Popham, W. J. (1975). Sampling strategies. In W. J. Popham *Educational Evaluation* (pp._218-233). Englewood Cliffs, NJ: Prentice-Hall.

CHAPTER TEN

Sampling Strategies

Everyone knows what sampling is. Almost all of us, in fact, have been involved in one or more sampling operations such as being interviewed by Gallup or Harris pollsters. Further, most people recognize why anyone adopts sampling procedures. By sampling a population of American voters, we can get an idea of how the whole group is apt to behave. Sampling should make a Scotchman's values vibrate. It is so terribly thrifty.

SAMPLING FOR RESEARCH

An educational researcher engages in sampling because the researcher wishes to generalize a set of findings based on the sampled data to a larger population. This is only natural because researchers are in the business of making generalizations. They want to discern the nature of relationships between variables. The more generalizable the finding, the better. A researcher's dream, in fact, would be to identify such a strong and generalizable relationship among a set of variables that predictions could confidently be made about one variable on the basis of another.

The educational researcher will often expend considerable energy on defining a population clearly, then sample from it carefully, since it is to that population that generalizations from the sample are made. Drawing a representative sample from that population, such as is used when one conducts an educational experiment on a small number of learners, is obviously important. If the sample is not truly representative of the population, the generalization from sample data to the larger population is clearly unwarranted.

SAMPLING FOR EVALUATION

Evaluators also will also engage in sampling operations. In many ways their reasons for doing so are identical to those of the researcher, but on

A far more limited scale. The "population" of learners with which evaluators are concerned is typically more modest. Such a population might be, for example, all the high-school seniors in a particular school district. The reasons that the evaluator engages in sampling are more often associated with questions of economy and practicality than of generalizability. There is a sense in which the educational evaluator needs to generalize to a population of future students who might interact with a specified instructional program. This point will be discussed in chapter eleven.

In many educational situations there is no feasible way for the evaluator to gather large amounts of data from each child. Teachers and administrators would balk, for example, if an evaluator asked for several hours of testing time each week from each pupil. Whether warranted or not, most educators believe that this much time away from their normal instructional endeavors would irreparably-damage pupils.

But evaluation is a data-based enterprise. Evidence regarding the status of learners (such as before and after instruction) is viewed as indispensable to well-conducted evaluations. Yet, because evaluators cannot intrude excessively on the ongoing school program, they must gather their data as economically as possible. Sampling is the answer.

Two major approaches to sampling will be described in this chapter. The first of these will consist of conventional sampling techniques used for several decades, most notably by survey researchers but also by experimentally oriented investigators. The second is a much newer approach to sampling, an approach ideally suited to a number of educational evaluation applications.

Customary Sampling Techniques

The most commonly used sampling procedures are typically those that, within the constraints of the situation at hand, offer the most likelihood of selecting a portion of a population that will be representative of that population. Obviously, not every situation will provide the evaluator with total freedom to choose any of the many sampling procedures available. Sometimes compromises will have to be made. In the past, educational researchers have generally opted for one of the following sampling plans.

SIMPLE RANDOM SAMPLING

A simple random sample is one in which each individual in the population has an equal chance of being included. Because of this equality of opportunity for inclusion in the sample, random sampling offers an excellent way to reduce the likelihood of a seriously unrepresentative population sample. We often refer to a *simple* random sample, not to denigrate the sample as unintelligent, but to signify that no additional sampling operation (such as stratification) is involved.

Although there are several ways of securing a random sample, the most popular is to employ a table of random numbers. Such tables will be found in the appendixes of most statistical texts. These tables typically consist of an extensive series of five-digit numbers, randomly generated by a computer, such as these:

10478
75510
22039
65959
64819
68531
09599
80130

To use a table of random numbers in constituting a simple random sample, one assigns sequential numbers to all members of the population at hand, then employs the tabled numbers to select the proportion of subject desired. To illustrate, suppose we had a group of 90 first-grade pupils from which we wished to select a sample of only 15. We would first number all our pupils, 01 through 90, then enter the table of random numbers at any point (from top to bottom, bottom to top, sideways, etc.) and pick the first 15 numbers which appeared. If we had started at the upper left two columns of numbers in this instance, we would select the students whose numbers were 10, 75, 22, 65, and so on. Numbers 91-99 (and 00) would be ignored.

Another commonly recommended technique for constituting a random sample is to place the number of all subjects on disks (statisticians appear to have access to an unlimited supply of disks), thoroughly mix the disks, then select the disks from a container until the appropriate number of disks has been chosen. Because of the imprecision associated with disk-mixing by hand (few mechanical disk-mixers have been financially successful), this procedure is less appropriate than employing random-number tables.

One of the writer's less-talented, but no-less-creative graduate students once suggested a variant on this procedure. When asked, on the final examination in a statistics class, to describe a common sampling procedure, she responded as follows.

> First *you* take small slips of paper and put every subject's name on them. Then *you* put all of the slips in a brown paper bag. Next you blindfold *yourself. Now* shake the bag *vigorously.* Then while still blindfolded, *you pull* out the *appro-*priate number of *slips. This* gives *you* an unbiased *rambling* sample.

As one might guess, sampling experts have never been sufficiently excited by this suggestion to substitute rambling for random samples in their recommended procedures.

STRATIFIED RANDOM SAMPLING

Although simple random sampling can provide us with representative samples, we can add greater precision to our sampling estimates by drawing a *stratified random sample.* Stratified random sampling utilizes supplementary information about the population in order to draw samples that are more apt to be representative of that population.

The population is first divided into any subpopulations whose nature would appear to be relevant in some way to the measures we are making. In other words, don't ritualistically subdivide a population into age, sex, and socioeconomic subgroups unless there is some reason to believe that these dimensions are relevant to the things you're measuring. This assumes, of course, that you have access to the information that will permit you to make these decisions. For example, you would have to know that in the population of learners under consideration there are approximately 47 percent males and 53 percent females. These subpopulations are called *strata* and from each *stratum* (singular) is typically drawn a proportional number of subjects by random sampling. Drawings are made independently in each of the strata. The strata selected, incidentally, should be mutually exclusive and exhaustive.

To illustrate the application of stratified random sampling procedures in an evaluation context, we might think of an "end-of-instruction attitude toward school" measurement that we wished to secure for an entire school district. Because there are many pupils in the district (10,000), the evaluator decides to draw a sample, thereby conserving both the students' time and the money required to administer and score tests. To make the sample more accurately representative of the entire district than a random sample might be, the evaluator decides to employ a stratified random sample. First, the variables in the population of potential relevance (to attitudes toward school) are isolated. Suppose previous studies in the district have suggested that ethnicity seems to be related to learners' attitudes toward school. Accordingly, the evaluator checks the central office to determine that the ethnic percentages of all schoolchildren in the district are as follows:

Anglo	42%
Chicano	21%
Black	20%
Oriental	11%
Others	6%

In constituting a 500-student stratified random sample of the district's 10,000 pupils, the evaluator then determines that 210 (42 percent) will be Anglo children, 105 (21 percent) will be Chicano youngsters, and so on. The pupils for each stratum, their number predetermined by the percentage of pupils in the population, are drawn randomly. The result is a more representative sample than would have been derived from simple random sampling, ignoring the ethnic strata. If we wish to make statements about all strata, then we may want to over sample the smaller groups and, in forming estimates for the total group, reduce the contribution of such over sampled strata accordingly.

It is possible, of course, to add additional strata to the sampling plan. If the stratified sampling plan becomes very complicated, the evaluator would be well advised to consult one of the references on sampling cited at the close of the chapter. Another alternative is to secure the services of a sampling expert who can aid in the design of a complex sampling plan. It should be noted that when we refer to a sample as a "simple random sample" or a "stratified random sample" we are really describing one *procedure* by which all samples are computed, not a characteristic of the sample. One can't tell whether a sample is random merely by looking at it--even looking at it closely. Because simple random sampling does not guarantee a sample that will be representative in terms of some important characteristic(s), stratified random sampling (which does) is seen as a more refined method of sampling.

CLUSTER SAMPLING

In *cluster sampling* the unit of sampling is not the individual but a naturally occurring group of individuals. This procedure is often used

when it is more convenient to select preformed groups of individuals than it is to select individuals from a population. To illustrate cluster sampling, imagine that you are attempting to draw a 300-pupil sample of ninth-grade English classes from a large district with 100 ninth-grade English classes. each of which has 30 pupils per class. To get the sample, you could draw 300 youngsters randomly from the 3000 pupils in all 100 classes. But, if that process is impractical because of the measurement interruption to all 100 classes, you might randomly select 10 of the 100 classes, thereby picking up your sample of 300 learners.

The disadvantage of cluster sampling is that it sometimes yields a less-accurate estimate of population's performance than would he provided via a random sampling scheme. On the other hand, the savings in time and money offered by this approach are almost always attractive trade-offs for the evaluator.

The decision to use cluster rather than random sampling often hinges on the unit of analysis to be employed when the data are analyzed. Recalling the brief discussion in chapter nine, there are instances when the appropriate unit of analysis must be the classroom, that a single average of pupil performance calculated for each class of pupils involved in the evaluation study. If this were the case, instead of having an n of 10, it might be preferable to randomly draw ten students from each of 30 randomly selected classrooms, then compute an average for each of the 30 classes lasses (represented by the ten pupils). If the pupil, rather than the classroom, is the appropriate unit of analysis, then cluster sampling is an economical way for the educator to round up a sample.

SYSTEMATIC SAMPLING

A convenient form of sampling involves drawing the subjects for the sample *systematically* from a list of the population. For example, if a 100-learner sample is to be drawn from a list of 1000 learners, then the evaluator might *randomly* decide which number (one through ten) to start with, then draw every tenth person from the list thereafter. For instance, if the number eight was randomly drawn, then the *systematic* sample would be constituted by the 8th, 18th, 28th, 38th (and so on) learners. Systematic sampling can be used instead of random sampling if the evaluator is certain that the population list is in a reasonably aperiodic order. If there is any possibility that there is *periodicity* in the list, i.e., if every *nth.* person on the list shares a characteristic not shared by others on the list, then simple random sampling should be preferred.

SAMPLE SIZE

The question of how to determine sample size is one of the evaluator's most difficult problems. The key, of course, is to draw a sample large enough so that those making the educational decisions will have confidence that the sample results are truly indicative of those that would be present had the entire population been measured.

A better way to put this might be to say that we must determine the uncertainty in the estimate that one would tolerate before changing a decision For example, would the decision be different if our estimates were off by *x* percent or *y* points? At the point at which the decision changes, sufficient precision to be sure we don't make "wrong" decision because of sampling variability. Looking at the problem this way, evaluators can often get by with much smaller samples than they might normally think.

Often, therefore, the evaluator will have to gauge the decision-maker's receptivity to various sample sizes. Suppose, for example, that on the basis of some small-scale statistical analyses it was determined that a sample of 100 would yield a reasonably accurate estimate of the student population's preferences regarding a particular school program. Yet, in telling the members of the school board (who would use the report to reach a program revision decision), it became apparent that they would be convinced only by a 200 or larger sample. If unable to persuade the board that the smaller sample would do the job, the evaluator ought to opt for the sample size that will make a difference, in this case the sample of 200.

Sometimes it is possible to gather measures for an entire population, then randomly select samples of different sizes from the population, with the purpose being to demonstrate (for future evaluations involving samples) what kind of accuracy is yielded by samples of varying sizes. For instance, in the previous example the evaluator might have been able to persuade the school board that the 100-student sample would be sufficient if it could be demonstrated that the mean performance for the entire population was 81.2 percent and three successively drawn random samples of 100 pupils fell within 3 percentage points above or below the population mean.

It is possible, as we shall see, to compute *confidence intervals,* which serve as ranges which (on the basis of sample data) span the "true" population performance at specified probability levels. This procedure will be examined in more detail in chapter eleven.

SAMPLE PRECISION

It is apparent that although we would prefer to be perfectly precise in estimating population data from our sample data, there is a degree of imprecision associated with sampling. An expression often used in connection with sampling operations is *standard error,* which provides an estimate of the possible magnitude of error present in some obtained statistic. The larger the standard error, the less confidence we have in the precision of our sample-based statistics. The smaller the standard error, the more likelihood that our sample estimates are precise reflectors of values in the population. Measures based on *samples,* such as the mean or standard deviation, are often referred to as *statistics.* Those same measures, when computed for the entire *population,* are referred to as *parameters.* It is important to determine how closely our sample's statistics match the population's parameters.

We can illustrate the meaning of standard error of measurement by referring to figure 10.1 in which two sets of score distributions are presented, where the *scores* are really *means* obtained by repeatedly sampling the population. Each of the distributions (A and B) has been produced by plotting the means yielded from each of 500 separate sampling operations. The population mean is, as seen, 20. Now, note distribution B, which results from sampling procedure B. See that its distribution is much more spread

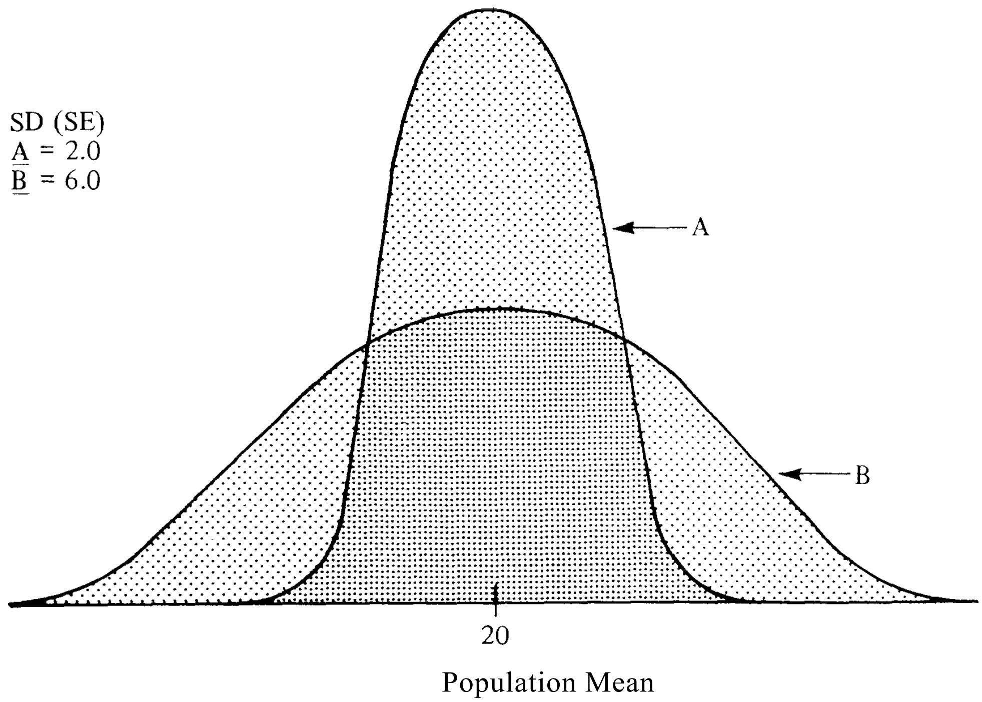

Figure 10.1 Fictitious distributions of sample means based on 500 replications of two different (A and B) sampling schemes.

out than that of distribution A. The standard deviation of distribution B is 6.0, reflecting the fact that it is more dispersed than distribution A, which has a standard deviation of only 2.0. This indicates that a greater number of the sample means drawn via sampling procedure B missed the mark (the population mean of 20) by any given amount than did those produced from procedure A. Sampling procedure A is the more precise.

The standard deviations of those two distributions are referred to as standard errors and it can be seen, therefore, that the smaller the standard error of a sample statistic, the greater confidence we can have in its accuracy.

We don't have to draw 500 samples to get an estimate of the sample's standard error. This can be calculated on the basis of a single sample. But it is a useful concept to employ when dealing with various types of sampling procedures. For example, cluster sampling typically yields a somewhat larger standard error than a simple random sample of the same size drawn from the same population. As we shall see in the next section, there are other sampling techniques that can shave the standard error even more.

Matrix Sampling

During recent years a new approach to sampling, which can be of particular use to the educational evaluator, has been devised. Known as *matrix* *sampling,* this procedure was developed chiefly by Frederic Lord[1] who in the mid- and late fifties authored a series of theoretical papers regarding the approach. In 1965 Lord spelled out even more clearly the rationale and potential advantages of matrix sampling.[2] Since that time a number of other psychometricians have advanced the technical procedures needed to employ matrix sampling methods more effectively.[3] As with most recently devised methodologies, the technology associated with matrix sampling is not yet complete. A number of difficult problems are now being wrestled with by matrix sampling methodologists. In some cases the problems are winning. The evaluator interested in applying matrix sampling procedures will want to consult current articles and monographs on this technique. Otherwise, in this fast-moving field, there is a danger that outmoded techniques will be inadvertently employed.

SAMPLING EXAMINEES AND SAMPLING ITEMS

In the conventional sampling procedures described earlier, it was apparent that we were selecting different individuals (or, in a measurement context, *examinees)* from a population. What Lord and his associates recognized is that for purposes of estimating a group's performance on a particular measure it is not necessary to have all the sampled examinees complete all of the items in a test. If this were done, that is, if every examinee completed every item, we would refer to it as *census testing.* Yet, not only can one engage in *examinee sampling* (only some examinees complete all items), but in *item sampling* (only some of the items are completed by all examinees) as well. For many people this is a revolutionary idea. We have been nurtured on the idea that sampling involves the selection of certain individuals who, having been selected, typically complete an entire test. After all, if such examinees didn't complete the entire test, how could we determine their score on the whole test? It's tough to do. But sometimes we don't need to.

The educational evaluator is often faced with situations where the task is not to establish how individuals perform on a test but, instead, to determine how a group performs. Indeed, the evaluator is characteristically more interested in the quality of educational programs as reflected by their impact on a group of examinees. Learner data, as noted previously, are gathered from individuals but treated collectively. To secure accurate estimates of a group's performance, it is absolutely unnecessary to have individuals complete the entire test. In fact, with a relatively large group of learners the most accurate way to secure an estimate of the group mean is to administer one test item to each examinee.

The writer has used this procedure for a number of years in large lecture classes at UCLA. When wishing to evaluate the effectiveness of 30-minute

[1] F. M. Lord, "Estimating Norms by Item Sampling."
[2] F. M. Lord, *Item Sampling in Test Theory and Research Design.*
[3] D. M. Shoemaker, *Principles and Procedures of Multiple Matrix Sampling. Also* K. Sirotnik, "An Introduction to Matrix Sampling for the Practitioner."

tape-slide instructional programs on a pretest-posttest basis, I would devise criterion-referenced tests of 10-20 items, one test per program, then have each of the items printed on a 3-by-5-inch card. Having shuffled the cards adroitly with all the flourish of a Las Vegas blackjack dealer, I would distribute them, *one card at a time,* to the 150-200 students enrolled. No one objected to these 30-second tests, since they took such a small bit of time. Yet, when I assembled the data collectively, they gave me a good idea of how the total class could perform before and after the instructional program. For each item, I typically had 10-20 individual responses, enough in fact to supply some evidence regarding how the entire class might perform on the item.

Because it is convenient to consider one's examinee-sampling and item-sampling options at one time, research specialists often display their data in a box or matrix such as that seen in figure 10.9, where the horizontal

Test Items

Examinees	1	2	3	4	5	6	etc.
1	1	0	0	1	1	0	
2	1	0	1	0	1	1	
3	1	0	1	1	0	1	
4	0	1	1	0	1	1	
5	1	1	0	1	1	1	
etc.							

Figure 10.2 *An* illustrative matrix representing examinee performance on test items.

axis refers to test items and the vertical axis refers to examinees. The zeros and ones in the matrix indicate whether the examinee answered each item correctly (1) or incorrectly (0). Different gradations of responses could also be used, such as is frequently seen with affective measures (3 = agree, 2 = neutral, 1 = disagree).

MULTIPLE-MATRIX SAMPLING

Now imagine a larger data matrix such as that seen in Figure 10.3 where we have 50 items across the top and 100 examinees on the vertical axis. It would be possible to sample some examinees and some items simultaneously.

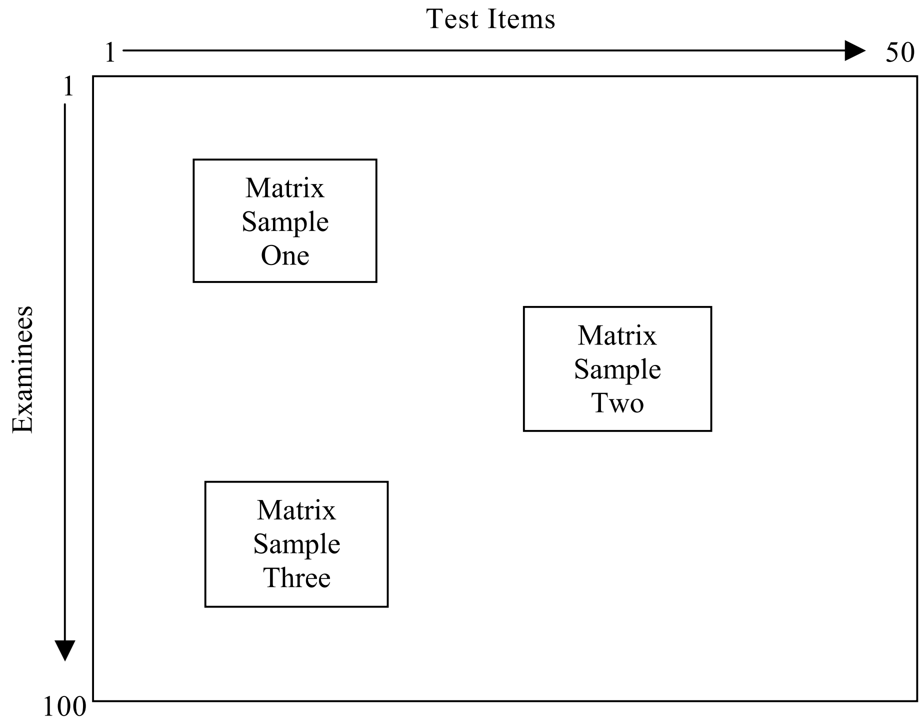

Figure 10.3 An illustrative data matrix reflecting multiple-matrix samples.

This results in an *item-examinee* sample, customarily referred to as a *matrix sample.* To illustrate, suppose we had randomly identified individuals 1, 4, 28, 37, 42, 50, 76, and 80 to represent our examinee sample. Then suppose that we randomly select items 4, 9, 17, 26, and 37 to represent our item sample. By having the selected examinees complete the selected items we constitute a matrix sample. If we repeated this process we would secure a second matrix sample. Several such matrix samples are symbolically represented in figure 10.3, although in reality if we sampled the examinees and items cited above the matrix, samples would be scattered out rather amorphously.

By repeating the initial operation several times, that is, drawing new item-examinee samples, we can constitute *a multiple-matrix sample.* Multiple-matrix samples are significant because they provide the most accurate estimate of population parameters of all the sampling procedures we have considered thus far. For example, if we were to compute the means for ten separate matrix samples used in a multiple-matrix sampling approach, then average out the ten separately derived sample means, the resulting value would be our best bet for estimating the population's mean.

There is some disagreement among matrix-sampling methodologists regarding whether the samples that constitute matrix sampling should be drawn *with* or *without replacement.* Most people appear to favor sampling without replacement, that is, once items or examinees are assigned to one matrix sample they are no longer available for assignment to another matrix sample.

ADVANTAGES OF MATRIX SAMPLING

Particularly for purposes of educational evaluation there are some advantages associated with matrix sampling that should be noted.

Reduced Testing Time per Student. Since today's educational evaluator relies very heavily on measurement data, typically measurement data derived from test administrations, any procedure to reduce the time needed to test learners is a boon. Matrix sampling really fills this need. Because not all examinees need be tested and those who are tested need not complete all the items, matrix sampling represents a considerable economy for evaluators, saving the testing time taken away from an individual learner's instructional activities and the supervisory time required of those administering the tests. Overall, however, more students must be tested or more items must be administered to achieve the same level of precision as if we were using sampling procedures other than matrix.

Smaller Standard Error of Estimate. It has been demonstrated both algebraically and empirically (post-mortem sampling for known populations to verify the accuracy of different sampling schemes) that *multiple-matrix sampling will* generally yield a smaller standard error of estimate than either examinee sampling, item sampling, or a single item—examinee or matrix sample. This assumes, of course, that the numbers of observations[4] are comparable in the different sampling approaches. We would almost always get a more accurate estimate from, for example, simple random sampling (a form of examinee sampling) if we used a large sample of the population—as opposed to a very small multiple-matrix sample. But, if the magnitude of the data gathered is comparable, multiple-matrix sampling gives us the best estimate available.

Measuring Large Item Domains. In some cases the number of test items constituting a domain will be so numerous that if we want to know how a group performs on the total domain there is no alternative to matrix sampling. For instance, suppose a foreign-language teacher wishes to test a class' knowledge of 600 vocabulary words. It would be unlikely that any procedure other than matrix-sampling could be practically used to secure information about all the words.

Less Threatening to Individual Examinees. There is plenty of evidence that some people become so threatened by the prospect of being tested that they perform below their capabilities on the examination. Since, in a matrix-sampling setting, it is obvious that most examinees cannot be compared (most of them are completing different sets of test items), those individuals who freeze under typical test conditions may thaw out a bit.

[4]Defined as the product of the number of examinees tested and the number of responses per examinee.

LIMITATIONS OF MATRIX SAMPLING

Not all is euphoric, however. There are some difficulties with matrix sampling that should be weighed along with its positive features.

Performances of Individuals Unknown. Although for many evaluation applications there is no need for data regarding individual learner performance, there are instances when it would be useful to have such data. For example, certain states are moving to a matrix sampling statewide educational assessment plan. This scheme provides the statewide decision-makers with the requisite information, but a good many teachers find themselves yearning for some sort of pupil-by-pupil test scores. The evaluator will have to reckon with the fact that no individual learner-performance data will be available from this group-oriented sampling procedure.

Logistics of Administering Multiple Tests. There is something incredibly straightforward about having every person complete every item in a test. Everybody gets the same test form. Every test is scored with an identical scoring key. Even students who lean toward cheating find real advantages in having all examinees receive the same tests. Those evaluators considering the use of matrix sampling procedures should recognize that the juggling of many different test forms—their administration, scoring, and assimilation—represents a nontrivial problem. For the very young learners (or foreign students), whose lack of reading skills requires tests to be administered orally, the time required to administer tests via matrix sampling is awesome. As our experience base with matrix sampling is expanded, we shall undoubtedly learn how to avoid some of these logistical problems, but for the moment they represent a distinct deficit for matrix-sampling enthusiasts.

Context Effects for Speeded Tests. There have been several studies conducted in which the way an examinee responds to an item in a total set of test items is contrasted with the way the examinee responds to items in a matrix sample. If there are substantial differences in the examinee's performance under the two conditions, then *context* effects are said to be present. For unspeeded tests, that is, those in which there is no maximum time limit set on student performance, there appears to be little or no context effects associated with matrix sampling items. For speeded tests, however, context effects logically appear to be operative. Hence, matrix sampling should typically not be used for speeded tests.

Data-Processing Requirements. For certain kinds of data-processing operations involving matrix sampled data the evaluator must employ electronic computers to treat the data. Although for the bulk of evaluation requirements these more sophisticated analyses will not be needed, the evaluator who contemplates use of matrix sampling should make sure that the requisite data-processing equipment is available. An abacus is usually not sufficient.

Impersonal Measurement and Optimal Performance. One advantage of matrix sampling is that individuals may not feel threatened by the items they are obliged to complete, inasmuch as the fact that other examinees complete other items renders comparisons impossible. But this positive feature has a negative flip side. Some people only perform their best when they have a personal stake in the testing situation. There is a possibility, therefore, that some examinees who sense the essentially noncomparative nature of matrix sampling may fail to perform at their optimal levels. This is an unstudied question thus far.

ON BALANCE

Although the application of matrix-sampling methodology is not without its problems, a consideration of its pros and cons would strongly suggest that there are a number of situations in which the evaluator will find it strategically advantageous to use this technique. Its time-saving for any one examinee, the ability to sample a domain widely, and its reduced standard error of estimate are particularly compelling advantages.

As observed previously, it is important for those who would employ matrix-sampling procedures to keep abreast of the technical developments associated with the procedure. The step-by-step procedures needed to decide, for example, how many examinees, items, or matrix samples to select are beyond the scope of this treatment. More technical references should be consulted. Used as a substitute for, or in connection with, some of the other sampling procedures described in this chapter, matrix sampling can prove a useful ally for educational evaluators.

Discussion Questions

1. Why do evaluators engage in sampling? Are their reasons for sampling basically different than those of educational researchers? If so, how?
2. Can you think of educational situations, not necessarily those associated with evaluation endeavors, in which different kinds of sampling procedures *would* be appropriate? What are the distinguishing features of such situations?
3. How would you present the major strengths and weaknesses of matrix sampling if you were reporting to a school board regarding this new sampling procedure?

Practical Exercises

1. Ten descriptions of sampling procedures follow. Decide which of the following descriptions is most appropriate for each:

simple random sampling	item sampling
stratified random sampling	matrix sampling
cluster sampling	multiple-matrix sampling
systematic sampling	census testing

a. The evaluator studies the status of 10 percent of the state's pupils by randomly sampling 10 percent of the state's school districts—measuring every pupil within each district sampled.

b. *A* class of 50 pupils is randomly assigned 25 of 500 spelling words to see whether they can properly spell the words after oral presentation by the test administrator.

c. Bob Harris, a high-school principal, draws a 20 percent sample of his school's student body by selecting every fifth name from the school's master roll book.

d. The 50 students whose I.D. numbers appear first in a table of random numbers are included in the sample.

e. The data base for this evaluation project involves all examinees. Each examinee is required Lo complete the entire test instrument.

f. In this random sample a provision is made to ensure proportional representation of students representing different socioeconomic status groups.

g. This sampling plan is accurately described as a single-item—examinee sample.

h. This sampling plan can be described as a series of item--examinee samples.

i. Pupils' names are placed on cardboard disks, placed in a box, then selected until the proper size sample has been chosen.

j. Some of the examinees and some of the items are randomly identified, then those selected examinees complete the selected items.

Answers to Practice Exercises

a. Cluster sampling; *b.* item sampling; c. systematic sampling; *d.* simple random sampling; *e.* census testing; *f.* stratified random sampling; *g.* matrix sampling; *h.* multiple-matrix sampling; i. simple random sampling; j. matrix sampling.

Selected References

COCHRAN WILLIAM *G. Sampling Techniques.* New York: John Wiley, 1953.

COOK, D. L., and D. L. Stufflebeam. "Estimating Test Norms from Variable Size Item and Examinee Samples." *Journal of Educational Measurement* 4 (1967): 27-33.

FELDT, L. S., and R. A. FORSYTH. "An Examination of the Context Effect in Item Sampling." *Journal of Educational Measurement* 2 (1974): 73-83.

HUSEK, T. R., and K. SIROTNIK "Item Sampling in Educational Research." Center for the Study of Evaluation, Occasional Report no. 2. Los Angeles: University of California, 1967.

JAEGER, RICHARD M. *A Primer on Sampling for Statewide Assessment.* Princeton, N.J.: Educational Testing Service, 1973.

JOHNSON, PALMER O. "Development of the Sample Survey as a Scientific Methodology." *Journal of Experimental Education* 27 (1959).

KNAPP, T. R. "Item Sampling." Unpublished monograph, 1973.

LORD, F. M. "Estimating Norms by Item Sampling." *Educational and Psychological Measurement* 22 (1962): 259-67.

———. *Item Sampling in Test Theory and Research Design.* Princeton, N.J.: Educational Testing Service, 1965.

LORD, F. M., and M. R. NOVICK *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison-Wesley, 1968.

SHOEMAKER, D. M. *Principles and Procedures of Multiple Matrix Sampling.* Cambridge, Mass.: Ballinger, 1973.

SIROTNIK, K. "An Introduction to Matrix Sampling for the Practitioner." In *Evaluation in Education: Current Applications.* Berkeley, Calif.: McCutchan Publishing, 1974.

———. "An Investigation of the Context Effect in Matrix Sampling." *Journal of Educational Measurement* 7 (1970) 199-208.

———. "Matrix Sampling in Educational Evaluation: Using the Standard Error of the Mean." Los Angeles: Instructional Objectives Exchange, 1973.

YATES, FRANK. *Sampling Methods for Censuses and Surveys.* London: Charles Griffing, 1949.

Instructional Aids

An Introduction to Matrix Sampling, Ken Sirotnik. American Educational Research Association, Washington, D.C. Training Tape, Series C. Audiotape.