DOCUMENT RESUME

ED 250 047 JC 840 584

TITLE	Learner Outcomes Handbook. Improving Community College Evaluation and Planning.
INSTITUTION	California Community Colleges, Sacramento. Office of the Chancellor.; Western Association of Schools and Colleges, Aptos, CA. Accrediting Commission for Community and Junior Colleges.
SPONS AGENCY	Fund for the Improvement of Postsecondary Education (ED), Washington, DC.
PUB DATE	84
NOTE	71p.; For related documents, see JC 840 576-584.
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*College Planning; Community Colleges; *Evaluation Methods; *Outcomes of Education; Research Design; Research Utilization; *School Surveys; *Student Records; *Test Construction; Testing Problems; Two Year Colleges

ABSTRACT

One of a series of papers resulting from a Fund for the Improvement of Postsecondary Education (FIPSE) project, this handbook describes procedures for documenting the learning students take with them from college. Introductory material on the FIPSE project and the learner outcomes approach to college evaluation is followed by a discussion of alternate assessment procedures, including testing, institutional records, and surveys. The next sections look at each of these evaluation methods in detail. The section on the direct assessment of learning through testing considers the use of tests developed locally by individual faculty members or faculty groups and externally developed tests; the development of short-essay test questions; common flaws to be avoided in test construction; the characteristics of effective short-essay questions; the aggregation of the results of multiple-choice questions; multiple-choice tests; and the development of multiple-choice questions. The section on the use of existing college records advocates the determination of patterns of course completion as a way of expanding information, and suggests steps to increase detail in inferences about student learning based on college records. The use of surveys to gather information on learning outcomes is discussed next, with particular emphasis on follow-up surveys and the construction of useful surveys. Finally, the uses of learner outcomes data are explored. Information on sampling theory is appended. (LAL)

Improving
Community
College
Evaluation
and Planning

LEARNER OUTCOMES HANDBOOK

Chancellor's Office
California Community Colleges
Western Association Accrediting Commission
for Community and Junior Colleges

FALL 1984

Table of Contents

LEARNER OUTCOMES HANDBOOK

PREFACE

This handbook, describing procedures for documenting the learning students take with them from college, is one product of a project carried out jointly by the Chancellor's Office, California Community Colleges, and the Western Association Accrediting Commission for Community and Junior Colleges. The project is partly supported by the Fund for the Improvement of Postsecondary Education (FIPSE).

This material draws on Working Paper #10 of the FIPSE project, Measuring Community College Learner Outcomes: State-of-the-Art. It also builds from another handbook on assessing student outcomes, Student-Outcomes Questionnaires: An Implementation Handbook, which has recently become available in its second edition from the National Center for Higher Education Management Systems and the College Board. The present handbook does not duplicate the NCHEMS/College Board publication, focusing more explicitly on student learning. The emphasis of the NCHEMS/College Board handbook is on student attitudes, satisfaction, job placement, and job success. It "does not attempt to measure changes in actual student skill levels or achievement" (p.7). That is the kind of outcome this handbook does address.

The focus on "learner" outcomes rather than "college" outcomes or even "student" outcomes is important. Community college outcomes include providing services to the community and servicing the training needs of local business and industry as well as providing educational programs for individual students. Student outcomes are less broad than college outcomes but still include information, as on post-college employment, that may be more dependent on extraneous influences, such as the local economy, than on the college. Learner outcomes refer to student learning, the primary activity with which colleges are concerned. They are intended in this handbook to reflect the products students take with them from their classes and related experiences. The setting in which faculty members lay before students the information, activities, and resources from which they are to learn is, after all, the classroom. It is the setting in which the curriculum is put into effect, where faculty and student capabilities and curricular structure all come together.

All learning, of course, does not occur in a classroom, even when the term classroom is extended to include laboratories, studios, theaters, and field work settings. Students also learn from conversations with other students, and from unplanned experiences to which their studies may lead. Nevertheless, a college's purposes, as they are translated into the effects it expects to produce in students, are accomplished predominantly in the activities associated with courses and classes. "Learner outcomes" in this handbook therefore refers to the learning students acquire in their classes and in class-related activities. The issue addressed can be stated in the following question. How can a college demonstrate or document what its students are learning?

Learner outcomes have been studied in a variety of forms for a long time. Pace's 1979 book on the outcomes of college and their measurement is revealing in its subtitle, "Fifty Years of Findings..." Working Paper #10 reviewed and illustrated the varied forms in which outcomes have been conceptualized and studied, from achievement test results to economic effects. The NCHEMS/College Board handbook also reviewed a variety of classification schemes for higher education outcomes. None of the classification or definitional schemes reviewed can be considered superior in any general sense. The choice of any scheme must be based on the purposes for which information about outcomes is to be gathered and the audiences to which it will be addressed. Information on the occupational patterns of graduates, which is addressed to potential students, will take a form different from information on the comparative academic accomplishments of students in different programs intended for the college's department heads and administrators.

The concern in the present handbook for documenting the products or consequences of course-related activities limits appreciably and usefully the kinds of information to be gathered. In particular, it places this handbook and the NCHEMS/College Board handbook side by side with little overlap.

This handbook and a related item bank on learner outcome questions will be used on a test basis by community colleges in California and Hawaii during the coming year. This experience will enable us to revise and refine these materials for distribution. Since this handbook is in draft,

we would appreciate your comments and suggestions regarding its contents.

We are indebted to Jonathan Warren for his work on this handbook and to others on our Learner Outcome Committee (Ernest Berg, Allan MacDougall, Nancy Renkiewicz, Mike Rota, and Jean Vincenzi) for their efforts. They, like we, have other responsibilities and were it not for their gracious help, this project would not be possible.

Chuck McIntyre Project Director	Robert Swenson Project Co-Director	Dale Tillery Principal Project Consultant
Director Analytical Studies Unit State Chancellor's Office California Community Colleges	Executive Director Western Accrediting Commission for Community and Junior Colleges	Professor Emeritus School of Education University of California, Berkeley

Alternate Assessment Procedures

Evidence on what students have learned can be gathered in several ways. The process that comes most immediately to mind is testing each student individually. That is by far the most commonly used procedure but also the most expensive, particularly if student time is considered. If the purpose is to serve the college by providing information about how well its instructional programs are working, every student need not be tested, testing that has already been completed for other purposes can be used, and procedures other than testing are possible. Tests are direct observations of students' accomplishments. They can provide more detail than other procedures and can be focused on whatever aspects of learning are of particular interest.

A less direct but valuable procedure relies on institutional records of students' educational experiences--courses taken, grades received, perhaps records of tests previously taken. These are more remote indicators of learning than tests, but they nevertheless permit defensible inferences about student learning, particularly when used with groups rather than individual students. While all indicators of learning are derived originally from observations of the performance of individual students, or of the learning typical of the members of a defined student group, indicators of the collective learning of various groups of students are adequate for most college purposes.

A third type of indicator consists of questionnaires or surveys completed by students, faculty members, employers, or others through which the persons surveyed give their impressions of what students have learned, either individually or collectively. A fourth type, also involving questionnaires, consists of information on students' post-college activities from which inferences about students' learning can be drawn. Continued employment as a legal secretary, for example, is justifiable evidence that that person has acquired the capabilities necessary for that job. A further inferential step is needed to attribute that learning to a college program, but that step is often reasonable.

Direct Assessment of Learning

Tests of students' accomplishments are direct measures of what they have learned. Two qualifications are necessary, though, if they are to be interpreted as indicators of educational quality. The first, and by far the more important one, is the degree to which the test reflects the purposes of the course or educational program of interest--the learning

expected as an educational result. Tests, whether they are essay tests, multiple-choice tests, performance tests, or some other form, are likely to be most relevant if they are constructed by the person teaching the course. They are likely to be less relevant when constructed by an agency unconnected to the person providing the instruction.

A commercial achievement test in U.S. history, for example, may or may not be relevant to a college course in U.S. history. It will almost certainly have some relevance to a sequence of courses in U.S. history, but the emphases in the test and the program of courses should be reasonably close if the test is to be useful in measuring the quality of the program. A test may be wholly relevant to a course or program and still not reflect the learning produced if it was poorly constructed. Scores on such a test will not permit inferences about the learning of students who have completed the course or program.

The second qualification necessary in using tests as indicators of educational quality is the degree to which the students' performance resulted from the course or program of interest rather than from some other set of experiences that occurred either before or concurrently with the course or program. It is less critical than the issue of relevance because in most circumstances students are not yet proficient in the knowledge and capabilities of the courses and programs in which they enroll. While that is not always true, the proportions of students who have already mastered the material of a course or program at entry are usually small enough that the college can comfortably infer that their students' collective end-of-course performance on course-related tests is a consequence of the college's instruction. The concern that has recently been expressed for measuring educational quality in terms of value

added--that is, ensuring that the learning demonstrated at the end of a course was for the most part achieved in that course--is not a serious problem. The more important requirement is that the tests used have the educational relevance and the technical quality to indicate the desired kinds of learning.

Externally developed tests may be provided by textbook publishers, by test publishers, by professional associations, or by faculty members in other institutions. Regardless of the source, even when it is the publisher of the textbook used in the course, the test must be examined carefully to be sure the distinctions it will make among students are related to the purposes of the course. If an important purpose of a course in U.S. history is to have students understand the interplay, change, and persistence of social tensions since the Civil War, while the test provided by the textbook publisher focuses on political and demographic issues, the test may not be useful. Any useful test of educational achievement must discriminate among students in ways that are relevant to the learning addressed in the course.

Commercial tests produced by test publishers such as Educational Testing Service, the American College Testing Program, or the California Test Bureau, because they must be designed for broad use across many colleges and courses, are rarely wholly appropriate for any particular course. Exceptions may occur in tests of freshman English and mathematics, where the number and variety of commercial tests is great enough that the likelihood of finding one closely related to the purposes of a course is reasonably good.

The American Chemical Society produces tests appropriate for undergraduate chemistry courses. Other professional associations may

also publish tests appropriate for community college use, as in nursing or computer science. In every case, though, the relevance of the test to the course must be examined critically.

The most effective kind of direct test of student achievement may consist of a combination of commercial and locally developed tests. The commercial tests will often provide comparative information on the performance of other kinds of students, something locally developed tests lack. The locally developed tests will provide the direct relationship to the particular course or program that the commercial tests usually lack.

Another valuable procedure is to form a group of four to eight neighboring colleges from which groups of faculty members teaching similar courses might collaborate in the development of end-of-course tests. Even where courses are not entirely equivalent, the sharing of parts of an examination can provide valuable comparative information not otherwise available.

Locally developed tests: Individual faculty members

The most common procedure by far for assessing students' learning is the administration at the end of a course of a test developed by the person teaching the course. It can, if well planned, be effective in telling the instructor how well his or her students as a group have acquired the knowledge, understanding of concepts and their relationships, and intellectual abilities toward which the course was directed. It also works well in helping the instructor rank the students for the assignment of grades. Its strength is in the close connection possible between the test and the course objectives.

Faculty-developed tests tend to suffer from three deficiencies. The first and most important is that they give almost no information to anyone other than the person teaching the course. Usually no one else sees the test results. Other faculty members in the department, department heads, and deans only learn the proportions of students in a class who passed and were awarded various grades. No one except the instructor knows what the students learned beyond the minimal information conveyed by the title of the course.

A second common deficiency is in the quality of the test. Contrary to popular belief, the reliability of faculty-developed tests is usually quite acceptable. The persons who do well on mid-terms in general do well on finals. Those who do well in one course do well in others. Even when that is not the case, the differences in performance can often be attributed to real differences in the students' learning. Some students may have overcome earlier deficiencies or lack of understanding to increase their relative performance, or they may have shifted their attention to other classes or activities and slipped in performance.

The critical question about the quality of a class-room test, developed either locally or externally, is whether the results indicate what the students have learned. If a faculty member has four major objectives for the students in his or her course, each with three to five subordinate objectives, a good test will indicate the collective performance of the students in the class on each of those main and subordinate objectives. A common and often justified criticism of faculty-developed tests is that they indicate factual recall but not depth of understanding or other intellectual capabilities of a higher order than recall, objectives usually more important than factual knowledge. A test in public

health nursing, for example, may show the students able to recite the health problems typical of Southeast Asian immigrants while failing to distinguish between students who can and cannot relate those health problems to cultural conflicts.

The third deficiency in locally developed tests is the lack of any reference group to provide a context in which the results of faculty-developed tests can be understood. When faculty members devise and administer tests that accurately assess their students' accomplishment of the objectives the instructors consider important, they still don't know whether similar classes elsewhere have accomplished far more than, or not nearly as much as, their own. If the course's objectives are known to be similar to those of similar courses elsewhere, this may not be a serious deficiency. Differences in faculty preferences or emphases, though, often create differences between apparently similar courses.

In summary, while faculty-developed end-of-course tests are usually relevant and reliable, their information on student learning is not transmitted beyond the class. They often miss the more complex kinds of learning desired and provide at best a limited context in which to evaluate the results. All a faculty member usually has as a context to give meaning to the results is a sense of how students in similar classes have performed in the past.

Locally developed tests: Faculty groups

On rare occasions that might well be more frequent, end-of-course tests are developed collaboratively by several faculty members and used in each of their classes. This kind of collaboration can improve the likelihood that the test will reflect the important course objectives and

will broaden its interpretive context. Even when the collaborative effort consists of no more than each of four faculty members contributing one-fourth of the questions on a test of achievement in a course each of them teaches, the gain in information over four separate tests seems obvious. Differences among the four in teaching emphases and preferences may lead to consistent differences across the four classes. Those differences, though, as well as the absence of differences in some parts of the test, will give each faculty member valuable comparative information. Department heads and deans can learn from tests given in more than one class the kinds of learning that are common regardless of which course is taken or which faculty member teaches it. They can also learn what curricular objectives are generally being missed.

More extensive faculty collaboration, when the faculty members discuss what they want their students to accomplish and what kinds of questions would indicate that accomplishment, can further improve the quality of the information provided. Not only are the objectives likely to be clarified but informed criticism by the collaborating faculty members will make the questions developed more accurate and comprehensive indicators of the desired learning.

Collaboration on test questions can extend to neighboring colleges and to somewhat different courses. Different biology courses, one for nursing majors and one for biology majors, for example, may have a few objectives in common that would permit several common questions on both end-of-course exams. The differences in the two courses' overall content, purposes, student characteristics, and instructors will probably produce differences in test results. The nature of those differences, and of the questions where differences do not appear, in the context of the differ-

ent courses and student characteristics, will give both faculty members information about their own students' performance they would not otherwise have.

Externally developed tests

Tests of course-related learning developed by agencies other than the teaching institution have one valuable feature and one clear disadvantage. Because they are used at more than one college, results at any individual college or in any particular course can be interpreted in the context of known student performance elsewhere. Student achievement in a community college course in American institutions, for example, could be compared with student performance in similar courses at other community colleges, in lower-division courses in open-door four-year institutions, and in lower-division courses in selective four-year institutions. Without that kind of comparative information, students' test performance and educational effectiveness is difficult to interpret.

Even with comparative information, test performance alone is still inadequate for an evaluation of educational quality. Students' prior learning, reasons for taking the course, the purposes of the course, and the capabilities of the students in the course all affect test results independently of the quality of instruction. But all that additional information, which is necessary to provide an interpretive context for test results, does little good if comparative information is not also available. Externally developed tests usually provide some comparative information, though often not as much as would be desirable.

Externally developed tests are most often deficient in their relevance to the learning expected in a course. When the test comes from the

publisher of a text on which the course is heavily dependent, or when the course is designed for certification by a professional association that provides the test, or when the content of both course and test are highly standardized, as in a first-year calculus course, the test can be expected to reflect the learning in the course. When none of those circumstances exists, the tests and course objectives will often match poorly and the tests will not indicate the learning that occurred.

Widely used achievement tests in English literature, mathematics, history, science, social science, or any other broad academic field must be carefully examined to determine the degree of overlap between test content and course content. Usually the test will include some content not present or given limited emphasis in the course and will lack content in some areas important in the course. Those dissimilarities in content are often great enough that a test with an appropriate title will not give acceptable information on the learning achieved in the course.

Test relevance involves importance as well as content. A question, for example, asking about the reasons Black migration from Southern farms to Northern cities was encouraged after 1915, while relevant in content to a course in American history, may reflect an unimportant element in the course. Good student performance on questions that require little more than the recall of information from the classroom or text will not indicate the accomplishment of more complex learning objectives. If an important goal of the American history course is to help students understand the interplay of economic and social forces, questions to assess that kind of achievement must require that understanding in their response.

This is not an argument against the use of multiple-choice examinations. When well constructed, they can indicate complex and subtle kinds

of understanding. In whatever form they may take, examinations based on the recall of information are easier to devise than those that will reveal a deeper understanding of an issue. Whether a test is multiple-choice, fill-in-the blank, short essay, take-home essay, or observation of performance, it must discriminate among students in their accomplishment of the important objectives of the course.

A common procedure in test development or selection is to construct a table of test questions and course objectives. Each objective in the course must find representation in one or more of the test questions, with more important objectives represented by a greater number of questions. This procedure is effective and easy to apply when objectives can be simply and clearly defined, as in the specificaton of facts that should be known or skills that should be mastered. The more complex objectives do not fit easily into that process because they tend to be difficult to assess. Even though they are often more important than the simpler objectives, the difficulty in formulating workable test questions for them means they will often be slighted in measuring student learning. The table matching test questions with course objectives can reveal that testing deficiency in both the development and selection of tests.

To avoid slighting course objectives that are important but difficult to assess, the objectives should be clearly defined before the test questions are linked to them. Fewer questions will appear for the complex objectives simply because they are more difficult to develop. Nevertheless, unless the effort is made, important kinds of learning will be missed.

Developing short-essay test questions

Short-essay questions are defined as test questions that require the students to write a paragraph or two in about 10 minutes. Essay questions that require much longer than that will probably produce responses that are too complex or that differ in too many ways to discriminate understandably among students. They may also not be worth the added time. Essays written in 20 or 30 minutes, for example, are not likely to produce results that carry much more information than ten-minute essays. Too much of the writing will be padding, and the responses will be so variable that comparisons will be difficult.

Two solutions are possible if a longer essay is considered necessary to permit students to demonstrate the depth with which they can treat a complicated set of issues. First, a long essay can be broken down into three or four major components, each of which is treated as described for short-essay questions. Second, the process can be reversed. The students can be given two or three five- or ten-minute essay questions and then be asked to write an additional ten-minute essay that builds on or integrates their early responses or treats one in relation to another. The second procedure helps the students organize a 20- or 30-minute essay. The first approach leaves the organization wholly to the students unless they are given guidelines in the essay instructions. Both procedures can be useful in different circumstances. The problem they avoid is finding a sensible way to get enough information from a simple long essay to justify the time it takes to write it and the complexities in grading it.

The grading of essay questions, which is one of their major problems, can be accomplished efficiently and accurately by assigning responses to predetermined categories. The unmanageably large number of categories

needed for the extensive information to be expected in responses to long essay questions is another reason to prefer shorter questions. Complexity in the responses, which is the only justification for committing a large amount of time to a single question, complicates grading regardless of what method is used. Responses will differ in the number, importance, and accuracy of the issues discussed, in how well the connections among the issues are developed, and in the discussion of implications or connections to broader issues. Problems occur, for example, in comparing a response that focuses on two or three issues but develops them fully with one that states briefly but accurately ten or twelve relevant issues without depth or elaboration.

A common procedure in grading essay questions is to award points for each issue accurately discussed and perhaps for the quality of the integration among the points. The effect is to reduce the single long essay question to a series of short questions that were merged into one. Because parts of a long response depend on the students' having provided other parts, more accurate assessment of the students' knowledge and understanding is achieved with several more focused, shorter questions, perhaps linked as noted above. The proportional amount of information about student learning that can be accurately inferred from essay responses declines with the amount of time allowed for the responses.

Every adequate test-development procedure takes time. It's time that is necessary if test questions are to be developed that give information about what students have and have not learned. Skipping any of the steps listed below will increase the uncertainty in the discriminations among students the questions are intended to provide. Following

the steps carefully will not guarantee that each question will work the way it is intended to. It will, however, give each question enough credibility that when unexpected results appear, careful study to determine their source will be valuable.

For example, questions related to similar content or requiring similar abilities should show similar results. A question that requires students to understand the relationship between two concepts in one context should show results similar to those of another question that differs only in the context to which it applies. When the same group of students produce responses to the two questions that are not related, in the sense that students who score well on one will score well on the other, those differences will probably identify a source of misunderstanding or some other reason for the failure of the concept to be generalized to a different context. In contrast to multiple-choice questions, the short-essay questions that seem not to work right should not be abandoned but should be examined for the reasons for their apparent failure. The following steps help produce informative questions.

1. Specify the variety of ways someone who is competent or successful in the course or educational program of interest differs from someone who is not. (This is almost the same thing as specifying course or educational objectives, but it sometimes produces more direct, concrete statements.)
2. Identify what a student might be asked to describe, explain, or discuss that would discriminate among students in the ways specified in Step 1.
3. Devise questions that require the students to carry out the exercise specified in Step 2.

4. Examine the resulting questions in relation to the following points, which are a restatement of the process through which the questions were initially formulated.

   What kind of distinctions are likely to appear among the responses?

   What would those distinctions imply about the students who give different responses? Would those implications be useful or informative about pertinent student abilities?

   What deficiencies in competence would the less-than-admirable responses indicate?

   Do those deficiencies involve the capabilities of interest?

   If indicated, clarify what the questions ask the students to do, or limit or broaden their scope, or revise them to bring them closer to an indication of the ability of interest.

5. Ask two or three faculty members and two or three students who have already had the relevant course to respond to each of the questions and then to tell you what they think is required to produce good answers. (Any individual should be asked to review no more than three or four questions.) If their responses or their perceptions of the issue a question illuminates do not match what you expect of the question, ask their help in revising it.

6. Use the questions in a class.

7. Collect the responses and, without regard for their merit, sort the responses to each question separately into groups that are qualitatively different--that have some set of identifying characteristics and lack others.

8. Examine the responses in each category for variations in merit. Use the characteristic that produces any variation in merit to

break up that category into two or more. For most short-essay questions that can be answered in about ten minutes, five to seven categories of response are usually sufficient. If more than that number appear, determine whether all the distinctions among responses are important. If not, merge some categories.

9. Now put the response categories for each question separately into a rough order of merit, allowing some categories to be tied with respect to merit. All qualitatively different responses to the same question do not necessarily differ in quality or merit.

10. Grade the responses to each question by assigning them the grade associated with the category to which they belong.

Common flaws to be avoided in essay questions

1. Lack of clarity in what the students are asked to do. When competent students are confronted by a question, they should know almost immediately what they are being asked to do, even if they aren't able to do it. Consider the following example from a course in public health nursing.

> Discuss two or three problems in public health that might be improved through better knowledge of public health principles.

This question is vague in what the students are asked to focus on with respect to public health problems--their causes, prevalence, severity, implications for nursing practice, solutions, or other aspects. It also suffers from the following problem, lack of boundaries.

2. Too much scope. The question is so broad that a wide range of responses could all fit within its demands. A good question will have a boundary around it, limiting as well as defining the area the student should give attention to, the scope of the desired response. Too much

scope to the question makes it hard to justify criticizing responses that are off the mark or miss the point. They may well have been within the boundary of the question as it was stated, even though not as the instructor intended. For example, in the question above from public health nursing, how can any response be criticized, i.e., differentiated from superior responses? All the students are asked to do is discuss a broad topic. Any set of statements can constitute a discussion.

3. Too little scope. The question is so limited in scope that many minimal responses will have to be accepted as equivalent to better, more elaborate responses that go beyond what was asked. There will be no way to infer that a student producing a minimal response would not be able, if asked, to produce a better one. The better response had not been asked for.

From a course in marketing for small businesses:

State at least two ways a marketing plan can help a small business.

Two brief statements would give a complete response. The question doesn't ask for anything more, which limits the inferences about learning that can be drawn from the responses.

4. Too complex. The question is too complex to permit informative evaluation of its responses. A full response will have so many elements contributing to it individually and in relation to each other that most responses will be incomplete simply because of the complexity of the question rather than limitations in the students' competence. This is similar to the problem of excessive scope. It is also a common problem with long essay questions.

From a landscaping course on woody plant materials:

> A Japanese-style garden is more than just a collection of plants native to Japan. Discuss the cultural, historical, and esthetic associations of five plants native to Japan in terms of their use in occidental and oriental landscape.

The questions calls for a discussion of three kinds of associations of five plants in two contexts. That three-by-five-by-two organization requires 30 elements in a complete response.

That question's complexity will collect its price when the question is graded, which is true of most of the deficiencies described. Any number of exam questions can look marvelously probing when first devised and then produce nothing but frustration when useful distinctions can't be found among the responses. The 30 elements in the Japanese garden question are likely to seduce instructors into grading it by counting how many of the 30 elements appear in the responses. Then they will feel uneasy about counting a poorly described element as much as one that is well and fully described, and they may complicate their grading scheme even further. At best, though, counting how many of a number of possible elements appear in a response is not likely to produce the kinds of discriminations the instructors want to make. Five or six important qualitative distinctions in the responses to a question are likely to be optimal in terms of the usefulness of the distinctions in learning they permit among students and the accuracy and ease of grading.

5. Testing recall rather than understanding. The inclination to grade questions by counting elements in the responses, as illustrated by the Japanese garden question, will often force the discriminations among the responses to reflect the simple recall of information rather than the exercise of a capability or understanding. An understanding demonstrated

by the ability to articulate relationships among elements or concepts is usually valued more highly than the ability to identify or even describe those elements or concepts. If the relationships of interest were discussed explicitly in class, though, their statement in response to a question may still only indicate that something said in class was remembered. The most useful questions will therefore require the active exercise of the ability or understanding being assessed rather than the recall of an instance of it.

6. Asking about opinions or beliefs. Occasionally a question will ask for students' opinions or beliefs. While they may be of interest for some reason, they provide useful information for assessing learning only when the students are asked to explain, defend, or support their opinions. The basis for inferences about capabilities is then the strength of the argument rather than the content of the belief. Opinions or beliefs can be held for any number of reasons unrelated to what the student has learned. Any accurate statement of the student's opinion would be right regardless of how little support could be provided for it unless the question required a sound defense of the opinion.

An effective short-essay question.

The following question is one that might be used in a course in modern American literature.

> A reviewer of James Gould Cozzens's Pulitzer-prize-winning novel, Guard of Honor, commented that it gave one of the most accurate pictures of the U.S. in World War II found in any novel to come out of that war. Yet the entire action of the novel occurs in a three-day period at an air base in Florida. Select an incident in the novel and describe how Cozzens's treatment of it might have led to the reviewer's comment.

The scope of the above question is clearly defined and limited but broad enough to allow room for students to demonstrate their understanding of the novel at a deeper level than recounting its narrative structure.

Either of two issues the novel treats would be expected--(1) the effects of war on the country's social fabric or (2) the demands being at war makes on various types of individuals. The key to a good response is the coupling of one of these two issues with an apt and accurate illustration from the novel. The following types or categories of response would be likely to appear.

1. A clear statement of one of the issues with an apt and accurate illustrative incident.

2. A clear statement of an issue treated in the novel that is not one of the two expected, with an apt and accurate illustrative incident.

3. A clear statement of one of the two issues with an inappropriate or inaccurate illustration, or none, which raises doubts about the student's understanding of the issue.

4. A clear statement of a secondary issue, as in Category 2, with an inaccurate or inappropriate illustration or none.

5. No statement, or an indefensible statement, of a reason for the reviewer's comment, regardless of the illustration.

The five response categories above are ordered roughly by merit. The distinction between the first two categories may be debatable if a case can be made for other reasons for the reviewer's comments. This illustrates a distinction that would become clear with the reading of 20 to 30 responses. Responses in categories 3 and 4 differ from those in 1 and 2 in the student's inability to mesh a specfic incident of the novel with the major issues the novel addresses. That distinction, contrasting 3 and 4 with 1 and 2, may indicate students who remember hearing in class

about the issues the novel illuminates but who don't understand how the novel does it.

Aggregating the results of essay questions

While tests of all kinds are commonly used to rank students for the assignment of grades, the use of test results in evaluating the learning that has occurred requires more information than the students' relative ranks. Grading essay responses by assigning them to categories having specified characteristics permits the performance of any group of students to be specified by reporting the percentages of the responses that fall in each category.

To illustrate, consider the following results from the question on Cozzens's Guard of Honor. The figures are the percentages among four groups of students in each of two different classes who gave responses in each of the five categories. Class A is a class in American literature in which lower-division credit is transferable to four-year colleges and universities. Class B is a similar class that does not carry transfer credit. In that class, the students are grouped by age--those 21 and under, and those over 21. The results are shown for that class as a whole and also grouped by age.

Student Group	Response Categories (Percentages)					
	1	2	3	4	5	Total
Class A	30	20	25	22	3	100
Class B	17	14	26	29	14	100
21 or Under	5	5	30	35	25	100
Over 21	33	20	24	20	3	100

In the transfer class, half the students respond in the top two categories, showing the ability to identify a major issue or theme

of the novel and relate an incident in the novel to it. Almost all the rest are able to state a theme of the novel but don't understand it well enough to illustrate it with a relevant incident. In the nontransfer class, only a third can relate a theme to an apt illustration, while 14 percent cannot state one of the novel's issues. When the nontransfer students are grouped by age, however, the older students are indistinguishable from the transfer students and the younger students look much worse, with only 10 percent able to illustrate an issue with a well-chosen incident. The difference between the two classes seems attributable to the characteristics of the students enrolled rather than the instruction. More importantly, though, the nature of the learning or understanding that produces the difference in performance is clear.

Multiple-choice tests

Externally developed tests usually consist of multiple-choice, matching, fill-in, or other types of questions that can be graded completely unambiguously, that is, "objectively". When well constructed, they are capable of assessing complex kinds of learning. They often, however, are limited to the assessment of the recall of unconnected small pieces of information. Their appeal is in their ease and speed of grading. That benefit is bought, however, at the price of difficult and extensive development and trying out of the questions. Without careful development, a set of multiple-choice questions is likely to give little accurate information.

The 40 to 50 questions on a one-hour multiple-choice test, if they are to give useful information about student learning, should be neither too difficult nor too easy for most of the students who take

the test. That assertion is challenged by proponents of mastery learning, who expect all successful students to respond correctly to at least 80 or 90 percent of the questions on a test. In most college courses, though, mastery cannot be that clearly defined. Students will typically vary widely within an acceptable level of achievement, and learning the nature of that variability is important to any educational evaluation. Test questions that almost all or very few students answer correctly will not provide much useful information. Good multiple-choice questions will therefore be answered correctly by roughly 20 to 80 percent of the students.

A second requirement of any test question, multiple-choice or essay, is that it be relevant to the learning expected in the course. This calls for the judgment of the instructor as to content and the nature of the learning required for an acceptable answer. It also implies that the students who have been most successful in learning the course material will most frequently respond correctly to each question. If students are ranked according to their grades on a multiple-choice test, more students in the top third than in the bottom third of the class should have the correct response to each individual question.

Each of these requirements is difficult to meet without trial and revision. No multiple-choice question can be relied on until at least 40 or 50 students in a relevant class have responded to it. Only then will the question's difficulty level and relation to the learning of the course be known. Students are ingenious in interpreting multiple-choice questions differently from the way intended, destroying the validity of any inference about the learning represented by responses to the questions. Usually several trials of groups of test questions are needed before

enough effective questions for a one-hour test can be devised. If several classes are available and the same test can be repeated with new classes, good multiple-choice tests can be developed. Too often, though, that developmental process is neglected, leading to multiple-choice tests that produce a spread in students' grades and have an appearance of relevance but that are not accurate indicators of student learning.

Developing multiple-choice test questions

The following steps are not intended as a recipe for multiple-choice tests. They are helpful, though, in illustrating potential difficulties and ways to prevent them.

1. Write a list of one-sentence statements a successful student in the course should be able to make. These can take a variety of forms, including statements of fact that simply require knowledge, statements that can be made only in the context of some given prior information and that require an inference from the given information, statements that can be made only after a problem is solved based on given information, or other forms. This step, if the statements are grouped according to the broader concepts or aspects of learning they represent, catalogs the learning expected in a course.
2. Reconstruct each statement into a question, leaving some aspect of the statement, or an inference from it, or a solution to a problem, to be provided by the students. With that as one response option, write three other plausible but incorrect response options. This is the most difficult step in writing good multiple-choice questions. All three wrong responses in a

four-option response must look reasonable to some students. A procedure that can provide good wrong options is to give the questions to a small group of students without providing any response options, asking the students simply to write out an answer to each question. The wrong responses provided by the students can then serve as good options in a multiple-choice format.

3. Ask several faculty members or advanced students to respond to the questions and tell you why they chose the response they did. Flaws in either the right or wrong options or in the wording of the question itself may appear.
4. After any revisions that may be necessary, tabulate the questions by the aspect or objective of the course each draws on. Collect 20 to 40 questions that span the desired kinds of learning into a 30- to 50-minute test. About 40 will usually make a 50-minute test. Fewer than 20 will not provide enough reliability for the test to be useful.
5. Give the test to one or two classes, providing at least 40 to 50 students.
6. Score the test, rank the students according to their total scores, and group the students into the lowest through highest quarters (or fifths) of the score distribution.
7. For each question, tally the number or percentage of students in each quarter (or fifth) of the score distribution who gave each response, as shown in the figure on the following page.

Percentages of Responses by Quarter

Response	Quarter				
	1	2	3	4	Total
A	2	2	0	0	1
B	32	52	68	82	58
C	36	22	12	10	20
D	30	24	20	8	21
Total	100	100	100	100	100

In the above example, Option B, which 58 percent of the students chose, is the correct response. The regular increase in the percentages choosing that response, from the lowest to highest quarter of the distribution of total test scores, shows that that question is related to the others in the test and contributes to a reliable score. The question could be improved if Option A were revised to make it a more plausible choice. In its present form, the question is functioning essentially as a three-option question. Because it is moderately difficult, even with only two effective wrong response options, the question still works well. If 70 or 80 percent were responding correctly, the improvement of Option A would be more important. Even when only 20 to 30 percent of the students choose the correct response, a question will work well if those 20 to 30 percent show a rising trend from the lowest to highest quarters of the total score distribution.

8. Rescore the test with the ineffective questions removed.

Rewrite or replace the ineffective questions before the next use of the test.

Existing College Records

Records that most colleges keep routinely can be drawn on for evidence of student learning. The records most directly related to learner outcomes are student transcripts, or the records of students' courses completed, and grades assigned. Enrollment data by course also give some information on learner outcomes, although it is more accurate when limited to courses completed, or enrollment at the end of a course.

The assumption that learning can be defined in terms of the courses students have completed underlies some of the procedures here proposed. When desired, that assumption can be verified with additional work. If the assumption is confirmed, the comparatively inexpensive procedures for documenting student learning through tallies of courses completed can be followed comfortably. If not, course descriptions may be revised to bring the assumption and actuality into agreement by specifying the learning that can reasonably be expected of students completing each course.

At the simplest level of analysis of courses completed, a college might list each course taught in a given semester with the numbers of students who completed each course. That would provide a crude but inexpensive picture of what the college had accomplished. Most people would want more information than that, probably requiring one or both of two ways to expand the tally of course completions.

Patterns of course completions

One way to expand the information in the numbers of students completing each course would combine individual courses into groups, such as several courses in practical nursing, in accounting, in real estate,

or in accounting and real estate combined. The numbers of students completing each group of courses, representing more extensive accomplishments than those associated with individual courses, would then be tallied. The numbers of courses per group can vary from two or three to all the required courses in a given program, depending on the level of detail required.

When the number of courses in a group is large, the tally of students completing that group differs only slightly from the tally of students completing a particular program. Smaller groups or clusters of courses are usually more informative than groups that incorporate an entire program because they show the variety of ways students can put together different combinations of courses to reach a degree or certificate in the same field. They have the further advantage of documenting the accomplishments of students who leave college without completing an established program.

The difficulty with tallying the numbers of students who have completed various course combinations is simply the large number of different course combinations that can occur. From only 200 courses, for example, more than a million groups of three courses and more than 64 million groups of four courses can be formed. In actuality, most of those possible groups will not appear in any tally of course completions, but the numbers of groups of interest and of various sizes that do occur will still be too large to be practicable. One solution is to identify in advance a small number of groups of courses that represent different aspects and different levels of completion of each program. Tallying the numbers of students who complete each of those course groups will then provide a useful indicator of what the college is accomplishing in terms

of student learning. The limitation to this procedure is that it will miss the ways students combine courses across different fields or programs, many of them common enough to be important, as well as the unexpected combinations that may occur within a program.

A second solution to the problem caused by the huge number of course combinations possible is to identify through an analysis of samples of transcripts the course combinations that most frequently occur, whether in a single field or program or in several. Such a procedure is not feasible with very large numbers of students or transcripts, but it can be carried out with samples of about 200 students, with each sample drawn from different programs or from among students not identifiable with any program. Identifying the groups of courses that appear in a series of such analyses will define the learning of students in terms of the common patterns of courses students take, whether identified with formal programs or not.

A number of analytical procedures are available that can identify patterns in a body of data like the courses completed by each member of a group of students. Cluster analysis, multidimensional scaling, and smallest space analysis would all work in slightly different ways but with essentially similar results. The starting point for any such analysis is a table of students and the courses they completed, with a one in the table where a student had completed that course and a zero for a course a student had not completed.

The usefulness of the patterns of courses identified as having been completed will depend on the way the group of students or transcripts analyzed is selected. Samples of transcripts of students who had just completed a program--an AA degree in business or computer science, or an

LVN certificate, for example--would indicate the variety of patterns students follow toward a common result. They will probably be more numerous in business than in computer science, and perhaps no more than one will appear in the LVN program. Some patterns will be common to business and computer science, reflecting the general requirements for the AA degree. Others will be common to those two programs because for some students they will overlap.

In a large program such as business, several groups of transcripts may be selected for analysis. One may consist of the transcripts of students completing the AA degree in business while others may consist of transcripts of students who have completed various numbers of units in courses that lead toward an AA in business. The patterns of courses appearing in the transcripts of students who have completed 15, 30, 45, and 60 units toward a business AA will indicate the nature of the learning at various stages toward that degree. Other groups may be formed from transcripts of students completing a specified degree or program who entered the college for the first time with various numbers of units completed elsewhere. Comparing those transcripts with those of students who had taken all or most of their work at the college awarding the degree would indicate differences in content or substance of student learning that result from attending more than one college and how the final college attended builds on the learning acquired elsewhere.

Assume a college offers programs in business and management that together offer a total of 48 courses. Within those general areas, ten curricular options are offered such as marketing, accounting, and small business operation. The college may want to know how those various options differ in terms of the learning students take from them. Are the

marketing and sales management options similar enough that they should be combined? Do the ten options in general reflect the most useful curricular arrangements for students' purposes? Are students completing any of the programs without having accomplished some of the key aspects of learning expected of them? Are some aspects of the desired learning treated repetitiously while others are slighted? What patterns of learning are the noncompleters taking with them? What are the stages of learning typically followed in completing the various programs? All of these are questions that can be answered by observing the patterns of courses that appear on the transcripts of various groups of students.

As a start, the college might study the transcripts of a sample of 200 students who had earned an AA in business in the last two to four years and tally the business courses each had completed. The different course patterns observed among those 200 students would indicate differences in the content of the learning acquired and how those differences matched the ten options. The results might, in part, show the following:

1. A group of three courses or their equivalents--Elements of Management, Principles of Accounting, and Introduction to Data Processing--had been taken by almost all the students.
2. A second cluster, primarily taken by students intending to transfer to a four-year institution, included Advanced Accounting, Business Law, and International Trade.
3. A third cluster consisted of at least two computer programming courses, Computer Operations, and Computerized Accounting Systems.
4. A total of seven clusters of three or more courses appeared that, in various combinations, described the ten options.

6. The learning associated with each cluster of courses could be summarized in a paragraph. The paragraphs could then be combined, with descriptions of a few individual courses added, to describe each of about 14 patterns of learning students showed on completion of an AA degree in business. Two or three of the planned options were followed by only a few students, with major parts of them appearing in conjunction with other substantive emphases.
7. Three of the patterns were characteristic of students intending to transfer to four-year institutions, although some students intending to transfer had shown other patterns.
8. Courses in banking and finance, although recommended in three of the options, were missing more often than the faculty thought desirable.

To determine how the business and management courses were being used by students who had not completed an AA degree and who may or may not complete it in the future, the college studied the transcripts of a new sample of 1,000 students who had enrolled in one or more business and management courses in the preceding eight semesters--125 students per semester. The sample had been selected to include equal numbers of students whose total records, as of the current date, showed 3-15, 16-29, 30-44, and more than 44 units completed in all courses. The patterns of courses completed by the last two groups were similar to those of the degree completers. Those of the first two groups were less so. The first group, because of the necessarily small number of courses completed, showed a large number of two- and three-course clusters that, while interpretable, matched the clusters of the completers in only a limited way. The course patterns of the third group, students who had completed

the equivalent of from two to three semesters, showed intermediate steps between the students with less than two full semesters and the completers.

The third and fourth groups showed patterns of learning that indicated the building of an integrated understanding of a number of definable aspects of business and management that had not been accomplished earlier. The students with the equivalent of less than a year, however, while lacking the integrated knowledge of the students who were approaching completion of a degree, nevertheless showed limited patterns of learning that could stand as useful accomplishments independent of any courses that might be completed later. The small clusters that showed some coherence, even though completed in less than two semesters, included Tax Accounting, Business Taxes, and Small Business Management; The Personal Computer in Business, Computer Programming for Business, Advanced BASIC for Business; Real Estate Management, Business Law; Legal Terminology, Legal Concepts, Office Procedures for Legal Secretaries.

When the numbers of students who completed each identified pattern of courses were tallied, the results provided moderately detailed descriptions of what students had accomplished at a college in one field in varying periods of enrollment. They can be expanded with parallel analyses of the courses the students in the various samples have accomplished outside the field of business. Those analyses would show the nature of the educational breadth typically achieved by students in satisfying the college's general education requirements. The usual result would be to show little coherence in the learning associated with those requirements.

The patterns of courses completed, identified as described above, will differ for different types of students. Students intending to

transfer to a four-year institution will show patterns that differ from those not planning on further education beyond the community college. Patterns will also differ by age and sex. The most useful differentiation of patterns of learning by types of students is likely to be in terms of students' educational objectives and later success. Students who achieve their planned objectives within a year or two of leaving the community college, whether they are educational or occupational, may show patterns of enrollment different from those of students who state the same objectives but fail to achieve them. Those findings would be useful in evaluating various course and curricular offerings as well as in advising students. They would not necessarily show one course pattern to be superior to another except for reaching particular objectives, and a course pattern that leads to success in one objective may be less valuable for another.

The analyses of patterns of courses completed go beyond the listing of enrollment figures by program in two ways. First, they show patterns of learning that are not necessarily described by program requirements. Second, they show patterns of learning achieved by students who do not complete any regular program. In describing what students with various numbers and types of courses completed can be expected to have learned, patterns of course completions can be informative and useful, both in documenting institutional accomplishment and in evaluating the college's curricular structure. They nevertheless lack direct evidence of student learning, relying instead on expectations of learning that will be met imperfectly at best.

Providing direct evidence of the learning typically associated with each course is the second way the information in courses completed can be expanded, in addition to identifying patterns of course completions.

Each of the following steps will increase the level of confidence and detail in inferences about student learning.

1. Elaborate on the course titles through the use of more detailed course descriptions, such as catalog descriptions or course syllabi.
2. Ask the instructors currently teaching the courses to state their primary objectives for each course, and use them to add detail to the accomplishments associated with transcripts.
3. Examine the examination questions and major assignments given in each course to identify the kinds of learning they call for.
4. Examine the collective results of each course's examinations and other important requirements as further indicators of the major kinds of student learning associated with each course.

Each of these steps in turn provides a closer approximation to what the college is accomplishing in terms of student learning. Each also requires greater effort than the preceding step. The last one completes a transition from the easy use of readily available college records to the combined use of records of course completions and direct evidence of course-related learning. This combination clearly provides the most detailed and comprehensive information on student learning of any of the procedures described, and far more so than any procedure now used.

Learning Surveys

While testing students' accomplishments is the most direct way to assess what they have learned, and existing records provide additional

information about the learning associated with different programs and combinations of courses, asking students or others who know them to report on their capabilities is often useful. When students are asked to describe what they know or have learned in a way that does not affect their grades, they tend to give information that is accurate and informative. Students who have just completed a course will report accurately how well they learned the material of the course in relation to the other students in the class. They will be less accurate in reporting how well they are prepared for a more advanced class or how well they understand the material in relation to students at other institutions. For those purposes, direct tests or judgments of others are needed.

The most effective use of student self-reports of learning is in comparing their accomplishments in different aspects of a course or program. If the questions are worded carefully, students can report accurately that they learned some elements of a course well and others poorly, identifying those parts of a course that might need greater emphasis. For such a purpose, asking students to report their achievement in relation to the others in the course would not be useful. Instead, they could be asked which of a list of course objectives they would feel comfortable explaining to other students and which they would have difficulty with. Others familiar with the capabilities of present or former students, such as faculty members or employers, can also accurately describe students' learning if they are asked appropriately--that is, if they are asked to report on student abilities they have had an opportunity to observe.

Two requirements are critical in getting useful information on students' accomplishments from their own self-reports or from reports

of faculty members or employers. First, the nature of the accomplishment being judged must be clearly stated. Asking students how well they mastered the learning expected of them in business management, for example, is less informative, since its referent is vague, than asking them how well they understand the preparation and analysis of financial statements.

The second requirement is that the students, or other persons making judgments, be given a measurement reference of some kind--a yardstick that they and others can understand. "How much do you know about the preparation and analysis of financial statements?" is more difficult to understand, by the person interpreting the response as well as the one giving the response, than, "In comparison with other students who have completed the second course in business management, would you put yourself in the top, middle or bottom one-third in knowledge of financial statements?"

Students can be asked how well they achieved clearly specified academic requirements in relation to (1) their own achievement six months ago or on entry into the course or program, (2) the ability of other students at their own level, (3) the instructor's expectations for the class, (4) the level of achievement they would like to have, or some other specifiable level. The way the question is asked will depend on the purposes for which the information is to be used. Assessing the overall achievement of all the students in a program might require a set of questions different from those used to evaluate the coherence of the existing curricular structure.

One procedure that has been used to assess students' perceptions of what they have accomplished in a particular field asks the students to complete two parallel self-rating sheets consisting of about 30 statements of program-related learning. The first asks them to judge their own

accomplishments relative to other students at their level. The second asks what portion of their current knowledge or understanding they had on entering the program. In conjunction with grades, other kinds of faculty judgments, or other kinds of information that would provide an interpretive context for them, students' self-ratings can be accurate and comprehensive indicators of educational accomplishment. Further, when judgments are made with respect to as many as 30 separate statements of learning, achievement on different course objectives can be compared.

Faculty judgments of the collective performance of their students with respect to various course or program objectives can be made fairly quickly and simply. Rather than making either a single overall judgment or a large number of individual judgments, faculty members may be asked to indicate where, among a large, representative sample of students, the best students in a class would rank, where the poorest would rank, and where the typical students would rank. These judgments would be most useful when applied to eight or ten important course objectives, indicating which parts of a course have been successful and which less so.

Employers can also give useful information on the capabilities of former students who are now employees, but that information is limited to areas of instruction that are related to the former student's current employment, as in nursing, secretarial work, or skilled trades. As with all surveys, employers should be asked to give their judgments of the abilities of former students only in those areas they can be expected to know about, that is, occupationally relevant performance. Preliminary interviews with a few employers are valuable in focusing a survey on useful issues. Since serving the local labor market is a common objective of community colleges, employers' judgments of former students are valuable sources of information.

Follow-up surveys

Follow-up surveys are defined in this handbook as surveys of former students that ask about their experiences since leaving college. In contrast, learning surveys ask about what students learned in college. Follow-up surveys are more common than learning surveys, but both can be useful.

A frequent flaw of follow-up surveys is that they provide information that is difficult to use. Asking whether students are in a skilled trade or office work, for example, can give a general but minimal indication of the degree to which the college courses taken were appropriate to the former students' current occupations. Similarly, former students' level of satisfaction with their courses may give college administrators some degree of satisfaction with their educational programs, but it too is minimal. Neither kind of information is clear enough to suggest an area of satisfaction or a concern explicit enough to indicate what the college might do to improve the quality of its programs.

More useful information would be provided by questions asking former students about particular aspects of their jobs, about the specific kinds of job-related knowledge or skills in which they felt weakest on entering the job, and about their current job activities in which they think college might have given them more preparation. Transfer students can similarly be asked about areas of strength and weakness they felt in their first term or two at a four-year institution. As with all surveys, the specific information desired must be carefully defined and the persons surveyed must have that information to give.

Surveys of former students are useful within a limited time frame. At least one year should elapse between leaving college and the time of

the survey to give the former students enough experience to be worth reporting. They will have had time to settle into a job or another college and will be able to report on how their community college experiences might have influenced their post-college experiences. From one to three years after leaving college may be the optimal period of elapsed time for follow-up surveys. Students who have left occupational programs will have had time to become established in their fields. Those who will persist in a field will have moved beyond entry level occupations while others will already have left the field.

Surveys of former students made five years or more after their leaving college have two problems. First, the influence of the former students' college experiences on their current activities will have been diluted by so many intervening effects that any inferences about college effects will be questionable. Despite exceptions in cases of people like nurses or dental hygienists, whose current occupational activities are clearly a product of their college studies, changes over a period of five years are typically too great for college effects to be clearly seen.

The second problem with long-term surveys--five years, for example--is that the relevant college experiences occurred five years or more in the past. The current college programs will already have changed, as will the economic and social climate entered by students currently leaving. If a college should change its curriculum or instructional procedures in the expectation of improving its success with students five years hence, another five years will elapse before comparable effects can be observed. The continual changes in a college's constituencies and environment and its own programs take most of the value out of long-term follow-up surveys.

Constructing useful surveys

The Item Bank: Learner Outcomes and Student Surveys, another product of this project, contains survey items in six broad areas that can be drawn on in constructing surveys. One of those areas is related to academic outcomes in the form of students' accomplishment of their academic goals and satisfaction with various aspects of their learning. Items such as those can be supplemented with more detailed questions about specific kinds of learning. Specificity, however, is achieved at the cost of a loss of scope, which makes clarity of the survey's purpose important in selecting or devising survey items.

Whether the persons surveyed are students, faculty members, or employers, and whether they are surveyed during or after college, several principles apply, two of which are basic. They are so obvious that mentioning them seems trivial, but they are frequently violated. First, ask questions that will tell you what you want to know. Second ask them of people who have the information. These principles are central to the following steps in the design of surveys.

1. Identify the purpose of the survey and the specific information that will serve that purpose.

   One typical purpose of surveys is to determine whether students in occupational programs are being adequately prepared for jobs. Another is to learn whether students are having unreasonably severe problems scheduling their needed classes. The statement of either of these purposes, or any other, will have arisen out of prior discussion of administrative or curricular needs, which will also be drawn on in specifying the information needed. The issue of job preparation may involve the availability of appropriate jobs, the success of former

students in finding initial employment in relevant jobs, the length of time former students stay in relevant jobs, their success on the job or simply the occupational qualifications of the students on leaving college regardless of whether or not they enter a job related directly to their college program. Each of these purposes, all related to the adequacy of students' occupational preparation, requires a different kind of information. The first requires a survey of potential employers. Success on the job requires a survey of actual employers. The others require surveys of different group of students--all students, all employed students, all students employed in jobs related to their college programs. The questions asked may refer to the former students' job history, job performance, or job satisfaction. All these and other variants of an occupational follow-up can be the basis for a useful survey. The specific information needed to answer the questions of primary interest must be clear before survey items can be selected or constructed.

2. Decide who can best provide the desired information.

Potential sources of survey information are current students, former students, faculty members, employers, possible employers, community members at large, and a variety of subgroups of each of these major groups, such as students or faculty members in various programs. Faculty members who have close relationships with employers can often give accurate information about occupational requirements and former students' occupational performance, eliminating the need for an extensive survey. Such information gathered systematically through interviews with faculty members may be followed by small, focused surveys to verify the perceptions of the faculty members or to fill in gaps in

their knowledge. When that is feasible, the cost of a major survey can be reduced.

In some cases, identifying the persons capable of giving the desired information may require a preliminary survey, making the full survey a two-step process. A post-card survey, for example, may identify the former students still employed in a job relevant to their college program a year or two after leaving college. A more extensive survey may then be addressed to just those students.

3. Ask questions that are explicit and that can be answered off the top of the head.

Questions that are vague, subtle, or general are likely to be interpreted in so many different ways that the summary of responses will be meaningless. They also aggravate the persons responding, sometimes to the point of throwing the survey in the wastebasket. Written surveys should be limited to simple questions of fact--how long have you been employed, what is your current job title, how many hours per week do you work. More complex or abstract information can best be gathered through interviews, perhaps after a factual survey has identified the people who can best give the desired information and are available for interviews.

Often the form of a question appropriate for a survey cannot be known without preliminary interviews. "How have you used what you learned in college in your present job?" is one such question that will require lots of thought and will produce such varied responses that their aggregation will not be useful. Asking that question in a small number of interviews, in which elaboration can be asked for, will produce a range of responses that can then be used to produce a survey

question of the following form: "In which of the following ways have you used what you learned in college on your present job?" (The question will, of course, be followed by a checklist derived from the interviews.) The persons responding will not have to guess at the kind of response desired or try to think up some way to describe what they learned that has been useful, but will simply check those that apply to them. Such a question might well be followed by one of a similar form, also derived from a small number of interviews, that asks what kinds of job requirements they did not get from college that would have been helpful.

In general, questionnaires used in surveys should require only checks as responses. They can be checklists, as in the example above, where the persons responding indicate which of a variety of options apply to them. They can be statements of possible experiences, such as number of different jobs held or number of weeks out of work, where the person checks some point on a scale of frequencies. They can also ask for qualitative judgments, as degree of satisfaction with the occupational training received at the college, on a four- or five-point scale from dissatisfied to satisfied. Statements of satisfaction, though of passing interest, are rarely useful unless elaborated with a list of possible reasons for satisfaction or dissatisfaction. After indicating the overall level, students can then be asked to indicate the reasons for their current attitudes. Again, the list of possible reasons to be checked should be derived from preliminary interviews with a small sample of the persons of interest.

4. Keep the survey questionnaire short.

A questionnaire that requires more than about 10 minutes to answer will usually not produce accurate information or a satisfactory response rate. The persons responding will tend to be those who are compliant, angry, or well satisfied. Many of the ordinary persons who went through a program in a satisfactory but routine way will not be interested enough to give more than five or ten minutes to a questionnaire. They will either stop part way through or not complete it at all.

If large amounts of information are needed, it should be separated into several questionnaires sent to several equivalent samples. At times, a two- or three-stage questionnaire can be sent to the same sample of people if the second and third stages are accompanied by results from the first or second stage, giving the persons something of interest in return for their continued effort. When more than one questionnaire is used with different samples, two or three questions of central importance might well be included on all questionnaires to get more accurate estimates of those pieces of information and to check the equivalence of the several samples.

5. Limit the scope of the questionnaire.

A common tendency to be avoided is to want to add just two or three more questions as long as the questionnaire is being sent out anyway. Everything worth knowing about a group of students or former students cannot be learned with a single questionnaire. A high value should be placed on restraint, keeping the questionnaire short and increasing the accuracy of the information returned. Questions should not be included unless a strong case can be made for the need for the

information. That it would be interesting to know is not sufficient reason. A clear use to be made of the information, as in modifying the curriculum or changing the scheduling of classes, should exist before a question is included on a questionnaire.

6. Having completed the planning of a questionnaire and produced a first draft, check to be sure it will give you the information you want.

Produce a set of imaginary results--the percentages you might expect in each response to each question. Give those speculative results to five or six faculty members and ask what they would infer from them. If they are uncertain about the inferences or disagree with the inferences you would like to make, the questions or the sample of respondents may need revision. If some inferences are uninteresting or trivial, the pertinent question should be deleted.

7. Distribute the revised questionnaire to a sample of from 50 to 200 persons.

Fewer than about fifty persons, unless a smaller number constitutes almost all the population of interest, are unlikely to give results stable enough to be useful unless very high percentages appear in single responses. Most information of interest in questionnaires is not so clearcut, and the results from fewer than 50 persons will be unreliable.

In contrast, samples of 200, or perhaps 300 when the accuracy of the information is critical, give results accurate enough for most purposes. Reducing errors by 3 or 4 percentage points is usually too trivial to justify the expense of getting responses from greater numbers. The accuracy of the estimate of a response percentage in an entire population increases only slightly as the number in the sample is increased beyond 200 or 300 persons. The size of the total

population has only a minor effect on errors in estimating the population percentage.

Larger sample sizes are needed when the entire sample is to be broken down into subsamples, perhaps by sex, age, program taken, number of terms completed, or type of job entered. In those cases, subsamples of around 50 are desirable, and the total sample should be large enough to provide 50 cases in the smallest subsample expected. If a large number of subsamples is desired, a survey may be carried out in several pieces with samples of manageable sizes. A survey in which comparisons are desired among men and women, from four separate programs, in three age groups will produce 24 subgroups and require responses from well over 1,200 persons if all the possible comparisons are to be made with subsamples of at least 50. As many as 2,000 responses may be needed to get 50 in every subsample, which typically would require questionnaires to be sent to 4,000 persons with a substantial follow-up effort. A more economical approach would be to identify the most important comparisons, which may exclude the need for sex or age comparisons in certain programs, and direct two or three separate surveys to smaller samples that will permit the desired comparisons.

When subsample comparisons are important, adequate subsample sizes can be assured and the accuracy of the overall estimates can be increased by selecting persons proportionately within each subsample in constituting the total sample. If a 50-percent nonresponse rate is anticipated, 100 persons should be selected in the smallest subsample, and the numbers in the other subsamples should be determined proportionately. Alternatively, if that produces a total

sample of unwieldy size, 100 persons can be selected in each subsample. Direct comparisons among subsamples will be possible, but combining subsamples, such as comparing results across two programs for men and women combined, will require adjustments in the observed percentages. Those adjustments are not complicated when the number of subsamples is small. Often, with numerous comparisons among groups and subgroups, they can be quite complicated.

In general, follow-up surveys are expensive, and because of low response rates and biased samples, the information they produce may have limited value. When the concern is with student learning, the optimal time for assessment is soon after the learning has occurred. When the interest is in the long-term effects of learning, the most useful approach may be a series of surveys spaced no more than a year or two apart to limit the intrusion of extraneous influences.

Uses of Learner Outcome Data

This handbook has focused on the learning characteristic of groups of students defined in various ways. It is not concerned with the learning of individual students except as elements aggregated to describe collective learning. Many procedures can be used to describe collective learning that would not be usable for individual learning.

Faculty members tend to have accurate though impressionistic knowledge of what their own students have learned. They have little if any knowledge of the learning of other students. Even with their own students, they don't know what similar students in similar classes elsewhere are learning. Department heads, deans, presidents, trustees, legislators, potential

students, employers, and taxpayers all have considerably less information than faculty members, and all have an interest in what students collectively have learned. The uses of information on learing will differ among those groups, which implies that different kinds of information will be required. Faculty members, department heads, and deans will want to know how well different aspects of courses and curricula are being learned, where in the curriculum and with what students learning is spotty, and where there is unnecessary redundance. Presidents want information they can present to the public, trustees, and legislators that will convince them the college is performing well. Most of the groups want to know for various reasons which programs are more effective than which comparable programs and in what ways. For each of these purposes, the particular groups of students or former students to be assessed and the nature of the assessment information desired will differ. For optimal usefulness, more than one procedure described might be selected and combined to suit the immediate purpose. An analysis of the course patterns taken by students identified in a follow-up study as either successful or not successful in a four-year college, for example, could be useful in guiding present students.

Cost is always a consideration, and the least costly method is to use existing institutional data. Even that cost increases, though, when the information is readily available only for students as a total group but is desired for certain subgroups of students. It increases further when the available institutional information, as on courses completed, must be supplemented with additional information, as on the learning expected in each course. A sufficiently wide range of procedures is available, though, that an acceptable compromise between comprehensiveness and detail on one hand and time and cost on the other can be reached.

SAMPLING THEORY AND SAMPLES

The attached material is taken from a Needs Assessment Handbook , authored by Jennifer Franz and myself in 1980. While the focus is on sampling from community populations, the same concepts apply to sampling from populations of students or former students. If necessary, the reader may want to refer to statistical tables which show the appropriate sample sizes from smaller populations; otherwise, the practical application is the same.

Chuck McIntyre
Director of Analytical Studies
State Chancellor's Office
California Community Colleges
September 1984

STEP 6: WHOM DO WE ASK?

Introduction: Sampling Theory and Samples

Clearly, conducting a survey of the entire community is only feasible in the smallest of areas - a town of a few hundred residents, for example. Happily, it is not necessary to do so in order to obtain results which *REPRESENT* the entire community. (The statistical wizards inform us that only communities of less than 15 residents need to be surveyed in their entirety.)

The question then becomes: how many people *DO* you need to survey, and who should they be? The answer to this question is based on the theory and practice of *SAMPLING* - that is, literally, selecting and viewing a sample of the community as being representative of the whole.

As Morris Slonim notes in his delightful treatise on the topic, Sampling, mathematicians, statisticians and researchers have obfuscated the subject of sampling to the point of terminal frustration among lay people. Technical terms and complex formulas abound, and it is easy even for the relative novice to bog down almost immediately in such complexities as "multi-stage stratified cluster sampling" or the "estimated sampling error for a binomial."

However, as Slonim so aptly points out:

> Everyone who has poured a highball into the nearest potted plant after taking one sip has had some experience in sampling. The abstemious reader doubtless has at one time or other pushed aside a bowl of tepid mush after swallowing a spoonful. He, too, has unwittingly employed a sampling technique. It is not necessary, one perceives, to have a graduate degree in mathematics to be reasonably proficient in sampling matters, in a practical sort of way.[32]

Fortunately for the non-mathmetically-minded reader, the author is of the same general opinion as Mr. Slonim. What you *DO* need to know about sampling in order to conduct a valid needs assessment survey is actually relatively simple and can be described without resorting either to esoteric jargon or to complex formulations.

To begin with, three general considerations. First:

- The confidence you have in your results will be directly related to the size of your sample.

HOWEVER

- In sampling, more is not necessarily better.

[32] Slonim (1960), p. 1.

People tend to think that more interviews or mailings means more representative results. This is simply not the case. If X interviews are sufficient under the laws of statistics, then X + 1 or even X + 100 will not be anything other than more than sufficient. More may, in fact, be worse. More cost will be incurred, more time will be taken (with the corresponding risk of changes which might affect responses), and more staff will be needed (with the corresponding risk of additional biases and/or the use of less than fully-qualified personnel).

Secondly:

- The confidence you can have in your results as well as the ways in which you will be able legitimately to analyze them will be directly related to the manner in which you select your sample.

Sampling theory is based on probability theory, which as anyone who has ever rolled dice, played cards, or even flipped a coin might suspect, has to do with the odds or chances of a particular event occurring.[33] When we flip a coin, for example, we know that we have a 50-50 chance that it will come up heads. Moreover, this chance will remain constant for every successive flip (to the dispair of those who live by hunches or patterns). We can therefore predict that if we flip said coin 100 times, the result will be heads 50 times and tails the other 50.

---

[33] It is perhaps illuminating in this regard that probability theory originated with a French libertine who wanted to find out what the odds were on a variety of his favorite gambling games.

This type of prediction only works, however, if we know what all the possibilities are (in this case just two - heads or tails) and we give each possibility an equal chance of occurring (that is, we flip the coin vigorously rather than tossing it carefully to get the heads we just bet on).[34] In short, we need to ensure that the outcome occurs at *RANDOM* rather than by intentional or unintentional design. All the predictions which sampling permits us to make and all of the methods of making them (i.e., the various statistical calculations which can be used) make this underlying assumption.

Third:

- The confidence you can have in your results and the extent to which you can legitimately subject them to statistical manipulation will be directly related to the manner in which you implement your sample.

In addition to the assumption of random selection, statistical procedures make the assumption that 100% of whatever sample was selected was in fact implemented (i.e., contacted and surveyed). It is therefore just as important that you survey all (or, to be practical about it, as many as humanly possible) of your initial sample as it is that the sample be of sufficient, randomly selected size.

---

[34] There is actually another way in which this type of prediction will work: if we know what any other-than-equal probability of a given occurrence is. However, this is as difficult to determine in most research situations as it is to assess just how carefully the coin was tossed.

This, then, is another reason why more interviews or more mailings are not necessarily better. You will actually be in a better position to have contacted and obtained responses from 90% of a minimum sample than to have surveyed 60% of some larger sample, even if the 60% represents a larger absolute number of responses.[35]

Even the uninitiated would suspect (and correctly so) that if we set out to flip a coin 100 times, we might not achieve a tidy 50-50 split between heads and tails. This is because we would be but one sample of 100 coin-flips. However, if we lined up ten students and set them to flipping coins 100 times each, the chances are good (to be nonstatistical about it) that the *AVERAGE* of their combined efforts would be extremely close to 50-50. It is this average of the results of multiple samples which is the foundation for all statistical calculations.

The population of coin-flips is infinite - our students could go on flipping coins forever (at least in theory) and never exhaust it. Most populations from which samples are drawn, however, are finite - at least in the social sciences. It is therefore at least theoretically possible to determine all the possible combinations of samples of various sizes which could be drawn from that population. Using this information, statisticians can predict just how much the results from any given sample are likely to deviate from the results we would get if we

---

[35] In addition to violating the mathematics which underlie statistics, low implementation rates can introduce serious biases into survey results. Those who are more difficult to reach differ in several respects from those who are readily contacted (Blankenship; Holmes and Glenn, Weaver), and non-respondents have been shown actually to have different perceptions than respondents (Blankenship; Donald; Erdos; O'Neil).

assessed the entire population (assuming random selection and assuming we used identical approaches to assessing the sample and the population). This deviation is referred to as the *SAMPLING ERROR*.

What this means is that we take a certain risk of being imprecise when we measure a sample rather than the entire population.[36] Typically, this imprecision is expressed by referring to whatever measure we derive from a sample as the *SAMPLE ESTIMATE* of the population value.

There are two components of this risk: *TOLERANCE* and *CONFIDENCE*. Tolerance refers to the degree of divergence we are willing to tolerate. We can specify, for example, that we want our results to be within 5% of the values we would get if we measured the entire population. Having made this tolerance specification, however, we have to concede that we might draw a peculiar sample and wind up with results that are more than 5% off. The extent to which we are willing to take that risk is called a confidence specification. We can express this in terms of betting odds that the chances are 99 to 1 or 95 to 5 (or whatever we decide upon) that our sample is off by no more than 5%.

---

[36] It is common practice to refer to a sample as being more or less precise rather than more or less accurate. The term *ACCURACY* refers to a "true" figure - which we might or might not get if we measured the entire population, errors of measurement being what they are. *PRECISION*, on the other hand, reflects the degree to which our sample approximates the results we would get if we measured the entire population in the same way - possible measurement errors included.

It is the assessment administrator who must set the tolerance and confidence specifications for the assessment results. In practice, most people (and all but one Project participant) set both at 5%, and the remaining discussion will use these figures. Those who wish a more precise estimate can find the corresponding numbers in most sets of statistical tables.[37]

Setting confidence and tolerance limits at 5% means you can be 95% confident that your results will be within 5 percentage points (plus or minus) of what the results would have been had you surveyed every resident: For example:

- 85% of your respondents can name your college.

THEREFORE:

- You are 95% confident that between 81% and 91% (inclusive) of all residents can name your college.[38]

This presumes, however, that your sample size is sufficient, that the sample itself was randomly selected and that it was fully implemented.

---

[37] Perhaps the best source is the Chemical Rubber Publishing Company's Standard Mathematical Tables, commonly referred to as the C.R.C.

[38] Actually, the greater the dichotomy of the response, the smaller the tolerance. In a sample of 400, for example, a response of 50% "yes" and 50% "no" is subject to a tolerance of ± 5%. A response of 90% "yes" and 10% "no," on the other hand, has a tolerance of ± 3%. As it happens, this represents the "estimated sampling error for a binomial." A full table of these estimates for various sample sizes and response levels can be found in Babbie, p. 376.

What is a sufficient sample size? As a general rule, any college or district with a service area population in excess of 30,000 can consider 400 surveys as constituting a sufficient sample. This is a rounded number, however, which is designed to be a convenient goal and to allow for some slippage due to unusable surveys. The actual numbers are displayed in Table III-1 below.

TABLE III - 1

POPULATION AND SAMPLE SIZES*

(95% Confidence; ± 5% Tolerance)

POPULATION	SAMPLE SIZE
10,000	370
15,000	375
20,000	377
30,000	379
40,000	380
50,000	381
75,000	382
100,000	383
500,000 to ∞	384

* Appropriate sample sizes are also dependent on the rate at which the characteristic(s) under examination occur in the population. The figures in this table assume the worst typical case in social research, which is a 50% rate of occurrence. (An example would be the respondent's sex.) Lower rates of occurrence - e.g., those who are attending a particular college (roughly 8% in most cases) - would allow smaller sample sizes. Inasmuch as the sample should be designed for the worst probable response, however, the figures in this table are the appropriate ones to use.

The astute reader will note that the relationship between population size and sample size represents an exceedingly flat curve. That is, the population can grow by leaps and bounds while the sample size increases by relatively minute increments. This is the primary reason why "more" is frequently an exercise in wasted resources rather than a guarantee of a more representative set of results.

Another way of viewing this phenomenon is to look at the increase in precision which accompanies an increase in sample size for a given population. Table III-2 below shows the tolerance specifications for various sample sizes for a population of 500,000 to ∞ at the 95% level of confidence.

TABLE III - 2*

TOLERANCE SPECIFICATIONS FOR VARIOUS SAMPLE SIZES
(Population ± 500,000; 95% Confidence)

SAMPLE SIZE	TOLERANCE
96	± 10%
119	± 9%
150	± 8%
196	± 7%
267	± 6%
384	± 5%
600	± 4%
1,067	± 3%
2,401	± 2%
9,604	± 1%

* Table adapted from the SAM Operations Manual (September 1978),p. 87.

As this table indicates, very little precision is gained from an increase in the sample size of as much as several hundred. Increasing the precision from ± 5% to ± 3%, for example, would mean almost tripling the sample size. For most purposes, the time and effort involved in surveying such a substantially larger sample would not be worth the relatively small gain in the precision of the results.

That having been said, however, there is a definite - and important - exception to this rule. This concerns what is referred to as *SUBGROUP ANALYSIS.*

The figures listed in Tables III - 1 and III - 2 are the sample sizes necessary and sufficient to draw conclusions about the population as a whole. In most instances, however, they will not permit conclusions to be drawn about population subgroups. Thus although we can be relatively definite about what all residents think, we would be on considerably more shaky ground if we attempted to compare the perceptions of men and women as distinct subgroups.

As a general rule of thumb, comparisons among subgroups should not be undertaken unless there are at least 100 responses in each subgroup.[39]

[39] Limiting subgroup analysis to groups of 100 or more gives a tolerance of roughly 10 percentage points at the 95% confidence level. Some researchers prefer to use a figure of 200, which gives a tolerance of around ± 7%, but this level of precision is usually not necessary given the kinds of objectives a needs assessment can reasonably address. Differences in responses of less than ± 10% will probably be of insufficient practical importance in most instances to warrant being acted upon.

The need for 100 responses poses no particular problem with respect to comparisons between men and women inasmuch as each group represents roughly half of the population. But suppose one of your objectives is to determine the relative needs of various ethnic groups in your service area. Consider the hypothetical example in Table III - 3 below.

TABLE III - 3

ETHNICITY - ALL GROUPS

Ethnic Group	Percentage of Population*	Number of Responses
White (Non-Hispanic)	74.8%	299
Mexican - American	10.1	40
Other Hispanic	3.0	12
Black	6.2	25
Asian	2.1	8
Native American	1.4	6
Other	2.4	10
	100.0%	400

* The actual figures used here and in subsequent examples are from the statewide survey and are not intended to be representative of any particular district's service area. The 1980 Census data will probably prove to be the most reliable source of comparable local data in the next five years.

Clearly, if you followed the "Rule of 100," you wouldn't be able to say anything about ethnic minorities. Not precisely what you had in mind, we're sure. Therefore: *BEFORE YOU DECIDE WHAT SIZE SAMPLE YOU NEED, YOU SHOULD DECIDE WHAT SUBGROUPS YOU ARE GOING TO WANT TO LOOK AT.* The problem with this is that you may find the number of responses you need to be prohibitive in terms of available resources. The example in Table III - 4 is illustrative.

TABLE III - 4

ETHNICITY - ALL GROUPS

Ethnic Group	Percentage of Population	Number of Responses
White (Non-Hispanic)	74.8%	5,341
Mexican - American	10.1	721
Other Hispanic	3.0	214
Black	6.2	443
Asian	2.1	150
Native American	1.4	100
Other	2.4	171
	100.0%	7,140

For purposes of illustration, the smallest ethnic groups are included in this example to show the total number of responses this approach implies. In practice, most researchers would probably concede that these groups could not be included in any subgroup analyses and either eliminate them, focusing only on the largest groups in the community,

or consolidate them into an "other" category. The results of two alternative approaches in this vein are shown in Tables III - 5 and III - 6 below.

TABLE III - 5

ETHNICITY - THREE MAJOR GROUPS

Ethnic Group	Percentage of Population	Number of Responses Needed
White (Non-Hispanic)	74.8%	1,268
Mexican - American/Hispanic	13.1	222
Black	6.2	105
Other	5.9	100
	100.0%	1,695

TABLE III - 6

ETHNICITY - TWO MAJOR GROUPS

Ethnic Group	Percentage of Population	Number of Responses Needed
White (Non-Hispanic)	74.8%	618
Mexican - American/Hispanic	13.1	108
Other	12.1	100
	100.0%	826

Finally, consider one of the more evenly distributed characteristics of respondents, namely household income. We will reinforce this characteristic's tendency to be evenly distributed by establishing categories so as to minimize small percentages, as shown in Table III - 7.[40]

TABLE III - 7

HOUSEHOLD INCOME

Gross Annual Household Income	Percentage	Number of Responses Needed*
Under $10,000	20.5%	100
$10,000 - $24,999	36.2	177
$25,000 and Over	27.5	134
Not Reported	15.8	77
	100.0%	488

*This calculation presumes we cannot say much about "not reported" as a subgroup in any event and are therefore not concerned about its being adequately represented.

[40] In practice, this approach to categorization is a risky business. Responses should be grouped in equal intervals (e.g., by ten-year increments for age) or in categories which reflect established standards (e.g., lower, middle and upper income groups in this example, if there were standardized cut-off points).

The third formulation of ethnicity and the somewhat contrived approach to income level are probably manageable given a reasonable assessment budget. More detailed analyses, however, are clearly beyond the scope of any reasonable assessment. What to do?

Most people don't. They simply report their data by subgroup and ignore the fact that the sampling tolerances will not permit meaningful comparisons. Readers are then free to conclude what they will. There is nothing inherently wrong with this approach *PROVIDED BOTH THE REPORT WRITER AND THE READER KNOW THE LIMITATIONS OF THE DATA.*

A somewhat more sophisticated alternative to this dilemma is to undertake what is called *DISPROPORTIONATE SAMPLING.* This approach, which usually follows the "main" survey, uses some type of screening mechanism (perhaps a brief interview) to identify potential respondents by subgroup. Respondents which fall into needed (smaller) categories are then questioned further, while those who fall into the larger groups are not. This technique can be frightfully costly, however, and will tend to be fruitless with respect to smallest subgroups (e.g., Native Americans). Furthermore, it poses problems with respect to analyzing all of the responses. In general, then, this approach may prove to be an unwise use of resources in all but the most critical of situations.