First publ. in: Handbook of Curriculum Evaluation / A. Levy (ed.). New York: Longman, 1977, pp. 258-295

12

Collecting and Analyzing Evaluation Data

This chapter deals with four major issues related to collecting and analyzing data. It is not intended to be a systematic guide for researchers; it is meant to serve those who already possess a basic knowledge of methodological problems of educational research. The chapter focuses on those unique features of data collection and data analysis that are characteristic for curriculum evaluation and are not treated satisfactorily in basic textbooks.

SAMPLING

In curriculum evaluation we are usually dealing with a grade group embracing thousands of children in hundreds of schools. It is in practice impossible, for example, to check the curriculum materials and procedures on all children, but yet we need to have some idea of how they would work on all children. We therefore select certain schools and children that resemble as closely as possible all schools and children. This small number of schools, classes, or children is known as a *sample.* What we need is evidence on how closely the sample mirrors the population, i.e., how large or small the error of sampling is.

This chapter was written by Michael Bauer, Wilhelm F. Kempf, Arieh Lewy, and T. Neville Postlethwaite.

The "Target Population"

The first prerequisite in any sampling procedure is that the group of persons to be sampled must be precisely described. The group of persons to be sampled is known in technical language as the *target population.* Let us take an example. If we say that the target population is "all ten-year-old children in the country," we must be careful to specify what we mean. Do we mean all ten-year-old children whether they are in school or out of school, or do we mean ten-year-old children in full-time schooling? If a school system has a grade-repeating system and if most ten-year-olds are in, say, grade 5, there will still be some ten-year-old children in grades 3, 4, and 6. Do we want to identify only the ten-year-olds in grade 5? Thus, it is possible to identify a target population and at the same time to agree to exclude certain children from that target population. In this case the target population is redefined to create a "sampled population," but we are very clear as to the types of ten-year-olds who have been excluded from the target population (in order to identify some of the population). We may then proceed to sample the sampled population and collect our data, but we know from experience that not all people in the sample will respond and therefore there will be a shortfall. The information we have at the end of our sampling consists of replies from what is often termed the "achieved sample." Whatever data we present, we must be clear as to what is the nature of the population that replied.

The definition of the target population is not always as easy a matter as it might seem, and it is necessary for the curriculum developers to agree among themselves (often requiring fairly lengthy discussions) as to the exact nature of the target population—whether this be of students, teachers, or of subsections of society.

The Sampling Unit

In some evaluation exercises we need to know the achievement levels reached by *each* student, or the opinions of *individual* politicians, or the suggestions of *individual* teachers. In other studies we may wish to know the performance of *classrooms* of students, or *subsets* of politicians, or different *categories* of teachers. In the first case, the unit of sampling will be the individual student, the individual politician, or the individual teacher; in the second instance, the unit of sampling will be the class, the subgroup, or the category. Thus, in order to draw a sample we would need in the first instance a list of every *student* in the target, or sampled population, whereas in the second we would need a list of all the *classrooms* of students in the population. The unit of sampling is highly related to the objectives to be evaluated and to the unit of statistical analysis to be used.

In curriculum studies where a method of instruction or a learning unit that is highly dependent on the teachers' manual is being tested, it is clear that it is the classroom (i.e., the aggregate of pupils within a class) that should be the unit of analysis and hence the unit of sampling.

On the other hand, if one is interested in identifying different subgroups of students mastering or not mastering particular objectives, the student will be the unit of analysis. In this case a sample of schools would be drawn, and a subsample of students in the target population within the sample schools would be drawn. Although it is often administratively advantageous to test whole classes within the sample schools since this causes little disruption in the school, there are two major disadvantages in doing this. First, it is cheaper to subsample pupils, and the sample will not be more representative if the whole class is taken than if a subsample of pupils is used. Second, if several classes in the school fall into the target population and if the classes are streamed, i.e., grouped by ability so that the level of classes differs, then a great deal of care has to be taken to ensure that the range of the "ability" classes is drawn from all the schools with more than one class where streaming takes place. This is a complex undertaking, and it is wiser to take a random subsample of students from across all the classes within the school.

Judgment Samples

One common occurrence in curriculum evaluation is to use judgment samples. This is a sample drawn by the use of judgment. In chapters 2, 3, 4, and 5 the use of judgment samples was suggested for polling the opinion of teachers, parents, industrial enterprises, and the like concerning the details of educational objectives, materials, and procedures. In the tryout of materials and methods, small judgment samples of six to eight classrooms are suggested.

As we have seen, the target population must be carefully defined. Let us assume that we wish to have a first tryout of some physics materials in seventh grade. We also know that there is considerable variance between schools in the social-class distribution of children attending the schools. From the hearsay of inspectors and teachers we know that certain schools are considered good (in that the pupils in those schools generally perform well in all subjects) and some schools are considered poor. More systematically, inspectors will be able to categorize schools into good, medium, and poor. Thus, using this categorization, the curriculum center personnel can select two or three classes from each group (good, medium, poor) for tryout purposes. As the curriculum center needs more and more schools for different tryout purposes, it should collect more detailed information

Table 12.1. **Grouping Schools According to Achievement Level**

Score Interval	*Code Numbers of Schools*
120–130+	617, 819, 403, 192, etc.
110–119	812, 414, 323, 216, etc.
100–109	798, 412, 375, 142, etc.
90–99	089, 917, 243, 172, etc.
80–89	014, 019, 961, 439, etc.

on the performance level of each school. The curriculum center will then be in a position to keep up-to-date lists of all schools and a finer differentiation will be possible, e.g., excellent, very good, good, medium, poor, very poor, exceptionally poor.

Where standardized tests are administered in all schools at regular intervals, for example, for national norming purposes, the overall range of performance in the country will be known, as well as the average performance of each school. This gives very detailed information and, in such cases, it is possible for the curriculum center to select the number of schools it requires in certain intervals along the range. Thus, for example, if the achievement results range from 80 to 130 (on some sort of standardized test), a table such as table 12.1 can be compiled.

The code numbers with the name and address of the corresponding schools are usually kept in a separate listing, but this makes it easy for the curriculum personnel to select their schools according to ability levels. If other criteria are associated with different levels of schooling (e.g., ethnic or religious or social grouping), it is also possible to categorize schools on this basis and use some combination of such criteria for selecting schools. However, it is judgment that is being used to select schools.

Quota Sampling

Quota sampling is a more refined form of judgment sampling. Again, this is frequently used for selecting samples of persons in the polling of opinions or attitudes. Throughout chapters 3–8 and also in chapter 11 there has been much discussion of gathering evidence of the opinions of teachers and parents on the appropriateness of certain educational objectives, certain materials, or certain teacher or pupil activities. Let us assume that we need to sample seventh-grade teachers, but that we have reason to believe that the opinions of these teachers will vary according to the sex and age of the teachers and according to the regional location of the schools. As a simple example let us assume that there are five main regions in the country. These variables characterizing the teachers are often called

Table 12.2. **Schema for Stratified Sample**

Age	*Sex*	*Regions*				
		1	*2*	*3*	*4*	*5*
20–29	M	x	x	x	x	x
	F	x	x	x	x	x
30–39	M	x	x	x	x	x
	F	x	x	x	x	x
40–49	M	x	x	x	x	x
	F	x	x	x	x	x
50–54	M	x	x	x	x	x
	F	x	x	x	x	x

"stratifiers"—a term borrowed from the world of geological strata. A grid may be set up as in table 12.2.

We have *four* major age groups, *two* sex groups, and *five* regions, giving 4 × 2 × 5 = 40 cells. We may then decide to have ten teachers in each cell, giving a sample size of 400. Alternatively, if we know beforehand the number of teachers in the target population in each cell and these numbers differ from cell to cell, then we may decide to take a fraction of the number in each cell. If the cells have fairly similar numbers, we may decide to use a constant sampling fraction, but if the numbers differ considerably, we may use different sampling fractions for each cell based on the total number of teachers our resources will allow us to have in the sample. Typically, we discover teachers fulfilling the characteristics of each cell and take them as they come until each cell is completed.

Probability Sampling

The advantage of probability sampling is that from the internal evidence of our sample we can estimate the error of sampling. The main difference between probability sampling and judgment or quota sampling is that in probability sampling random selection is used, such that every pupil, or teacher, or class, or parent (or whatever our unit of sampling) in the target population has a specified nonzero chance of entering the sample. This implies that a listing of teachers or classes or parents exists. Let us take an example.

In chapter 7 emphasis was placed on the use of probability samples for the quality control of the implemented curriculum. Let us assume that a seventh-grade physics course has been implemented across all schools in the country. We need to draw a sample of students in all seventh grades. First we must draw up a sampling frame. In order to do this we again need to think of the criterion we are trying to estimate (in this case achievement in physics) and to bring wisdom and/or evidence to bear on what factors are likely to be associated with large differences in performance, in order to create the strata for the sampling frame. Typical stratifiers are type of school, sex of children in the school, and urban/ruralness of the school. Where there is a multipartite school system, the selection of students into different school types is usually on the basis of general performance and hence there is usually a considerable difference (often as much as one and a half standard deviations) between the high- and low-scoring school types. There are three sexes of schools—boys, girls and mixed; or there is one sex only, mixed but taught separately, or mixed and taught together—the difference in performance is often marked between the sexes of children in schools. There are also often large differences between urban and rural communities. However, it is up to each country to determine either from the wisdom of inspectors or the country's examination unit, or from previous surveys the important characteristics distinguishing between school performance.

Within each stratum it is necessary to know the number of schools and pupils in that stratum in the target population. Thus, a sampling frame similar to that presented in table 12.3 can be formed. This works well where there are good national statistics; but even in the best of cases the statistics are often two years out of date, and in the worst cases they either do not exist or are badly out of date or unreliable. In addition, some of the schools listed in the national statistics may have ceased to function or may have been merged with other schools or have changed their function (in terms of type of school). In this case, it is advisable to take the latest national statistics, but when making a random selection of schools within each stratum to draw three parallel random samples of schools such that three lists exist (i.e., two substitute schools for each school drawn). Hence, if a school drawn in list 1 no longer exists, it can be replaced by a school in list 2.

Table 12.3 gives an example of a sampling frame. In this example, there is only one *type of school,* but the stratifiers are size of school (large, medium, small), sex of school (boys, girls, mixed), and rural/urban. If there were two types of school (say academic and nonacademic), then strata 1–18 would be academic and these 18 strata would have to be repeated below for the nonacademic school type, forming strata 19–36.

Table 12.3. **Example of Sampling Frame**

	Population		*Sample*	
Stratum	*Schools*	*Pupils*	*Schools*	*Pupils*
1 Large, boys, urban	X_1	Y_1	x_1	y_1
2 Medium, boys, urban	X_2	Y_2	x_2	y_2
3 Small, boys, urban	X_3	Y_3	x_3	y_3
4 Large, boys, rural	X_4	Y_4	x_4	y_4
5 Medium, boys, rural	X_5	Y_5	x_5	y_5
6 Small, boys, rural	X_6	Y_6	x_6	y_6
7 Large, girls, urban	X_7	Y_7	x_7	y_7
8 Medium, girls, urban	X_8	Y_8	x_8	y_8
9 Small, girls, urban	X_9	Y_9	x_9	y_9
10 Large, girls, rural	X_{10}	Y_{10}	x_{10}	y_{10}
11 Medium, girls, rural	X_{11}	Y_{11}	x_{11}	y_{11}
12 Small, girls, rural	X_{12}	Y_{12}	x_{12}	y_{12}
13 Large, mixed, urban	X_{13}	Y_{13}	x_{13}	y_{13}
14 Medium, mixed, urban	X_{14}	Y_{14}	x_{14}	y_{14}
15 Small, mixed, urban	X_{15}	Y_{15}	x_{15}	y_{15}
16 Large, mixed, rural	X_{16}	Y_{16}	x_{16}	y_{16}
17 Medium, mixed, rural	X_{17}	Y_{17}	x_{17}	y_{17}
18 Small, mixed, rural	X_{18}	Y_{18}	x_{18}	y_{18}

For each stratum we need the number of schools and pupils within those schools in the target population in each stratum. The number of schools for stratum 1 is designated X_1 and the number of pupils Y_1. A rough estimate has then to be made of the number of schools and pupils to be included in the total sample. This will then determine the sampling fraction. Thus, if there are 2,000 schools in the total population and one decides to test in 200 schools, the sampling fraction for schools is 1 in 10.

It may happen that a particular stratum has very few schools and pupils. What should be done in such a case? If the stratum is considered educationally of little significance, it could be dropped or merged with another similar stratum. If, on the other hand, it is considered an important but numerically small type of school, it would be worth "oversampling" that stratum. That is to say, in order to get reliable estimates (i.e., with small sampling error) for that stratum, it might be desirable to take half or all the schools in the stratum. However, when it comes to calculating national estimates (i.e., for all strata together), the stratum must be reduced back to its size proportional to all other strata. In this case two sets of weights are required—one for looking at the stratum by itself and a second for weighting it in its proper proportion in the total sample.

The selection of schools may be done on the basis of random numbers. All schools should be ordered accordingly by strata. It is convenient to assign running serial numbers to all schools where the numbering of schools in each stratum will start by following the last number of the previous stratum. The sampling will, then, be done by selecting schools at equal intervals from the numbered lists of the schools starting with a random number within the sampling fraction. Such sampling procedures take care of unequal multiples of the sampling proportion in each cell. Thus, for example, if the sampling proportion is 1/20, it would be difficult to decide how many schools should be sampled from a stratum which contains 35 schools or from a stratum which contains 19 schools. The principle of the previous ordering of all schools from all strata and the selection at equal distances provides a solution for such a situation.

After the schools have been drawn in each stratum it is possible to use various methods of making a random selection of students within those schools. Three possibilities are:

1. Working through the list of pupils in the target population with a constant sampling interval with a random start less than the interval, i.e., if one third of the students are required and there are 30 of them in a class, a number between 1 and 3 is selected at random. If it is 3, then the third, sixth, ninth, twelfth child, etc., are selected.
2. Selecting the pupils whose surnames begin with certain letters of the alphabet.
3. Selecting the pupils whose birthdays fall on certain days, spread uniformly around the year.

In the first case, this random draw must be strictly respected, and no replacements chosen by the head teacher, who could be biased. In the second case, care must be taken that there is no association between the initial letter of surnames and ethnic or other groupings in the society.

A fourth possibility exists which is that of selecting a whole class within the school to represent the totality of students within the target population. As mentioned, this has the advantage of being administratively easy. It will work well if each class is equally heterogeneous. However, if the classes are grouped by ability, it will be difficult for one class to represent the school. Furthermore, it may prove to be more expensive to test a whole class of, say, 40 students when only 20 students need to be subsampled for that school.

A further variation in selecting pupils within schools occurs if one draws a sample where schools are drawn with a probability proportional to the size of the school. If we take our small, medium, and large schools in table 12.3, and if the medium schools were *twice* the size of the small schools and the large schools *three* times as large as the small schools, we could allocate the weights 3 to large, 2 to medium, and 1 to small, allowing large schools three times as much chance to enter as a small school, and so on. However, when it comes to drawing pupils within schools, we have to invert the number so that all pupils in the population have the same chance of entry to the sample. Thus, we might take 1/1, i.e., all the pupils in the small schools, 1/2 the pupils in the medium schools, and 1/3 pupils in the large schools. The random-selection techniques of pupils within schools could be the same as mentioned above.

The sampling frame given in table 12.3 depicts a two-stage (complex) sample, i.e., first schools and then pupils within schools. This differs from a simple random sample (s.r.s.). In a simple random sample all the students in the target population would be listed and a random sample of pupils drawn from the list. It is rare for this to be feasible, and most samples are two-stage. In large countries three-stage sampling is often necessary. The first stage may be the state, province, or administrative unit; the second stage is the schools within the selected states; and the third, the pupils within the selected schools. However, a multistage (complex) sample is bound to be larger, in terms of students, than a simple random sample where both have the same standard errors of sampling.

THE "LOGIC" OF EXPERIMENTAL DESIGN

The Need for Comparisons

Curriculum evaluation may require comparisons in three different settings. First, two or more alternative programs or curriculum packages may be compared. Most often, one of these is the program already in use in the system, and the curriculum developers wish to determine if the new program can attain certain objectives more effectively than the old program. Of course such comparisons are of interest only if all alternative programs have the same objectives. Quite often, different curricula have different objectives, and even if they have some common objectives, these often get different levels of emphasis within the program, and therefore comparison of the various curricula does not yield useful results. Comparison in such cases may be justified if one has reason to suspect that a program does not yield better results, even on those objectives that are highly emphasized in them, than the alternative programs in which these same objectives are not included at all. These considerations limit the value of comparison of educational outcomes of two different programs.

Second, one may compare the outcomes of the same program *used in different ways*. Thus, one may compare the outcomes of a program used without running special teacher-training courses with the outcomes obtained after having such training courses. By combination of program components such as textbooks, enrichment television programs, films, complementary reading material, teacher-training programs, and so on, one may construct different sets of actual programs based on a single curriculum kit. One may be interested in comparing outcomes of these different combinations of program components.

Third, alternative approaches or components within a given program may be examined; during program development a decision must be made as to which of several alternatives, performing the same function, is most suitable. Thus, for example, during the process of program development one compares the students' reaction to two alternative sets of illustrations, or to two alternative modes of presentation of the same idea, such as by diagrams or descriptive passages.

It is likely that comparisons of the first and second type will take place at the implementation stage of a program, while comparisons of the third type will take place at the early stages of program development.

In any comparison the important point is that the criterion variable to be compared must be the same for all treatments (programs or elements). Even if educational or learning objectives are not stated explicitly in the original program description, the evaluator has to specify all objectives to be used in the comparison. This is the first step in making a comparison.

Experimental Design

The objective of an experiment is to determine if a new procedure will produce the desired outcomes for the student population for which it was designed. Since it is neither desirable nor feasible to apply the procedure to the entire population, it is necessary to experiment initially with a relatively small group of students from the population. The problem of ensuring that this small group is as representative as possible of the entire population has already been discussed.

A second problem, besides representativeness, does arise. How does one determine if any observed effects are indeed produced by the new procedure and not by some other external factor? Consider a new program that has been assigned to a group of students. At the end of the course this group obtained a certain average on a test measuring the outcomes of the program. In this case it seems likely that the test results are the consequence of using the new program, since the test was administered at the end of the course; but there is no certainty that the measured outcomes are in fact caused by the new program. Some other factors, such as the students' access to television or parents' tutoring or some unknown external factor, may well have generated the measured outcomes. Furthermore, in examining the situation one cannot fully exclude the possibility that the test results measured at the end of the program were caused by the same factors that affected the assignment of the program to that particular group of students. This may be the case in sample surveys when one comes to examine the outcomes of different innovative educational programs that are introduced in several schools. It may happen that the new program has been introduced in schools having better facilities, more adequate equipment, better trained teachers, or more able students than the comparison or control schools have. The researcher is inclined to conceive the test results as outcomes of utilizing the innovative programs whereas in reality the programs were assigned to schools where initial achievements were already very high. Tests designed to measure outcomes of the program do reflect differences that are not outcomes of the program.

Thus, in order to make valid comparisons, one needs to set up an experiment. In the simplest case, two groups are formed, the control group (existing curriculum) and the experimental group (new curriculum). Again, attention must be paid to external factors in order to ensure that they are not the real cause of any observed outcome differences. Consider, for example, the case where two classrooms are chosen and the new curriculum is used in one of them. Some external factor, such as having well-trained teachers, may have determined that certain students found themselves in each of the two classrooms, and this same factor may be important in producing the observed differences in outcome. Thus we cannot decide whether the program is the cause of the test results or if the external factor causes both the assignment of a particular class to the innovative program and the observed differences in the results. Even with the utmost care in choosing groups that seem the same, some unforeseen factor is always possible.

How then can such external factors be excluded? The best way is to choose a random sample, large enough to cover both the control and experimental groups. Then the members of this sample must be assigned randomly, using a table of random numbers (cf. Fisher and Yates 1966), to each of the two groups. When this has been done, we know that no external factor is the cause of both the students' following a given program and/or the test results, since we know the exact cause of the first: the students find themselves in one of the two groups because of their corresponding random number in the table, and nothing else. Thus overall differences between the control and experimental groups can be caused only by difference in curriculum. Of course, internal differences within each group will be caused by a variety of external factors, but the objective of the experiment has been attained and the comparison of the two groups' means will yield valid results.

Analysis of Variance

The relative efficiency of two different programs can be compared by the technique of one-way analysis of variance. In a more complicated situation evaluators may wish to compare simultaneously two curricula and two teaching methods in order to determine not only which curriculum is better for the stated goals and which teaching method is better, but also if one teaching method is better with one curriculum than with the other. The solution of such a problem is done by a two-way analysis of variance. Although much more complicated designs of experiments are not usually encountered in curriculum evaluation, many such classical designs are available from other fields of research (Cochran and Cox 1957).

One-way design. This statistical technique is used to test to what extent group differences can be attributed to chance variation. An illustration of the use of analysis of variance is given in table 12.4.

Table 12.4 contains the means and the standard deviations of three groups on an arithmetic test. Group 3 utilized a new program in arithmetic, and groups 1 and 2 utilized a particular conventional program. It can be seen that the average achievement level of group 3 is higher than the average achievement levels of the other groups. Since the three groups were randomly assigned to the three different treatment groups, one may assume that the higher scores of group 3 are the result of using the new program. In order to test whether or not differences between the groups reflect chance variation only, an analysis of variance has been performed.

Table 12.4. **Performance on an Arithmetic Test**

Group	*Mean*	*S.D.*	*N*
3	22.10	5.63	1015
2	18.68	6.24	263
1	16.91	5.79	703

Table 12.5. **Analysis of Variance**

Source	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F*	
Between	11147	2	5573.5	167.1	significant
Within	65953	1978	33.3	—	at .001 level
Total	77100	1980	—	—	

Results of this analysis are presented in table 12.5. It can be seen that the observed differences between the groups are highly significant and are not the result of chance variation.

Two-way analysis of variance. In many experiments the evaluator is interested in assessing the effects of two different factors or treatments. Thus, for example, in the previously mentioned three groups, some teachers received in-service training and another group of teachers did not. Considering both the types of the program used in school and the variation with regard to in-service training, one may identify six different groups, as presented in table 12.6.

A two-way analysis of variance tests the significance of the two-factors: type of program and in-service training. The results of this analysis are presented in table 12.7. The data reveal that both factors contribute to the improvement of the students' achievement level. It is of interest to note that beyond the significance of the two factors mentioned before, a third significant factor emerges, which is labeled "interaction." This factor means that the effect of the in-service training was not the same for all three program groups. Indeed, an inspection of the results in table 12.6 reveals that the in-service training did not have any effect in the program of group 1, but it beneficially affected the results of groups 2 and 3.

Table 12.6. **Performance of Groups According to Program and Type of Training**

Program	*Training*	*Mean*	*S.D.*	*N*
1	1	16.9	6.1	379
1	2	16.8	5.4	324
2	1	15.3	5.0	60
2	2	19.6	6.2	203
3	1	21.6	6.2	134
3	2	22.0	5.5	881

Table 12.7. **Two-Way Analysis of Variance**

Source	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F*
Factor 1	11148	2	5573	169.1 signif. at .001 level
Factor 2	141	1	141	4.2 signif. at .05 level
Interaction	732	2	366	11.1 signif. at .001 level
Within	64979	1975	33	—
Total	77100	1980	—	—

Nonexperimental Designs

In educational practice it might be impossible to assign a new program or a particular treatment to a randomly selected group of students. It may happen that an educational system decides to introduce a promising new program in the whole system, and therefore no control groups are available. Such a situation frequently occurs in the evaluation of literacy campaigns. Obviously, it is very difficult to justify keeping a control group illiterate. Another case might be a situation where a new program has been introduced in a small number of classes on a tryout basis and where no other groups exist with similar learning objectives. In this case a comparison would be meaningless.

In both cases mentioned above the assessment of program outcomes has to be made without comparisons with control groups. However, it may also happen that a new program has been introduced only in some schools, not on a basis of random selection but rather on the basis of the schools' willingness to participate in the new program. In this case it may be possible to identify a control group similar in many respects to the experimental group, but since the assignment to experimental and control groups has not been done on a random basis one cannot be sure that the differences appearing between the two groups are caused by the utilization of the experimental program.

To summarize, it may be necessary to examine the effect of a program without using control groups or to utilize control groups that were not selected on the basis of random assignment. Evaluation studies of these types represent weaker designs than real experiments, and the user of such weak designs should be aware of their pitfalls.

The following section discusses problems related to program evaluation without using control groups or using nonrandomly assigned control groups.

Preexperimental Designs: The One-group Case

The one-shot case study. As indicated, in the context of curriculum evaluation it may be necessary to evaluate the effect of a program without being able to utilize a control group as a basis for comparison. Thus, for example, in a group of classes, a new program of biology has been introduced in which new types of educational objectives and learning activities appear. It may be inappropriate to compare at the end of the course the cognitive achievements of a student who participated in the program with the achievements of students who did not. It is very likely that those who learned a certain type of new materials will know these materials better than those who did not follow that particular course. Comparison of the two groups in this case is irrelevant for assessing the value of the program. Evaluators should here limit their interest to the performance of students who participated in the new program, and should satisfy themselves by examining whether these students attained the learning objectives of the program. Study designs that do not utilize control groups as the basis for comparison are termed *preexperimental designs.* Frequently, only results at the end of a study period are examined, using a single achievement test. Such designs are called *one-shot case-study designs* (Campbell and Stanley 1963). If the assessment of program results is performed on the basis of an end of course examination, it may be necessary to conduct a thorough observation of classroom activities in order to determine which classes implemented the program properly and to restrict the examination of end of course results to those classes where the program has been properly implemented.

A one-shot case-study design may be justified if the evaluator knows that at the beginning of the course students did not have mastery of program objectives, and if he is quite confident that the mastery of the skills demonstrated at the end of the course is the result of the program tested. This may be the case, for example, if the new program deals with teaching a foreign language that the students have not had the opportunity to study before.

The one-group pre- and posttest design. In some cases an educational program deals with skills in a domain that is not entirely new for the students. Thus, for example, a new reading program may strive to improve the reading comprehension skills of students.

In such a case the end-of-course examinations cannot by themselves reveal how much the student learned through participating in the program. It may be necessary to administer a pretest, i.e., to test the students' mastery of reading skills at the beginning of the course and then at the end of the course to administer the same test or a parallel test again. The difference in the achievement levels of the pre- and posttest should provide evidence that the students did indeed improve their skills in reading comprehension. The significance of pre- and posttest differences should be tested through a "test of correlated observations." It should be noted, however, that a significant difference between pretest and posttest results does not necessarily prove the success of the program. It is possible that a small increment in a certain skill attained by the overwhelming majority of the students yields statistically significant differences, but the intellectual gain is so small that the program's contribution is negligible.

Quasi-experimental Design

Quite frequently, educational innovations or experimental curricula are introduced in a small number of classes volunteering to use the new program. The evaluator who wishes to assess the efficiency of the program may identify a control group that has not received the innovative treatment and may compare the achievements of the two groups. However, since the students were not assigned to the experimental or the control groups on a random basis, such a design of comparison is not considered a real experiment. Campbell and Stanley (1963) term such comparison "quasi-experimental design."

In quasi-experimental designs there is less certainty than in a real experiment that differences appearing at the end of the course between the two groups are actually caused by the experimental treatment. There are at least two contaminating effects that threaten the validity of conclusions based on the comparison. First, it may well be that there were initial differences between the two groups with regard to mastery of relevant skills or the aptitude to learn certain skills. Second, it may well be that the group that participated in the experiment was better equipped, had better teachers, and so forth, which contributed to their progress more than the experimental program itself.

It is important to consider the possible impact of such contamination effects. There are no formal ways to control the second contamination effect. All that the evaluator can do is examine the situation thoroughly and try to draw a control group that operates in a setting similar to that of the experimental group. As to the control of the first contaminating effect, the evaluator may administer some kind of relevant aptitude or achievement test before the start of the program and compare the two groups in terms of initial performances. If the two groups were found equal on the initial measures, one may attribute the end-of-course differences to the differential treatments the groups received. If, on the other hand, initial differences appeared between the two groups, no meaningful comparison

Table 12.8. **Aptitude and Achievement Statistics of the Three Groups**

Group	*Aptitude*		*Achievement*		*N*
	Mean	*S.D.*	*Mean*	*S.D.*	
3	20.17	4.99	22.10	5.63	1015
2	20.53	5.04	18.68	6.24	263
1	19.21	4.75	16.91	5.79	703

of end-of-course scores can be made unless one eliminates the effect of initial aptitude differences. Fortunately this can be done through the method of analysis of covariance.

In order to perform an analysis of covariance to test the effect of a particular treatment, one needs to obtain data about the initial performance of each person in the two groups and a measure of their performance after the treatment has been given to the experimental group. Schematically such a design can be represented in the following way:

experimental group—observation 1	treatment	observation 2
control group—observation 1		observation 2

The numerical example given in tables 12.3 and 12.4 to illustrate the technique of analysis of variance is extended here for the sake of demonstrating analysis of covariance (table 12.8). The same three groups are compared. Group 1 received the experimental treatment, and groups 2 and 3 are two different control groups. To the data presented above, measures of aptitudes were added. It can be seen that on the aptitude measure no considerable differences between the three groups were observed. Nevertheless, group 2 had slightly better average scores on the aptitude measure than the other two groups.

In order to control these initial differences, an analysis of covariance was performed, the results of which are reported in table 12.9. It can be seen that even after eliminating the effect of the slight initial differences, the experimental group did significantly better on the end-of-course test

Table 12.9. **Analysis of Covariance**

Source	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F*
Between	9253	2	4626.7	176.7 signif. at .001 level
Regression	14193	1	14193	542.1 signif. at .001 level
Within	51760	1977	26	—

than did the control groups. These results provide evidence about the effectiveness of the experimental treatment.

It should be added, however, that in quasi-experimental design one can never be sure that one has succeeded in controlling all relevant initial differences between the groups being compared.

Even if one controlled initial differences in a certain aptitude, it might well be that groups differed also on other unidentified variables and therefore some initial differences remained uncontrolled. Thus it may well be that groups differed from the point of view of motivation and that the end-of-course differences were caused mainly by differences in motivation and not by differences in treatment. This is the major shortcoming of quasi-experimental designs in comparison to real experiments, where individuals are assigned to a particular treatment on a random basis. For this reason one should use quasi-experimental design only if there is no possibility of assigning individuals randomly to different treatments.

The Classroom as a Unit of Analysis

Almost invariably, new programs are introduced in entire classrooms and not to particular individuals. The performance of students studying in the same class are not independent of each other. They are exposed to the same teacher and to the same positive and negative experiences. The presence of highly aggressive children or the lengthy absence of a particular teacher equally affect all students studying in the same class.

It has therefore been emphasized frequently that in curriculum experiments one should consider the class as a unit of observation (see Wiley 1970). What does this mean in practice? Let us take an example of a new program that has been introduced in several classes. The experimental design contained treatment classes and control classes. If we take the class as the unit of observation, then the comparison of the groups will be based on the comparison of the class means only. In such a case all intra-class differences will be disregarded.

An example of such analysis is presented in table 12.10. For the sake of comparison, the analysis of the same set of data performed on the basis of class means is also presented (table 12.11).

Such an approach is justified only if the variances within the classrooms are not significantly different one from the other. If there are significant differences between the variances of the classes, the proper solution is to use a nested or a two-folded hierarchical classification analysis of variance (Kempthorne 1952). In this analysis the variation among students within classrooms within treatments provides the measure of error for the subsequent *F* tests of significance of effects of treatment.

Table 12.10. **Means of Class Means**

Group	Mean	S.D.	Number of Classes
3	21.33	4.07	45
2	20.27	2.82	13
1	17.53	3.31	38

Table 12.11. **Analysis of Variance of Class Means**

Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Between	303.0	2	151.5	11.45 signif. at .01 level
Within	1231.2	93	13.2	—
Total	1534.2	95	—	—

With the use of classroom means only, this is not possible since variation among such means will not be a good estimate of the error variance. The set of data analyzed above serves as input for an example of such analysis as presented in table 12.12.

It can be seen that by using the class average as a unit of analysis, one may observe significant differences, and that the *F* ratio has a value of 11.45. Applying the nested approach one should use as error term the mean square between classrooms. The *F* ratio obtained in this analysis is slightly larger and has the value of 29.92. It should be noted, however, that this value is considerably lower than the *F* ratio obtained in the case of

Table 12.12. **Analysis of Variance, Nested Design**

Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Between groups	11147	2	5573.5	29.92 signif. at .001 level
Between classes	17885	96	186.3	—
Within classes	48068	1882	25.5	—
Total	77100	1980	—	—

individual comparison. Thus results obtained on the basis of the analysis of individual scores and on the basis of the analysis of classroom averages may differ considerably.

Summary

Three types of designs were described above: preexperimental, quasi-experimental, and experimental designs. A schematic representation of these designs is given in table 12.13.

Preexperimental designs are used when no control groups are available. In such cases a posttest only, or pre- and posttests, will be used to determine the outcomes of a program.

Quasi-experimental designs are used in cases where control groups are available but where students cannot be assigned to groups on the basis of random selection. In such cases the utilization of analysis of covariance

Table 12.13. **Study Designs and Conditions for Their Use**

Type of Design		*When to Use*	*Type of Statistics*
Preexperimental	One-shot	No control group available Assumed students have no preliminary mastery in domain of program objectives	Posttest scores—mean percent of correct responses
Preexperimental	Pre- and post-	Students have some preliminary mastery of objectives	Significance of difference between correlated observations
Quasi-experimental	Experimental and control groups No random assignment	No random assignment possible	Analysis of covariance
Experimental	Randomly assigned experimental and control groups	Random assignment possible	Analysis of variance and covariance

is recommended as the most appropriate technique for comparing group results.

The most powerful technique for group comparison is the utilization of experimental design. It implies random assignment of students to various groups. Although experimental design is the most valid among the different designs mentioned here, in actual educational settings it is often the case that no random assignment of students is possible and the evaluator has to employ weaker models, such as quasi-experimental and preexperimental ones.

SUMMARIZING EVALUATION DATA

One major task of the curriculum evaluator is to analyze the data at the various stages of program development and summarize them for the curriculum team. Once the data have been collected, the analysis and data summary should be very rapid so that the developers may have the results within days after the data collection. This section deals with the analysis of data from typical evaluation instruments, namely, tests, scales, observation schedules, and questionnaires. The amount of analysis will depend on the types of technical aid available for data processing. Depending on the resources of a curriculum center, there may be either no equipment, or only a sorter, or a computer. Some centers have good desk calculators that allow quite an amount of computation. Also, the type of computer owned by the center or available to it will vary in size and quality. Calculation by hand is very time-consuming, and computational errors frequently remain undetected. The use of sorters, desk calculators, and computers not only speeds up work considerably, but it also ensures the accuracy of computational results. What follows are tasks that are typically performed at centers on data from various types of instruments.

Tests

Hand-scoring of Small-scale Data Sets

At the small-scale trial stage it is typical to produce summaries of formative data. Where the pupils do not number more than about 200, the analysis can be done by hand. In this case the developers will require the percentage of correct responses by all students for each item, or also by specified subgroups of students, e.g., boys vs. girls, urban vs. rural, or children from privileged homes vs. children from underprivileged homes.

Typically the students record their responses in a test booklet or on a special answer sheet. If the students' answer sheets are manually scored, it is useful to prepare a summary sheet in which the correct responses are marked 1 and the wrong ones are marked 0.