE BOARD."

"THE CORRECT ANSWER TO QUESTION 2 IS '14', ETC. (Read the rest of the answers for Test 1 from the Answer Key which appears on Page 6 of these directions.)

After the correct answers have been read for the 45 questions in Test 1, say to the pupils:

"NOW COUNT THE NUMBER OF C's AND WRITE THAT NUMBER OPPOSITE 'TEST 1 SCORE' AT THE TOP OF THE ANSWER SHEET."

Step 3

"NOW I WILL READ THE CORRECT ANSWERS FOR TEST 2. LOOK AT THE FIRST QUESTION. THE CORRECT ANSWER IS '1'. MARK EACH QUESTION IN THE SAME WAY YOU MARKED THE QUESTIONS ON TEST 1."

"THE ANSWER TO QUESTION 2 IS '9', ETC." (Read the rest of the answers for Test 2 from the Answer Key.)

After the correct answers have been read for the 50 questions on Test 2, say to the pupils:

"NOW COUNT THE NUMBER OF C's. WRITE THAT NUMBER OPPOSITE 'TEST 2 SCORE' AT THE TOP OF THE ANSWER SHEET."

Step 4

When all have finished, have the pupils exchange papers a <u>second time</u>, making sure that no one has his own paper. Have the counting and recording of "C's" checked. Any errors in original counting should be corrected by changing the number recorded at the top of the answer sheet.

The translations of raw scores into equated scores were made according to published directions.[1] This step was carried out by the examiners and by the regular teachers of the

---

[1] <u>Stanford Advanced Language Arts Test: Directions for Administering</u>, p. 4. Chicago: World Book Co., 1941.

groups tested. Equated scores were first recorded on the pupils answer sheet and then transcribed to specially prepared class record sheets. All original data used in this study were therefore obtained in the form of individual pupil scores recorded on class record sheets.

All original pupil answer sheets were also collected for the purpose of making a sampling check on the accuracy of scoring. Several samples of answer sheets consisting of about 50 each were re-scored. Comparisons of the obtained results with the recorded scores for the same papers revealed a few mistakes in the original scoring. The errors were compensating, however, and it was deemed unnecessary even to continue the process of re-checking additional samples.

An IBM card was punched for each pupil directly from the class record forms. Every card was first verified, then the squares of individuals' scores were punched by running the cards through the multiplier. All sums and sums of squares for both schools and classes were obtained from the tabulating machine.

APPENDIX B

SAMPLE COPIES OF READING TEST

AND

TEST ANSWER SHEET

TANFORD ACHIEVEMENT TEST

By TRUMAN L. KELLEY, GILES M. RUCH, and LEWIS M. TERMAN

ADVANCED LANGUAGE ARTS TESTS FORM DM

Adv. Lang. Arts **DM**

For Use with Separate Answer Sheet

Do not open this booklet or turn it over until you are told to do so.

Samples

1-2 Dick and Tom were playing ball in the field. Dick was throwing the —1— and —2— was trying to catch it.

3-4 Mary's mother gave her a little garden for her own. In it she planted some lettuce and some onions. Soon Mary hopes to gather lettuce and —3— from her —4—.

Answers

1	1 bat 3 field	2 toy 4 ball
2	5 Dick 7 field	6 Tom 8 she
3	9 flowers 11 onions	10 vegetables 12 radishes
4	13 garden 15 seeds	14 work 16 plants

5 A rose is a — 1 box 2 flower 3 home 4 month 5 river 5

6 A roof is found on a — 6 book 7 person 8 rock 9 house 10 word . . . 6

7 Apples 1 is / 2 are good. 7

8 He 3 told / 4 telled me. 8

9 A black 1 catt / 3 katt 2 kat / 4 cat ran across our path. 9

10 What is 5 hiz / 7 hiss 6 his / 8 hizz name? 10

Published 1941 by World Book Company, Yonkers-on-Hudson, New York, and Chicago, Illinois
 SAT: ADV. LANG. ARTS: DM-

PRINTED IN U.S.A.

TEST 2. READING: WORD MEANING

RECTIONS. In each exercise one of the five numbered words will complete the sentence correctly. Note the number of this word. Then mark on the answer sheet the space which is numbered the same as the word you have selected. Study the samples and answer the other questions in the same way.

MPLES.

A **rose** is a — 1 box 2 flower 3 home 4 month 5 river 5

A **roof** is found on a — 6 book 7 person 8 rock 9 house 10 word 6

To be **content** is to be — 1 satisfied 2 angry 3 awake 4 faithful 5 bold 1

To have **sympathy** for is to — 6 rejoice 7 praise 8 refuse 9 pity 10 shame 2

A **safeguard** is a — 1 plague 2 bureau 3 defeat 4 protection 5 pause 3

A **communication** is a — 6 palace 7 message 8 companion 9 struggle 10 memory 4

Gloomy means — 1 heroic 2 fragrant 3 dismal 4 gorgeous 5 majestic 5

Behavior refers to — 6 courage 7 conduct 8 appearance 9 effort 10 features 6

Intelligence means — 1 wisdom 2 justice 3 anger 4 liberty 5 praise 7

Disobedient means — 6 clumsy 7 critical 8 credulous 9 grotesque 10 unruly 8

A **situation** refers to a — 1 rival 2 majority 3 capture 4 position 5 strain 9

A **counselor** is a — 6 beggar 7 carpenter 8 lawyer 9 dragon 10 chariot 10

6➡

11 **Rotation** means in — 1 agreement 2 attendance 3 production 4 shipment 5 succession . . . 11

12 An **abode** is a place where one — 6 earns 7 dwells 8 bakes 9 parks 10 swims 12

13 An **emperor** is a kind of — 1 priest 2 robber 3 official 4 witch 5 beggar 13

14 To **bleach** is to — 6 harden 7 darken 8 lighten 9 soften 10 sharpen 14

15 **Capacity** refers to — 1 climate 2 vanity 3 habit 4 poverty 5 volume 15

16 To **insinuate** is to — 6 accustom 7 suggest 8 counsel 9 surround 10 injure 16

17 **Violence** often causes — 1 wisdom 2 respect 3 justice 4 knowledge 5 harm 17

18 To **detect** is to — 6 discover 7 lower 8 hide 9 practice 10 reply 18

19 **Lasting** means — 1 literal 2 specific 3 unnatural 4 durable 5 stagnant 19

20 A **sprig** is a — 6 thief 7 ditch 8 pillow 9 prophet 10 twig 20

21 To **shrivel** is to — 1 stumble 2 stagger 3 rustle 4 wrinkle 5 waver 21

22 **Crafty** means — 6 shrewd 7 bashful 8 confident 9 forlorn 10 envious 22

23 A **hillock** is a — 1 memorial 2 mound 3 nerve 4 knave 5 patron 23

Go right on to the next page.

STANFORD ACHIEVEMENT TEST

By TRUMAN L. KELLEY, GILES M. RUCH, and LEWIS M. TERMAN

ADVANCED LANGUAGE ARTS TESTS
FORM DM

Adv. Lang. Arts
DM

For Use with Separate Answer Sheet

Do not open this booklet or turn it over until you are told to do so.

Samples

1-2 Dick and Tom were playing ball in the field. Dick was throwing the —1— and —2— was trying to catch it.

Answers		
1	1 bat 3 field	2 toy 4 ball
2	5 Dick 7 field	6 Tom 8 she

3-4 Mary's mother gave her a little garden for her own. In it she planted some lettuce and some onions. Soon Mary hopes to gather lettuce and —3— from her —4—.

3	9 flowers 11 onions	10 vegetables 12 radishes
4	13 garden 15 seeds	14 work 16 plants

5 A rose is a — 1 box 2 flower 3 home 4 month 5 river 5

6 A roof is found on a — 6 book 7 person 8 rock 9 house 10 word . . . 6

7 Apples 1 is / 2 are good. 7

8 He 3 told / 4 telled me. 8

9 A black 1 catt / 3 katt 2 kat / 4 cat ran across our path. 9

10 What is 5 hiz / 7 hiss 6 his / 8 hizz name? 10

Published 1941 by World Book Company, Yonkers-on-Hudson, New York, and Chicago, Illinois

SAT: ADV. LANG. ARTS: DM-

PRINTED IN U.S.A.

TEST 2. READING: WORD MEANING (*Cont'd*)

24 Bondage means — 6 bravery 7 amusement 8 slavery 9 instinct 10 haven

25 Monopoly means exclusive — 1 control 2 custom 3 fashion 4 expense 5 judgment

26 A petition is an — 6 assembly 7 offense 8 estate 9 embrace 10 appeal

27 Ensnared means — 1 struck 2 shocked 3 trapped 4 weary 5 unknown

28 Abashed means — 6 amiable 7 capacious 8 woeful 9 unreasonable 10 embarrassed

29 Minimum means the — 1 largest 2 least 3 most 4 newest 5 oldest

30 Ceaseless means — 6 entirely 7 flexible 8 incessant 9 elaborate 10 formidable

31 A vulgar deed is — 1 honest 2 base 3 friendly 4 noble 5 generous

32 Prior means — 6 flowery 7 pious 8 jolly 9 parallel 10 former

33 Stagnant means — 1 stormy 2 foul 3 sober 4 cunning 5 sandy

34 Tumultuous means — 6 jocund 7 lowly 8 boisterous 9 unspeakable 10 thoughtless

35 To repulse is to — 1 isolate 2 repel 3 rebuild 4 exaggerate 5 replenish

36 To decompose is to — 6 carve 7 astonish 8 excite 9 decay 10 overcome

37 A loathing is a — 1 clatter 2 dislike 3 revel 4 lamentation 5 rebellion

38 Demeanor refers to — 6 conduct 7 speech 8 property 9 influence 10 fortune

39 To empower is to — 1 authorize 2 conjure 3 comprise 4 submerge 5 stimulate

40 To thrive is to — 6 draw 7 endeavor 8 supply 9 grow 10 waste

41 An affront is an — 1 insult 2 amendment 3 expedition 4 ancestor 5 ordinance

42 A stratagem is a — 6 greeting 7 doctrine 8 scheme 9 miracle 10 ceremony

43 Opportune means — 1 reasonable 2 foremost 3 uncertain 4 suitable 5 apparent

44 Lithe means — 6 flexible 7 massive 8 somber 9 eloquent 10 tremulous

45 To apprehend is to — 1 stifle 2 perceive 3 venture 4 provoke 5 betray

46 Modesty is lack of — 6 vigor 7 patience 8 vanity 9 charity 10 appeal

47 Enmity is — 1 hatred 2 contempt 3 cruelty 4 bliss 5 reverence

48 Cherubim are — 6 bushes 7 chickens 8 bottles 9 curtains 10 angels

49 Paltry means — 1 fickle 2 odious 3 sumptuous 4 thrifty 5 trivial

50 Conformity means — 6 contempt 7 extent 8 accord 9 contrast 10 doctrine

End of Test 2. Look over your

TEST 2. READING: WORD MEANING

DIRECTIONS. In each exercise one of the five numbered words will complete the sentence correctly. Note the number of this word. Then mark on the answer sheet the space which is numbered the same as the word you have selected. Study the samples and answer the other questions in the same way.

SAMPLES.

5 A **rose** is a — 1 box 2 flower 3 home 4 month 5 river 5

6 A **roof** is found on a — 6 book 7 person 8 rock 9 house 10 word 6

1 To be **content** is to be — 1 satisfied 2 angry 3 awake 4 faithful 5 bold 1

2 To have **sympathy** for is to — 6 rejoice 7 praise 8 refuse 9 pity 10 shame 2

3 A **safeguard** is a — 1 plague 2 bureau 3 defeat 4 protection 5 pause 3

4 A **communication** is a — 6 palace 7 message 8 companion 9 struggle 10 memory 4

5 **Gloomy** means — 1 heroic 2 fragrant 3 dismal 4 gorgeous 5 majestic 5

6 **Behavior** refers to — 6 courage 7 conduct 8 appearance 9 effort 10 features 6

7 **Intelligence** means — 1 wisdom 2 justice 3 anger 4 liberty 5 praise 7

8 **Disobedient** means — 6 clumsy 7 critical 8 credulous 9 grotesque 10 unruly 8

9 A **situation** refers to a — 1 rival 2 majority 3 capture 4 position 5 strain 9

10 A **counselor** is a — 6 beggar 7 carpenter 8 lawyer 9 dragon 10 chariot 10

6→

11 **Rotation** means in — 1 agreement 2 attendance 3 production 4 shipment 5 succession . . . 11

12 An **abode** is a place where one — 6 earns 7 dwells 8 bakes 9 parks 10 swims 12

13 An **emperor** is a kind of — 1 priest 2 robber 3 official 4 witch 5 beggar 13

14 To **bleach** is to — 6 harden 7 darken 8 lighten 9 soften 10 sharpen 14

15 **Capacity** refers to — 1 climate 2 vanity 3 habit 4 poverty 5 volume 15

16 To **insinuate** is to — 6 accustom 7 suggest 8 counsel 9 surround 10 injure 16

17 **Violence** often causes — 1 wisdom 2 respect 3 justice 4 knowledge 5 harm 17

18 To **detect** is to — 6 discover 7 lower 8 hide 9 practice 10 reply 18

19 **Lasting** means — 1 literal 2 specific 3 unnatural 4 durable 5 stagnant 19

20 A **sprig** is a — 6 thief 7 ditch 8 pillow 9 prophet 10 twig 20

21 To **shrivel** is to — 1 stumble 2 stagger 3 rustle 4 wrinkle 5 waver 21

22 **Crafty** means — 6 shrewd 7 bashful 8 confident 9 forlorn 10 envious 22

23 A **hillock** is a — 1 memorial 2 mound 3 nerve 4 knave 5 patron 23

Go right on to the next page.

TEST 1. READING: PARAGRAPH MEANING

DIRECTIONS. Read each paragraph below. Decide which one of the four words at the right is the b… for each blank. Make a mark on your answer sheet in the space which is numbered the same a… choice. Study the sample below, and answer the other questions in the same way.

SAMPLE.

1-2 Dick and Tom were playing ball in the field. Dick was throwing the —1— and —2— was trying to catch it.

	Answers	
1	1 bat 3 field	2 toy 4 ball
2	5 Dick 7 field	6 Tom 8 she

1-2-3 In olden days men made their own pens from the quills of feathers. It required considerable skill to cut a pen properly so as to suit one's individual taste in writing. Students were always on the lookout for good goose, swan, turkey, or other bird feathers. Goose quills made the most satisfactory —1— for general —2—, but schoolmasters liked pens made from the —3— of swan feathers because they fitted best behind the ear.

1	9 quills 11 feathers	10 writin… 12 pens
2	13 taste 15 students	14 use 16 Washin…
3	17 feathers 19 quills	18 tails 20 wings

4-5 In this country we seldom hear of duels today, but in colonial days, and for some time after, the duel played a considerable part in American political and social life. Many great names have been connected with the story of the duels in this —4—, including one President of the United States. The most famous of all American —5— was the tragic meeting of Hamilton and Burr.

4	21 time 23 country	22 way 24 period
5	25 duels 27 Presidents	26 people 28 men

6-7-8 One of the paradoxical developments of the machine age is increased leisure-time interest in handicrafts. Such activities as weaving, woodworking, and knitting are carried on by large numbers of persons. By decreasing the amount of time necessary to perform many kinds of labor, the widespread use of —6— has increased the desire of people to do work with their —7— in their —8— time.

6	29 handicrafts 31 knitting	30 machin… 32 labor
7	33 weaving 35 machines	34 hands 36 knitting
8	37 olden 39 labor	38 leisure 40 early

9-10-11 There were three great philosophers and leaders of thought in ancient Greece who are still revered by students everywhere. Socrates, the first of these, was put to death because of his teachings. Plato, a young student of Socrates, took up the latter's work after his death and taught the people what he had learned from his great instructor. After Plato came Aristotle, who, though he belonged to a different school of thought from that of —9— and —10—, was an equally great —11— and teacher.

9	41 Greece 43 Aristotle	42 philoso… 44 Plato
10	45 Aristotle 47 Greece	46 Socrat… 48 war
11	49 instructor 51 man	50 philoso… 52 person

12-13 In speaking of gold, the term "carat" is used to indicate the proportion of gold in a given mass. A carat is one twenty-fourth of the whole mass. Thus, a fourteen-carat ring is one with fourteen parts of pure gold and ten parts of some other metal, usually copper. A twenty-four-carat watch chain is pure —12—. A bracelet that is half gold and half copper would be called a —13—-carat gold bracelet.

12	53 mass 55 copper	54 carat 56 gold
13	57 twelve 59 fourteen	58 half 60 twenty…

Go right on to the next page.

14-15-16 Samuel Weller is a character in Dickens's well-known book, *Pickwick Papers*. He was a servant to Mr. Pickwick and was devoted to his master. He is a very entertaining character, combining wit, simplicity, humor, and fidelity. When *Pickwick Papers* came out, people were greatly amused by Mr. Pickwick and —14— —15— and were eager to read other books written by —16—.

14	**1 Dickens** / **3 Pickwick**	**2 master** / **4 Samuel**
15	**5 papers** / **7 Weller**	**6 Dickens** / **8 characters**
16	**9 Pickwick** / **11 Dickens**	**10 them** / **12 Weller**

17-18-19 "Blue stocking" means, figuratively, a female pedant, or one who emphasizes learning unduly. It derives its name from literary societies of the seventeenth and eighteenth centuries whose members wore —17— —18— as a distinguishing mark. In present-day usage the term is applied to women who make a display of their —19—.

17	**13 blue** / **15 long**	**14 silk** / **16 white**
18	**20 pedant** / **22 stockings**	**21 clothes** / **23 shirts**
19	**24 clothes** / **26 legs**	**25 society** / **27 learning**

20-21-22-23 The National Gallery in London is one of the most famous art galleries of the world. It is full of masterpieces by the world's greatest —20—. These pictures have been garnered chiefly from the private collections of England, either through bequest or by purchase. When a member of the nobility dies without an heir, he bequeaths the —21— treasures collected by his ancestors to this —22—. If, on the other hand, family fortunes are depleted and an heir finds himself in need of money, his art treasures often go to the auction room and are —23— for the Gallery.

20	**28 galleries** / **30 masterpieces**	**29 arts** / **31 painters**
21	**32 gallery** / **34 art**	**33 masterpiece** / **35 national**
22	**36 heir** / **38 country**	**37 gallery** / **39 generation**
23	**40 purchased** / **42 sold**	**41 collected** / **43 auctioned**

4→

24-25 What makes a farmer decide to grow wheat instead of cabbage in a certain field? Although many factors enter into his decision, probably the most important is the kind of soil. Drainage is frequently also a limiting factor, but —24— more often than —25— is the main factor.

24	**44 location** / **46 drainage**	**45 soil** / **47 climate**
25	**48 wheat** / **50 soil**	**49 drainage** / **51 rain**

26-27 Crimes may be classified as either misdemeanors or felonies. The more serious ones fall into the latter class. Murder is a —26—; bribery, no matter how strongly society condemns it, is usually classed legally as a —27—.

26	**52 crime** / **54 felony**	**53 misdemeanor** / **55 killing**
27	**56 felony** / **58 crime**	**57 misdemeanor** / **59 murder**

28-29 A habit is a tendency to respond in a particular manner to a given situation that has become fixed through repetition. The more these responses are —28—, the more —29— they become.

28	**60 habit** / **62 fixed**	**61 practiced** / **63 satisfied**
29	**64 habitual** / **66 undesirable**	**65 particular** / **67 popular**

30-31 Chile is a country of great versatility and wonderful natural gifts. In the mountains are rich mines, and in the lower regions the soil is fertile and productive. In the sea, rivers, and lakes all kinds of fish are to be found. This explains why —30— is able to satisfy the most varied —31—.

30	**68 Chile** / **70 soil**	**69 fish** / **71 nature**
31	**72 people** / **74 gifts**	**73 products** / **75 needs**

Go right on to the next page.

32-33 The Frenchman Descartes won fame both as a philosopher and as a mathematician. His ideas and principles are known as the Cartesian system of philosophy. Eminent as a —32—, Descartes, as the founder of analytical geometry, must also be regarded as a great —33—.

32	1 mathematician 3 philosopher	2 Frenchm 4 Cartesi
33	5 philosopher 7 teacher	6 astron 8 mathem

34-35-36 Aggravate means to make worse or intensify, while alleviate means to lighten or mitigate. Allay, although similar to alleviate in meaning, is used more in the sense of to put at rest or quell. We would say, for example, that the man sought to —34— the burden of his responsibilities; or, in another case, that his fears regarding the future were —35— by the fortune he inherited. On the other hand, a man's illness would be —36— by the shock of bad news.

34	9 allay 11 alleviate	10 aggrava 12 intensify
35	13 alleviated 15 mitigated	14 aggravat 16 allayed
36	17 allayed 19 aggravated	18 mitigated 20 quelled

37-38 The story of early Greek life is a tale of moderation and simplicity. Their food and clothes were simple, and they disliked the possession of elaborate material things. Above all, they wanted to be free both in mind and body. The Greeks —37— gaudy display, and with their love of liberty, encouraged —38— speech.

37	21 possessed 23 disliked	22 liked 24 encourag
38	25 loud 27 simple	26 good 28 free

39-40-41 *Candor* impels us to acknowledge even that which may militate against ourselves, *openness* obliges us to say whatever passes through the mind, and *sincerity* prevents us from saying what we do not believe. In other words, —39— is unguarded, —40— is free from dissimulation, and —41— is disinterested and impartial.

39	29 openness 31 sincerity	30 candor 32 truth
40	33 candor 35 openness	34 sincerity 36 speech
41	37 candor 39 truth	38 sincerity 40 openness

42-43 To pant for recognition, to yearn to impress one's personality upon one's fellow-men, is the essence of ambition. The ambitious person may think that he merely thirsts to "do something" or "be somebody," but really what he craves is to figure potently in the minds of others, to be greatly loved, admired, or feared. To reap even a great success which no one —42— does not satisfy the yearnings of the —43— individual.

42	45 craves 47 yearns	46 recognizes 48 fears
43	49 average 51 admired	50 other 52 ambitious

44-45 Among the most characteristic and amazing properties of bacteria is their capacity for rapid multiplication. It has been estimated that the descendants of one bacterium under continued favorable conditions would in two days number 281,500,000,000 and in three days weigh about 7000 tons. Fortunately, under ordinary conditions —44— does not proceed unchecked at such a —45—.

44	53 bacterium 55 reproduction	54 capacity 56 characteri
45	57 rate 59 property	58 condition 60 time

End of Test 1. Look over your work.

NAME .... AGE ...... GRADE ...... BOY ...... GIRL ......

SCHOOL .... TEACHER ....

CITY .... STATE .... DATE ....

TEST	Paragraph Meaning	Word Meaning	Language Usage	Spelling
AGE EQUIV.				
GRADE EQUIV.				

SAMPLES

TEST 1

TEST 1 (Cont'd)

TEST 1 (Cont'd)

TEST 2

TEST 2 (Cont'd)

Number Right — Test 1 / Test 2

Equated Score

Published by World Book Company, Yonkers-on-Hudson, New York, and Chicago, Illinois

a

IBM FORM I.T.S. 1010 B 1371

APPENDIX C

SUMMARIES OF TEST DATA

FOR

INDIVIDUAL SCHOOLS AND FOR INDIVIDUAL CLASSES

SUMMARIES OF TEST DATA FOR INDIVIDUAL SCHOOLS AND FOR INDIVIDUAL CLASSES

Tables XLIV and XLV give the basic data for the 97 schools and for the 237 classes, respectively. The sums of scores and the sums of squared scores shown in the two tables were obtained from the original pupil scores (i.e. equated scores) recorded on 7724 IBM cards. No "grouping" of data was involved in any of the computations.

TABLE XLIV

Summary of Test Data for Each of 97 Schools--Number of Pupils Enroled Number Tested, Sum of Scores, Sum of Scores Squared Mean and Standard Deviation

School	No. Enroled	No. Tested	Sum of Scores (Σx)	Sum of Squared Scores (Σx^2)	Mean Score ($\bar{x}$)	S.D. of Scores (σ_x)
01	39	37	2,381	155,505	64.35	7.86
02	41	39	2,746	195,550	70.41	7.52
03	31	27	1,619	98,125	59.96	6.22
04	34	31	1,832	109,854	59.10	7.16
05	50	47	2,641	151,225	56.19	7.75
06	48	45	2,946	195,120	65.47	7.08
07	32	30	1,917	124,999	63.90	9.13
08	14	13	952	70,154	73.23	5.81
09	89	88	6,441	476,969	73.19	7.93
10	39	35	2,350	160,282	67.14	8.45
11	42	39	2,864	212,520	73.44	7.51
12	30	27	1,664	104,928	61.63	9.38
13	92	91	6,372	451,260	70.02	7.47
14	34	33	2,171	143,789	65.79	5.40
15	55	50	3,453	241,425	69.06	7.70
16	50	49	3,054	193,790	62.33	8.39
17	50	47	3,420	251,836	72.77	7.96
18	66	66	4,422	302,076	67.00	9.38
19	32	32	2,305	167,743	72.03	7.31
20	06	05	362	26,350	72.40	5.31
21	159	149	8,576	502,102	57.56	7.55
22	52	47	2,769	165,689	59.81	7.37
23	55	53	3,820	278,918	72.08	8.23
24	48	46	2,985	197,131	64.89	8.64
25	76	75	4,876	322,474	65.01	8.54

TABLE XLIV (Continued)

School	No. Enroled	No. Tested	Sum of Scores (Σx)	Sum of Squared Scores (Σx^2)	Mean Score ($\bar{x}$)	S.D. of Scores (σ_x)
26	81	80	5,616	399,328	70.20	7.97
27	39	37	2,423	161,361	65.49	8.52
28	55	53	3,533	238,081	66.66	6.96
29	112	110	7,865	570,757	71.50	8.74
30	54	53	3,586	247,536	67.66	9.62
31	47	47	3,216	222,948	68.43	7.84
32	67	66	4,885	365,143	74.02	7.36
33	36	32	2,140	144,318	66.88	6.14
34	17	16	980	61,302	61.25	8.93
35	78	78	5,230	355,296	67.05	7.69
36	20	20	1,021	61,591	60.06	3.99
37	52	52	3,749	274,499	72.10	9.00
38	37	34	2,161	138,535	63.56	5.90
39	48	47	3,112	210,462	66.21	9.68
40	42	40	2,470	154,448	71.75	6.94
41	54	51	3,291	216,023	64.53	8.47
42	47	45	2,803	176,499	62.29	6.50
43	36	36	2,291	147,543	63.64	6.96
44	29	28	1,893	130,097	67.61	8.69
45	30	29	1,941	130,973	66.93	6.05
46	30	29	2,059	148,403	71.00	8.74
47	34	32	1,977	123,575	61.78	6.69
48	38	36	2,539	181,157	70.53	7.61
49	30	28	1,742	110,822	62.21	9.34
50	24	19	1,143	69,769	60.16	7.29
51	54	53	3,567	245,197	67.30	9.84
52	20	19	1,183	74,607	62.26	7.07
53	43	42	2,794	188,962	66.52	8.58
54	16	16	1,102	76,908	68.88	7.94
55	23	22	1,435	94,793	65.23	7.36
56	32	29	1,930	131,248	66.55	9.83
57	33	30	1,967	131,459	65.57	9.11
58	105	102	7,179	512,913	70.38	8.65
59	27	22	1,448	97,254	65.82	9.41
60	65	65	4,613	331,817	70.97	8.26
61	32	30	2,021	138,237	67.37	8.34

TABLE XLIV (Continued)

School	No. Enroled	No. Tested	Sum of Scores (Σx)	Sum of Squared Scores (Σx^2)	Mean Score ($\bar{x}$)	S.D. of Scores (σ_x)
62	60	57	3,865	265,689	67.81	7.96
63	59	56	3,872	272,218	69.14	8.96
64	68	68	4,401	290,783	64.72	9.35
65	52	51	3,237	211,671	63.47	11.04
66	44	42	2,536	155,020	60.38	6.71
67	32	32	2,058	134,106	64.31	7.40
68	24	24	1,531	99,147	63.79	7.86
69	32	32	1,762	98,252	55.06	6.20
70	36	34	2,397	172,527	70.50	10.20
71	21	21	1,363	90,645	64.90	10.14
72	24	19	1,163	72,225	61.21	7.39
73	20	19	1,275	86,661	67.11	7.62
74	49	47	3,267	231,225	69.51	9.38
75	35	34	2,204	144,870	64.82	7.67
76	75	75	4,987	337,705	66.49	9.02
77	70	66	4,438	303,290	67.24	8.59
78	08	08	432	23,816	54.00	7.81
79	03	03	208	14,630	69.33	8.34
81	298	273	18,470	1,265,296	67.66	7.58
82	211	208	13,629	905,305	65.52	7.68
83	227	218	13,652	868,176	62.62	7.79
84	159	150	9,503	611,011	63.35	7.73
85	459	435	31,484	2,313,058	72.38	8.88
86	301	282	18,909	1,285,295	67.05	7.85
87	164	155	9,351	575,355	60.33	8.51
88	152	146	9,243	592,377	63.31	7.03
89	233	221	15,217	1,064,459	68.86	8.69
90	423	396	27,745	1,969,767	70.06	8.08
91	309	270	17,919	1,204,897	66.37	7.62
92	189	181	10,992	678,268	60.73	7.70
93	214	198	13,086	875,050	66.09	7.17
94	240	228	15,754	1,104,394	69.10	8.34
95	204	191	13,196	925,056	69.09	8.36
96	146	139	8,242	496,940	59.30	7.69
97	401	385	26,559	1,860,661	68.98	8.60
98	246	234	15,673	1,068,085	66.98	8.85
Total	8139	7724	515,463	35,031,605	6,399.77	772.85
Mean	83.9	79.5	66.74	......	65.98	7.97

TABLE XLV

Summary of Test Data for Each of 237 Classes--Number of Pupils Enroled Number Tested, Sum of Scores, Sum of Scores Squared and Mean Score

Class	School	No. Enroled	No. Tested (N)	Sum of Scores (Σx)	Sum of Squared Scores (Σx^2)	Mean Score ($\bar{x}$)
001	01	39	37	2,381	155,505	64.35
002	19	32	32	2,305	167,743	72.03
003	22	32	29	1,624	91,762	56.00
004	22	20	18	1,145	73,927	63.61
005	24	48	46	2,985	197,131	64.89
006	50	24	19	1,143	69,769	60.16
007	03	31	27	1,619	98,125	59.96
008	16	40	39	2,373	146,929	60.85
009	16	10	10	681	46,861	68.10
010	18	41	41	2,739	186,143	66.80
011	18	25	25	1,683	115,933	67.32
012	28	36	34	2,203	144,257	64.79
013	28	19	19	1,330	93,824	70.00
014	31	37	37	2,599	184,587	70.24
015	31	10	10	617	38,361	61.70
016	32	41	40	3,099	241,637	77.48
017	32	26	26	1,786	123,506	68.69
018	59	27	32	1,448	97,254	65.82
019	73	20	19	1,275	86,661	67.11
020	07	32	30	1,917	124,999	63.90
021	26	39	39	2,811	204,589	72.08
022	26	42	41	2,805	194,739	68.41
023	27	39	37	2,423	161,361	65.49
024	34	17	16	980	61,302	61.25
025	43	36	36	2,291	147,543	63.64
026	45	30	29	1,941	130,973	66.93
027	48	38	36	2,539	181,157	70.53
028	61	32	30	2,021	138,237	67.37
029	63	39	38	2,739	199,471	72.08
030	63	20	18	1,133	72,747	62.94
031	64	34	34	2,175	141,517	63.97
032	64	34	34	2,226	149,266	65.47
033	70	36	34	2,397	172,527	70.50
034	75	35	34	2,204	144,870	64.82
035	77	37	36	2,419	165,439	67.19

TABLE XLV (Continued)

Class	School	No. Enroled	No. Tested (N)	Sum of Scores (Σx)	Sum of Squared Scores (Σx^2)	Mean Score ($\bar{x}$)
036	77	33	30	2,019	137,851	67.30
037	08	14	13	952	70,154	73.23
038	09	39	39	2,924	221,480	74.97
039	09	40	39	2,869	213,207	73.56
040	09	10	10	648	42,282	64.80
041	10	39	35	2,350	160,282	67.14
042	11	42	39	2,864	212,520	73.44
043	13	40	40	2,930	215,892	73.25
044	13	39	39	2,560	170,164	65.64
045	13	13	12	882	65,204	73.50
046	15	45	41	2,816	195,748	68.68
047	15	10	09	637	45,677	70.78
048	17	50	47	3,420	251,836	72.77
049	20	06	05	362	26,350	72.40
050	23	40	38	2,854	215,934	75.11
051	23	15	15	966	62,984	64.40
052	29	37	37	2,640	190,600	71.35
053	29	38	37	2,675	196,195	72.30
054	29	37	36	2,550	183,962	70.83
055	35	41	41	2,875	204,001	70.12
056	35	37	37	2,355	151,295	63.65
057	37	43	43	3,101	227,011	72.12
058	37	09	09	648	47,488	72.00
059	39	37	36	2,451	170,143	68.08
060	39	11	11	661	40,319	60.09
061	41	39	37	2,458	165,580	66.43
062	41	15	14	833	50,443	59.50
063	53	43	42	2,794	188,962	66.52
064	56	32	29	1,930	131,248	66.55
065	74	49	47	3,267	231,225	69.51
066	79	03	03	208	14,630	69.33
067	05	40	37	2,067	117,659	55.86
068	05	10	10	574	33,566	57.40
069	21	39	37	2,253	138,075	60.89
070	21	40	36	2,034	117,436	56.50
071	21	40	36	2,027	116,107	56.31
072	21	40	40	2,262	130,484	56.55
073	25	39	38	2,495	166,207	65.66
074	25	37	37	2,381	156,267	64.35
075	40	42	40	2,470	154,448	61.75

TABLE XLV (Continued)

Class	School	No. Enroled	No. Tested (N)	Sum of Scores (Σx)	Sum of Squared Scores (Σx^2)	Mean Score ($\bar{x}$)
076	65	45	44	2,763	178,611	62.80
077	65	07	07	474	33,060	67.71
078	69	32	32	1,762	98,252	55.06
079	76	38	38	2,701	194,857	71.08
080	76	37	37	2,286	142,848	61.78
081	04	34	31	1,832	109,854	59.10
082	30	40	40	2,729	190,061	68.23
083	30	14	13	857	57,475	65.92
084	36	20	17	1,021	61,591	60.06
085	44	29	28	1,893	130,097	67.61
086	49	30	28	1,742	110,822	62.21
087	54	16	16	1,102	76,908	68.88
088	57	33	30	1,967	131,459	65.57
089	67	32	32	2,058	134,106	64.31
090	71	21	21	1,363	90,645	64.90
091	72	24	19	1,163	72,225	61.21
092	06	39	36	2,295	147,795	63.75
093	06	09	09	651	47,325	72.33
094	12	30	27	1,664	104,928	61.63
095	14	34	33	2,171	143,789	65.79
096	33	36	32	2,140	144,318	66.88
097	38	37	34	2,161	138,535	63.56
098	42	47	45	2,803	176,499	62.29
099	47	34	32	1,977	123,575	61.78
100	52	20	19	1,183	74,607	62.26
101	55	23	22	1,435	94,793	65.23
102	62	43	41	2,792	192,472	68.10
103	62	17	16	1,073	73,217	67.06
104	66	36	35	2,147	133,301	61.34
105	66	08	07	389	21,719	55.57
106	02	41	39	2,746	195,550	70.41
107	46	30	29	2,059	148,403	71.00
108	51	39	39	2,609	178,839	66.90
109	51	15	14	958	66,358	68.43
110	58	40	39	2,857	212,279	73.26
111	58	35	34	2,476	181,502	72.82
112	58	30	29	1,846	119,132	63.66
113	60	33	33	2,478	187,918	75.09
114	60	32	32	2,135	143,899	66.72
115	68	24	24	1,531	99,147	63.79
116	78	08	08	432	23,816	54.00
117	81	43	38	2,557	173,723	67.29
118	81	43	40	2,691	182,387	67.28

TABLE XLV (Continued)

Class	School	No. Enroled	No. Tested (N)	Sum of Scores (Σx)	Sum of Squared Scores (Σx^2)	Mean Score ($\bar{x}$)
119	81	42	41	3,006	221,954	73.32
120	81	44	43	2,790	183,684	64.88
121	81	42	39	2,623	177,977	67.26
122	81	42	36	2,566	184,622	71.28
123	81	42	36	2,237	140,949	62.14
124	82	45	45	3,199	229,511	71.09
125	82	44	42	2,764	185,200	65.81
126	82	42	42	2,668	171,438	63.52
127	82	43	42	2,621	164,743	62.40
128	82	37	37	2,377	154,413	64.24
129	83	39	38	2,659	188,389	69.97
130	83	41	37	2,418	159,446	65.35
131	83	40	40	2,416	153,573	61.53
132	83	34	34	2,043	123,929	60.09
133	83	37	35	2,077	124,829	59.34
134	83	36	34	1,994	118,010	58.65
135	84	40	40	2,700	184,498	67.50
136	84	41	38	2,472	162,418	65.05
137	84	39	35	2,137	132,133	61.06
138	84	39	37	2,194	131,962	59.30
139	85	40	38	2,932	227,590	77.16
140	85	40	40	3,014	229,010	75.35
141	85	34	32	2,458	189,854	76.81
142	85	35	34	2,429	175,769	71.44
143	85	43	39	3,146	255,920	80.67
144	85	41	41	3,089	234,649	75.34
145	85	41	40	2,970	222,366	74.25
146	85	41	40	2,865	207,703	71.63
147	85	38	38	2,589	178,417	68.13
148	85	36	31	1,992	130,216	64.26
149	85	38	32	2,092	138,754	65.38
150	85	32	30	1,908	122,810	63.60
151	86	35	34	2,459	179,297	72.32
152	86	38	37	2,546	176,128	68.81
153	86	38	37	2,705	199,821	73.11
154	86	41	37	2,521	173,149	68.14
155	86	42	40	2,644	176,780	66.10
156	86	38	36	2,298	148,418	63.83
157	86	36	33	2,090	133,866	63.33
158	86	33	28	1,646	97,836	58.79
159	87	33	31	1,961	126,943	63.26
160	87	38	34	2,135	136,069	62.79
161	87	31	29	1,684	99,702	58.07

TABLE XLV (Continued)

Class	School	No. Enroled	No. Tested (N)	Sum of Scores (Σx)	Sum of Squared Scores (Σx^2)	Mean Score ($\bar{x}$)
162	87	33	32	1,905	115,433	59.53
163	87	29	29	1,666	97,208	57.45
164	88	40	37	2,249	138,071	60.78
165	88	41	41	2,741	184,819	66.85
166	88	35	33	2,000	122,722	60.61
167	88	36	35	2,253	146,765	64.37
168	89	39	37	2,866	223,382	77.46
169	89	41	38	2,723	197,669	71.66
170	89	42	42	2,818	191,726	67.10
171	89	40	36	2,380	159,412	66.11
172	89	35	32	2,131	143,645	66.59
173	89	36	36	2,299	148,625	63.86
174	90	39	39	2,504	162,844	64.21
175	90	35	33	2,178	145,896	66.00
176	90	37	34	2,156	138,156	63.41
177	90	39	38	2,690	193,356	70.79
178	90	40	37	2,635	188,907	71.22
179	90	40	33	2,361	170,365	71.55
180	90	41	39	2,805	203,861	71.92
181	90	37	35	2,620	198,144	74.86
182	90	38	37	2,643	190,123	71.43
183	90	36	35	2,577	191,955	73.63
184	90	41	36	2,576	186,160	71.56
185	91	41	38	2,789	205,997	73.39
186	91	42	39	2,704	189,404	69.33
187	91	40	38	2,555	173,641	67.24
188	91	40	36	2,368	157,278	65.78
189	91	37	28	1,797	116,94[illegible]	64.18
190	91	39	31	2,008	131,54[illegible]	64.77
191	91	34	30	1,923	125,101	64.10
192	91	36	30	1,775	105,983	59.17
193	92	40	40	2,666	179,638	66.65
194	92	37	37	2,262	140,012	61.14
195	92	39	36	2,171	132,221	60.31
196	92	38	34	2,073	128,005	60.97
197	92	35	34	1,820	98,392	53.53
198	93	35	30	1,946	127,486	64.87
199	93	38	36	2,474	171,828	68.72
200	93	38	34	2,263	151,573	66.56
201	93	34	33	2,067	130,589	62.64
202	93	35	33	2,230	152,556	67.58
203	93	34	32	2,106	141,018	65.81
204	94	36	35	2,717	211,823	77.63

TABLE XLV (Continued)

Class	School	No. Enroled	No. Tested (N)	Sum of Scores (Σx)	Sum of Squared Scores (Σx^2)	Mean Score ($\bar{x}$)
205	94	38	37	2,652	191,090	71.68
206	94	36	35	2,473	176,629	70.66
207	94	35	34	2,357	165,753	69.32
208	94	33	31	1,993	129,323	64.29
209	94	32	28	1,858	124,496	66.36
210	94	30	28	1,704	105,280	60.86
211	95	40	39	2,950	225,428	75.64
212	95	41	41	2,924	210,558	71.32
213	95	42	39	2,725	192,229	69.87
214	95	41	39	2,536	166,964	65.03
215	95	40	33	2,061	129,877	62.45
216	96	35	34	1,926	110,448	56.65
217	96	08	08	477	28,581	59.63
218	96	34	30	1,668	94,506	55.60
219	96	32	31	1,975	127,367	63.71
220	96	37	36	2,196	136,038	61.00
221	97	42	40	3,099	241,125	77.48
222	97	42	41	3,013	224,223	73.49
223	97	43	43	3,045	217,231	70.81
224	97	41	40	2,824	201,178	70.60
225	97	43	42	2,914	205,198	69.38
226	97	42	41	2,862	201,900	69.80
227	97	40	39	2,596	175,298	66.56
228	97	40	39	2,462	157,630	63.13
229	97	34	31	2,001	130,677	64.55
230	97	34	29	1,743	106,201	60.10
231	98	35	34	2,337	162,541	68.74
232	98	33	33	2,426	181,264	73.52
233	98	39	39	2,658	183,782	68.15
234	98	38	33	2,110	136,608	63.94
235	98	33	32	2,109	140,913	65.91
236	98	34	31	1,856	113,106	59.87
237	98	34	32	2,177	149,871	68.03
Total	....	8139	7724	515,463	35,031,605	15,737.50
Mean	....	83.9	79.6	66.74	....	66.40

APPENDIX D

PROCEDURES USED IN SELECTING SAMPLES OF INDIVIDUAL PUPILS

PROCEDURES USED IN SELECTING SAMPLES OF INDIVIDUAL PUPILS

Identification of Pupils by Code Numbers. The original data for each pupil were reported on class record forms. For example, in a school having three classes of grade 8A pupils, the data were recorded on three separate forms--one for each class. Pupils names were listed on the record form in alphabetical order and each pupil was then given a code number representing his alphabetical position within the class. The first pupil name appearing on each class list was coded "01", the second name was coded "02", etc. These pupil codes were punched on the IBM cards along with the class code and the school code. Each individual was thus identified by school, by class and by alphabetical position with the class.

Method Used in Coding Test Scores. Although coded test scores were not used in any of the computations involving sampling by schools or by classes, "grouped" or coded data were used in deriving mean scores for samples of individual pupils. The coding plan called for a three-point step interval. Scores of 42-44 were coded "01", scores of 45-47 were coded "02", and so on up to 90-92 which were coded "17". These codes were punched on the cards along with the actual scores.

The example given below illustrates a portion of the class record for Class 002 in School 19.

ILLUSTRATION OF CLASS RECORD FORM

School 19 Class 002

Name of Pupil (Col. No. on IBM Card) →	Pupil No. (8;9)	Total Score (18;19)	Coded Total Score (20;21)
G.A.	01	65	08
L.A.	02	76	12
....	....	....	....
B.E.	08	83	14
W.E.	09	70	10
....	....	....	....
R.O.	18	70	10
B.P.	19	72	11
....	....	....	....
E.T.	31	71	10
R.W.	32	68	09

Method Used in Drawing a 10 per cent Sample. It will be noted in the illustration given above that "coded score" is punched in Columns 20 and 21 on the cards. The 7724 cards were sorted first by these two columns. The result was 17 different stacks of cards representing the distribution of total scores for all pupils. There were 11 cards in the smallest stack and 996 cards in the largest stack. (See Table I).

The sorting machine was then set to sort by the second digit of the pupil code number. It will be seen from the illustration of the record form above that this is "Column 9" on the IBM card. All cards that fall in pocket "1" on this sort represent pupils whose code numbers are either 01 or 11 or 21, etc. The cards that fall in pocket "2" represent pupils whose code numbers are either 02 or 12 or 22, etc.,-- and so on up to pocket "0" representing pupils whose codes are 10 or 20 or 30, etc. Since there are 237 classes, there will be 237 pupils with the code number "01". There are fewer than 237 pupils with the code "11", however, because not all classes contained as many as 11 pupils.

With the machine set for Column 9, each of the 17 stacks of cards described above was run through separately and the number of cards falling in each of the 10 pockets was recorded in a prepared table. This gave 10 separate distributions of scores, each distribution including roughly 10 per cent of the pupils. Five of the distributions are shown in Table XLVI.

The distributions in Table XLVI thus represent five different groups of pupils obtained by statematic subsampling of individuals within each class. The mean and standard deviation for each sample were computed from the grouped data as shown in the table. There is, of course, some possibility of error in the obtained means and standard deviations due to using a three-point step interval rather than the actual scores as were employed in all computations involving sampling by schools and by classes. It seems extremely unlikely that such possible errors would be large enough to call for any qualifications of the interpretations made in Chapter V.

TABLE XLVI

Scores for Five Groups of Pupils Drawn by Systematic Sub-Sampling of Individuals Within Each Class*

Actual Score	Col. 20;21 (Coded Score)	Column 9 (Second digit of Pupil Code Number) 1 No. Pupils	2 No. Pupils	3 No. Pupils	4 No. Pupils	5 No. Pupils
90-92	17	1	2	1	3	1
87-89	16	2	9	9	9	9
84-86	15	25	12	20	16	14
81-83	14	33	27	31	37	24
78-80	13	59	56	39	39	49
75-77	12	79	77	68	62	68
72-74	11	78	81	75	84	69
69-71	10	84	79	84	70	76
66-68	09	100	116	103	106	105
63-65	08	116	107	104	110	99
60-62	07	79	90	89	83	82
57-59	06	76	79	79	64	68
54-56	05	55	57	63	57	57
51-53	04	41	39	40	37	35
48-50	03	26	12	16	14	15
45-47	02	8	3	6	5	4
42-44	01	...	3	3	1	...
Total		862	849	830	797	775
Mean Score		66.69	66.71	66.30	66.77	66.62
Stand. Dev.		9.30	8.91	9.21	9.13	8.96

*Ten Samples of individuals were actually drawn. Data for samples 6 through 10 are not shown in this table.

BIBLIOGRAPHY

BIBLIOGRAPHY

Black, Bertram J. and Olds, Edward B. "A Punched Card Method for Presenting, Analyzing, and Comparing Many Series of Statistics for Areas." Journal of the American Statistical Association, Vol. 41 (1946), pp. 347-355.

Bowley, A. L. "The Application of Sampling to Economic and Sociological Problems," Journal of the American Statistical Association, Vol. 31 (1936), pp. 474-480.

Chaddock, Robert Emmet. Principles and Methods of Statistics. Boston: Houghton Mifflin Company, 1925. Pp. xvi ≁ 471.

Cochran, W. G. "Relative Accuracy of Systematic and Stratified Random Samples for a Certain Class of Populations." The Annals of Mathematical Statistics. Vol. 17 (1946), pp. 164-177.

__________. "The Use of Analysis of Variance in Enumeration by Sampling," Journal of American Statistical Association, Vol. 34 (1939), pp. 492-510.

Courtis, S. A. "The Derivation of Norms," Journal of Experimental Education, Vol. 2 (1934), pp. 237-242.

Croxton, F. E. and Dudley, J. C. Applied General Statistics. New York: Prentice-Hall, Inc., 1939. Pp. xviii ≁ 944.

Deming, W. Edwards. Statistical Adjustment of Data. New York: J. Wiley and Sons, Inc., 1943. Pp. x ≁ 242.

__________. "On Classifying Errors in Surveys," American Sociological Review, Vol. 9 (1944), pp. 359-369.

__________. "On Training in Sampling," Journal of the American Statistical Association, Vol. 40 (1945), pp. 307-316.

Deming, W. Edwards and Simmons, Willard. "On the Design of a Sample for Dealers' Inventories." Journal of the American Statistical Association, Vol. 41 (1946), pp. 16-33.

Fisher, R. A. *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd, 1946. Pp. 354.

Fisher, R. A. and Yates, F. *Statistical Tables for Biological Agricultural and Medical Research.* London: Oliver and Boyd, 1938. Pp. viii ≠ 90.

Garrett, Henry Edward. *Statistics in Psychology and Education.* New York: Longmans, Green and Co., 1947. Pp. xii ≠ 487.

Good, Carter V., Barr, A. S. and Scates, Douglas E. *The Methodology of Educational Research.* New York: D. Appleton-Century Company, 1936. Pp. xxi ≠ 882.

Hagood, Margaret J. and Bernert, Eleanor H. "Component Indexes as a Basis for Stratification in Sampling," *Journal of the American Statistical Association*, Vol. 40 (1945), pp. 330-341.

Hansen, M. H. and Hauser, P. M. "Area Sampling--Some Principles of Sample Design," *Public Opinion Quarterly*, Vol. 9 (1945), pp. 183-193.

Hansen, M. H. and Hurwitz, W. N. "On the Theory of Sampling from Finite Populations," *Annals of Mathematical Statistics*, Vol. 14 (1943), pp. 333-362.

______________________. "Relative Efficiencies of Various Sampling Units in Population Inquiries," *Journal of the American Statistical Association*, Vol. 37 (1942), pp. 89-94.

Hansen, M. H., Hurwitz, W. N. and Gurney, Margaret. "Problems and Methods of the Sample Survey of Business." *Journal of the American Statistical Association*, Vol. 41 (1946), pp. 173-189.

Hilgard, E. R. and Payne, S. L. "Those Not at Home: Riddle for Pollsters," *Public Opinion Quarterly*, Vol. 8 (1944), pp. 254-261.

Hoel, Paul G. *Introduction to Mathematical Statistics.* New York: John Wiley and Sons, Inc., 1947. Pp. x ≠ 258.

Huffaker, C. L. and Douglass, H. R. "On the Standard Error of the Mean Due to Sampling and to Measurement," *Journal of Educational Psychology*, Vol. 19 (1928), pp. 643-649.

Kelley, Truman Lee. Fundamentals of Statistics. Cambridge: Harvard University Press, 1947. Pp. xvi + 755.

King, A. J. and Jessen, E. J. "The Master Sample of Agriculture: I. Development and Use; II. Design," Journal of the American Statistical Association, Vol. 40 (1945), pp. 38-56.

Lindquist, E. F. "Factors Determining the Reliability of Test Norms," Journal of Educational Psychology, Vol. 31 (1930), pp. 512-520.

__________. Statistical Analysis in Educational Research. Chicago: Houghton Mifflin Company, 1940. Pp. xi + 266.

McNemar, Quinn. "Opinion--Attitude Methodology," Psychological Bulletin, Vol. 43 (1946), pp. 289-374.

__________. "Sampling in Psychological Research," Psychological Bulletin, Vol. 37 (1940), pp. 331-65.

Madow, Lillian H. "Systematic Sampling and Its Relation to Other Sampling Designs," Journal of the American Statistical Association, Vol. 41 (1946), pp. 204-217.

Marks, Eli S. "Sampling in the Revision of the Stanford-Binet Scale," Psychological Bulletin, Vol. 44 (1947), pp. 413-434.

__________. "Selective Sampling in Psychological Research," Psychological Bulletin, Vol. 44 (1947), pp. 267-275.

Neyman, Jerzy. "Contributions to the Theory of Sampling Human Populations," Journal of the American Statistical Association, Vol. 35 (1938), pp. 101-116.

__________. "Contributions to the Theory of Small Samples Drawn from a Finite Population," Biometrika, Vol. 17 (1925), pp. 472-479.

__________. "On Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection," Journal of the Royal Statistical Society, Vol. 97 (1934), pp. 558-606.

Odell, C. W. An Introduction to Educational Statistics, New York: Prentice-Hall, Inc., 1946. Pp. xii + 269.

Peatman, John Gray. Descriptive and Sampling Statistics. New York: Harper and Brothers, 1947. Pp. xviii + 577.

Peters, Charles C. and Van Voorhis, Walter R. Statistical Procedures and Their Mathematical Bases, New York: McGraw-Hill Book Company, 1940. Pp. xvii + 516.

Smith, J. G. and Duncan, A. J. Sampling Statistics and Applications. New York: McGraw-Hill Book Company, Inc., 1945. Pp. xii + 498.

Smith, John N. Tests of Significance and How to Use Them: Studies in Business Administration, Vol. 10, No. 1. Chicago: University of Chicago Press, 1939. Pp. ix + 90.

Snedecor, George W. "Designs of Sampling Experiments in the Social Sciences," Journal of Farm Economics, Vol. 21 (1939), pp. 846-855.

__________. Statistical Methods. Ames, Iowa: The Collegiate Press, Inc., 1946. Pp. xvi + 485.

Strayer, George D. A Report of a Survey of the Public Schools of St. Louis, Missouri. New York: Bureau of Publications, Teachers College, Columbia University, 1939. Pp. xxiv + 468.

__________. Report of a Survey of the Public Schools of Boston, Massachusetts. Boston: City of Boston Printing Department, 1944. Pp. xxxii + 1127.

__________. The Report of a Survey of the Public Schools of Pittsburgh, Pennsylvania. New York: Bureau of Publications, Teachers College, Columbia University, 1940. Pp. xvii + 564.

Tepping, Benjamin J., Hurwitz, W. N. and Deming, W. Edwards. "On the Efficiency of Deep Stratification in Block Sampling," Journal of the American Statistical Association, Vol. 38 (1943), pp. 93-104.

Walker, Helen M. "The Sampling Problem in Educational Research," Teachers College Record, Vol. 30 (1929), pp. 760-774.

Walsh, J. E. "Concerning the Effect of Intra-class Correlation on Certain Significance Tests," Annals of Mathematical Statistics, Vol. 18 (1947), pp. 88-96.

Webb, John N., Northrop, M. Starr and Payne, Stanley L. "Practical Applications of Theoretical Sampling Methods," Journal of the American Statistical Association, Vol. 38 (1943), pp. 69-77.

Wilks, S. S. "Representative Sampling and Poll Reliability," Public Opinion Quarterly, Vol. 4 (1940), pp. 261-269.