2

A Conceptual Model for Assessment

How can assessment be used to enlighten and inform college faculty and administrators? If an educator is interested in using assessment to learn how effective particular educational practices or programs are in developing student talent, it is not enough simply to go out and collect some "outcomes assessments." Unfortunately, a good deal of educational "evaluation" is done in this fashion, and as a consequence, it ends up shedding little light on the educational questions being investigated.

For several decades now we have been using what is called the input-environment-outcome (I-E-O) model as a conceptual guide for assessment activities in higher education. The I-E-O model is simple, yet it provides a powerful framework for the design of assessment activities and for dealing with even the most complex and sophisticated issues in assessment and evaluation. Because this model is something that evolved during Astin's early years as a higher education researcher, we will introduce it by providing a brief autobiographical account of how it originated.

EARLY LESSONS IN EDUCATIONAL ASSESSMENT (A. W. ASTIN)

My doctoral training in psychology and my early employment as a clinical and counseling psychologist in a variety of medical settings conditioned me to look at human behavior in a developmental framework; people come to you for help in a certain condition, and you strive to work with them in such a way as to improve their condition. The success of the treatment that you provide is thus judged in terms of how much the patient or client is able to improve.

Because some clients are in much worse shape than others when you first see them, you cannot judge the efficacy of your treatment simply in terms of the outcome (the condition of the patient at the termination of treatment); on the contrary, the effectiveness of treatment has to be judged in terms of how much *improvement* takes place.

My initial exposure to educational research occurred when I accepted a position as a research associate at the National Merit Scholarship Corporation (NMSC). Moving from clinical to educational psychology represented a major shift in orientation, but the problems in education seemed at least as interesting as—and probably more tractable than—those in the mental health field. The move also gave me an opportunity to work with a former mentor—psychologist John L. Holland—whom I greatly liked and admired.

My first research project at NMSC was concerned with something called *Ph.D. productivity*. The study was supported by the National Science Foundation, which, at the time, was concerned with finding ways to encourage more undergraduates to pursue graduate work, especially in the sciences. Researchers at Wesleyan University and the University of Chicago (Knapp and Goodrich, 1952; Knapp and Greenbaum, 1953) had found that certain colleges were much more likely than others to produce graduates who eventually went on to win graduate fellowships and to earn the Ph.D. degree. Because the highly productive colleges also tended to have larger libraries, smaller student-faculty ratios, and more faculty who themselves had Ph.D.s than did the less productive colleges, the researchers concluded that these superior facilities and resources were somehow responsible for the colleges' higher productivity.

Holland and I noticed that the highly productive colleges tended to be the same ones that the Merit Scholars preferred to attend. This fact prompted us to ask a rather simple question: Could a college's output of Ph.D.s be explained simply in terms of its initial input of talented freshmen? To test this possibility, we conducted a series of studies which showed that, as far as Ph.D. output is concerned, the student input is by far the most important determining factor (Astin, 1962, 1963). It turned out that, when you took student inputs into account, some of the so-called highly productive institutions were actually *underproducing* Ph.D.s, whereas some of those with more modest outputs were actually producing more than one would expect from their student inputs.

These early studies were critical in teaching us three fundamental lessons about assessment in higher education:

1. The output of an institution or program—whether we measure this in terms of how many graduates earn advanced degrees, how much money the alumni earn, or whatever—does not really tell us much about its

educational *impact* or educational *effectiveness* in developing talent. Rather, outputs must always be evaluated in terms of inputs. This is a particularly important principle for U.S. higher education, given the fact that the four thousand institutions in our system differ so greatly in the kinds of students they enroll.

2. An output measure such as Ph.D. productivity is not determined solely by a single input measure such as student ability. On the contrary, even in our earliest studies of this phenomenon we found that input variables such as the student's sex and intended major field of study are at least as important as ability in determining Ph.D. outputs.
3. Even if we have good longitudinal input and student output data, our understanding of the educational process will still be limited if we lack information on the college *environment*. Thus, it is one thing to know that your college overproduces or underproduces Ph.D.s, but quite another to understand why. What is it about the environment of a college that causes it to over- or underproduce? This last lesson suggests that input and output data, by themselves, are of limited usefulness. What we need in addition is information about the student's educational environment and experience (i.e., the particular courses, programs, facilities, faculty, and peer groups to which each student is exposed).

THE I-E-O ASSESSMENT MODEL

These early studies convinced us that any educational assessment project is incomplete unless it includes data on student inputs, student outcomes, and the educational environment to which the student is exposed (Figure 2.1). *Outcomes*, of course, refers to the "talents" we are trying to develop in our educational program; *inputs* refers to those personal qualities the student brings initially to the educational program (including the student's initial level of developed talent at the time of entry); and the *environment* refers to the student's actual experiences during the educational program. Environmental information is especially critical here, since the environment includes those things that the educator directly controls in order to develop the student's talents. A fundamental purpose of assessment and evaluation, it should be emphasized, is to learn as much as possible about how to structure educational environments so as to maximize talent development.

To place the I-E-O in a more familiar terminological context, we could also refer to the outcome variables as dependent variables, criterion variables, posttests, outputs, consequents, ends, or endogenous variables. Both environmental variables and input variables are types of independent variables, antecedent

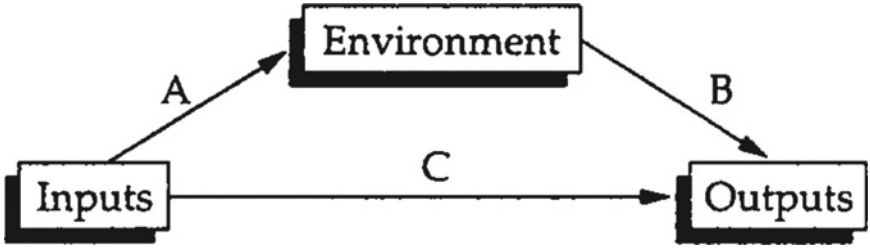

Figure 2.1 The I-E-O model.

variables, or exogenous variables. Inputs could also be called control variables or pretests. Environmental variables might also be referred to as treatments, means, or educational experiences, practices, programs, or interventions.

The three arrows in Figure 2.1 (A, B, and C) depict the relationships among the three classes of variables. Assessment and evaluation in education are basically concerned with relationship B—the effects of environmental variables on outcome variables. However, as the history of research on Ph.D. productivity shows, the relationship between environments and student outcomes cannot be understood without also taking into account student inputs. Student inputs, of course, can be related to both outputs (arrow C) and environments (arrow A). Another way of saying this is, first, that differences among students tend to show some consistency (i.e., correlation) over time (arrow C), and second, that different types of students often choose different types of educational environments (arrow A). The fact that inputs are thus related to both outputs *and* environments means that inputs can, in turn, affect the observed relationship between environments and outputs.

This problem can be illustrated with a simple example. Suppose we are concerned that many of our students do not seem to have very well-developed skills in English composition by the time they graduate and that we decide to try to learn whether there are particular course-taking patterns that facilitate or inhibit the development of writing talent among our students. Accordingly, we administer a test of skill in English composition to all graduating seniors (outcome measure) and compare the average test performance of students who took different patterns of courses (different environments). We might find, for example, that students who majored in engineering do relatively poorly on the test, whereas those who major in journalism do substantially better. Would such a finding justify the causal conclusion that majoring in engineering is detrimental to the development of talent in writing and that majoring in journalism facilitates the development of writing talent? Probably not. Is it not reasonable to suppose that students who as freshmen choose to major in

journalism *already* have better-developed writing skills when they first enter college (input) than do students who choose engineering? If so, we would expect journalism majors to score better than engineering majors on the senior test of writing skill, *even if the different course-taking patterns had identical effects on the development of writing talent!*

The basic purpose of the I-E-O design is to allow us to measure relevant input characteristics of each student and then correct or adjust for the effects of these input differences in order to get a less biased estimate of the comparative effects of different environments on outputs. (Details of how to make such adjustments and interpret the results are given in chapter 6 and the appendix.)

Perhaps the need for these three kinds of data can be better understood with an analogy from the field of horticulture. Suppose we go to a county fair and examine the different entries in a rose contest. Although it might be interesting to observe that some people's roses are bigger, more beautiful, or more fragrant than the roses of others, such output information, by itself, is not very useful in telling us how to grow roses successfully. We might improve our understanding somewhat if we also had input information on the types of seeds or cuttings that each grower had used. But would we be justified in concluding that output differences in rose quality were simply a matter of input differences in the seeds or cuttings from which they grew? Clearly we would not. What is missing here, of course, is environmental data concerning the conditions under which the different roses were grown (e.g., type of soil, method of planting, light, fertilizers, watering schedule, and fungicides and pesticides used). These environmental factors are important considerations in how effectively the grower can develop the rose's "talent."[1]

In other words, simply having input and outcome data of a group of students over a period of time is of limited value if you do not know what forces were acting on these students during the same period of time.

Perhaps an even better analogy can be found in the field of health care. The basic evaluation problem in medical research is to learn which treatments (environments) are most effective. If we were trying to enhance our understanding of how best to treat patients in a hospital, imagine how difficult it would be if all we did was to collect output information on how long patients stayed, whether they lived or died, and what their condition was as they left the hospital. We would improve the situation considerably if we also got input (diagnostic) information on the patients' condition at the time of admission. But we would still be greatly handicapped without environmental data. That is, how could we expect to learn much about how best to care for our patients if we did not know which patients got which therapies, which operations, or which medications? This is the equivalent of studying student development

with no environmental data on what courses they took, where they lived, how much they studied, and so on.

There is nothing magical or even necessarily real in the I-E-O model. For us it represents nothing more than a convenient way of looking at phenomena that interest us—a tool for trying to understand why things are the way they are and for learning what might be done to make things different if we feel the need to change them. The model seems applicable to almost any social or behavioral science field—history, anthropology, economics, sociology, psychology, or political science—as long as the interest is in studying the development (input to output) of human beings or groups of human beings and in understanding more about factors (environments) that have influenced (or might influence) that development. We are using education (especially higher education) as the focal point for most of our discussion of the model, but we see no reason why it could not be used just as readily in any of these other fields.

Although most of the illustrations and applications of the model used in this book are *quantitative* (i.e., they involve quantifiable measures of inputs, environments, and outcomes and statistical analyses of the data), the logic underlying the model would seem to apply equally to *qualitative* problems. Qualitative research, like quantitative research, ordinarily seeks to identify causal connections between certain antecedent events or conditions (environments) and certain subsequent events (outcomes). Even if no quantitative data are involved, the qualitative investigator who is striving to understand why a certain event (outcome) occurred would be well advised to consider the possible contribution of inputs as well as environment.

Let's take as an example one of the most primitive forms of qualitative assessment: the testimonial. A testimonial is a verbal statement by an individual that is basically causal in nature. In effect, a testimonial attributes a particular outcome to the effects of a particular environment: "that teacher [environment] really helped me to understand calculus [outcome]." Note that the testimonial always implies an environmental *variable*, in the sense that it implicitly argues that some other environment (e.g., no teacher or a different teacher) would have produced a different outcome (less knowledge of calculus).

Often the testimonial also implies an input "pretest" condition (e.g., the student's lesser knowledge of calculus prior to encountering the teacher), but it usually ignores other inputs that might have had an important bearing on the outcome (e.g., the student's degree of determination to learn calculus).

What are we really trying to accomplish by applying such a model? First, it is important to keep in mind what higher education is attempting to accomplish, that is, to enhance the educational and personal development of its students and faculty. (To simplify this particular discussion, we shall focus on

student development, but remember that practically everything said applies equally to faculty development.) Taken together, student input and student outcome data are meant to represent student development—*changes* in the student's abilities, competence, knowledge, values, aspiration, and self-concept that occur over time. Because the notion of change is so basic to the purposes of higher education, we need to have at least two (and probably more) snapshots of the student taken at different times to determine what changes have actually occurred. At the same time, knowing what particular environmental experiences each student has had helps us to understand why some students develop differently from others.

Input and outcome refer simply to the state of the person at two different time points, and environment refers to the intervening experiences. We are particularly interested in learning about environmental experiences that can be controlled or changed because it is these experiences that offer the possibility of improving outcomes in the future. Some environmental experiences, of course, can't be controlled. It is one thing to know that a death in the family (environmental event) contributed to a decline in a student's performance, but quite another to know what could have been done to prevent it or what can be done to prevent such events from occurring in the future. By contrast, if we know that a particular teaching method or particular curriculum is better than others, we are in a much better position to use such findings in designing educational environments that will produce more favorable outcomes in the future.

Nothing in human experience is intrinsically an input, an output, or an environment. How we should assign these labels depends entirely on what aspects of experience we choose to study and how we formulate the questions we wish to answer. To see why this is so, we can look at a single variable: the student's score on the SAT taken in the senior year in high school. We might want to know why students score as they do on this test and to find the most effective ways to help future generation students achieve better scores. In such a case the SAT score would probably constitute an *outcome* measure. For possible environmental measures we have an almost infinite range of possibilities: the type of secondary school students attended, the kinds of courses they took, the quality of teaching they received, whether they took preparation courses for the SAT (and which ones they took), how they prepared for the test, what kind of peer group stimulation they had, and what kind of home environment they had. For input measures we would, obviously, need some sort of pretest, maybe the PSAT or a special previous administration of the SAT. Our choice here would be determined in part by the period of time covered by our environmental variables (e.g., the last year of high school, the last two years, or whatever). We would also need to assess a variety of other input

variables (e.g., sex, ethnicity, or socioeconomic status) that might affect SAT performance, especially if these variables could also affect the set of environmental variables to which the students were exposed.

But we might have a different interest in the SAT. Perhaps we wish to evaluate its usefulness in college admissions. From this particular perspective we would probably consider it an *input* variable and select such variables as college GPA, retention, or GRE performance as our outcome measures. Even in this situation, we would no doubt want to include environmental variables such as college major and place of residence in our analysis, since the effects of SAT on certain outcomes may be mediated by such variables. For example, it has been well established that student retention (completion of the baccalaureate) is facilitated by living in a campus residence hall during their freshman year (Astin, 1975, 1977, 1982, 1993; Astin and Oseguera, 2005; Chickering, 1974; Pascarella, 1985). Because students with high SAT scores are more likely to live on campus than students with low scores (Chickering, 1974; Karabel and Astin, 1975), it may well be that the "effect" of SAT on retention is an indirect one which is, in fact, mediated by campus residence and possibly other environmental variables.

Still another perspective on SAT scores is to use them to construct an *environmental* variable. It has long been recognized (Astin, 1993; Feldman and Newcomb, 1969) that one of the most important sources of environmental influence on students is the peer group. We could define each student's peer group as all the students who were majoring in the same field. If we were to compute the average SAT scores of the students separately by major, we would then have an estimate of the average ability of the peer group within each major field. A variation on this idea would be to use the SAT score of each student's roommate (or the average score if there is more than one roommate) as an environmental measure (see chapter 5).[2]

COMPARATIVE EVALUATION AND THE CONTROL GROUP CONCEPT

Educational policy making and educational decision making in general inevitably involve choices among alternatives. The student can decide to go to college instead of going to work, joining the military service, becoming a homemaker, traveling, or just loafing. The student can choose college A over college B or C or decide to live on campus rather than at home. The student can also pick a particular college major over dozens of others or decide to put off the decision for a few years. Finally, the student can decide how to go about studying and how much effort to devote to it. For their part, college

officials make choices when they decide to offer particular programs, to hire particular faculty or staff, or to set particular standards of performance. Similarly, the faculty make choices when they decide what to teach, how to teach it, how to counsel and guide students, what to read or what problem to study, and how to treat their colleagues.

The point is a simple one, but one that is frequently overlooked in treatises on assessment and evaluation: all educational evaluation is "comparative," in the sense that whatever is being evaluated is being compared with something else. Often these comparisons are implicit rather than explicit, and often neither the evaluator nor the decision maker is aware of what the (implicit) comparison really is.

To take the most primitive kind of educational assessment and evaluation as an example, a common educational assessment practice is to administer some kind of standardized competency test to students as they reach an important educational transition point. Such testing has been especially popular in the elementary and secondary schools for many years and has lately been gaining popularity in higher education. Many community colleges, for example, require such a test (the ACT's Collegiate Assessment of Academic Proficiency) as an exit exam for their associate degree programs. Although the original intent of such testing was that of quality control—to establish and maintain minimal performance standards for persons before conferring an academic degree or certificate—the temptation to aggregate such scores for certain groups of students (e.g., by institution) may eventually prove too great for many educators to resist. Suppose we were trying to evaluate a particular course and that we used this approach by administering some kind of test to the students just as they completed the course. As with any other evaluation problem, our ultimate interest is in decision making: to continue the course as is, to make certain modifications, to revise completely and even abandon it, to recommend or not recommend it to others, and so on. By looking at how well or how poorly the students do on the end-of-course test, we then make a judgment about the effectiveness of such things as the syllabus, homework assignments, the teacher, the teaching method, and the course in general.

To simplify this discussion, suppose we are the teachers and we do this evaluation to determine whether changes in the course are needed. The wary reader may by now have already detected a flaw in our approach: we have an output measure but no input measure. How can we know how much our students have learned during the course if we don't know how well they were doing at the start of the course? But suppose we are sophisticated enough to have also administered a pretest (input measure) at the beginning of the course. Now we can determine how much *improvement* took place between the beginning and end (input to output).

Let us assume further that we are not satisfied with the amount of improvement that took place in a certain aspect of the students' performance, and on the basis of that evaluative judgment, we decide to change something about the way we teach the course. In effect, what we are doing here is comparing the actual improvement that occurred under our current teaching approach (environment A) with what we expect to happen under the new approach (environment B). That (implicit) comparative evaluation has led us to conclude that the outcome would be better under the new approach. Note that we must assume not only that the particular outcome performance that concerns us under the old approach will improve under the new approach, but also that all other aspects of the student's outcome performance will be at least as good under the new approach (i.e., that there will not be any undesirable side effects caused by the new environment).

Exactly the same kind of comparative judgment would be involved if our evaluation led us to conclude that nothing needed to be changed. In effect, such a decision is based on the assumption that our current method of teaching the course (environment A) produced an overall outcome performance that is just as good or better than what could be produced under all other approaches that we might consider (environments B, C, D, and so on).

All educational choices, regardless of whether they result in a decision to change something or a decision merely to keep things as they are, involve comparative judgments such as the one just discussed. A decision to change something implies that the new environment is expected to produce a better outcome than the current environment. When the decision is not to change anything, the current environment is judged to be equal to or better than all possible alternative environments.

Control Groups and the True Experiment

Comparisons of the type discussed here are similar in principle to what experimental scientists call the *control group* approach. In experimental science, we try to understand the effects of a particular environment by simultaneously studying the effects of at least one other environmental situation and comparing the results. Typically these two situations are called the experimental condition and the control condition, respectively. One group of subjects or cases is exposed to the experimental condition and a second, equivalent group is exposed to the control condition. The idea is to try to make the environments of the two conditions identical in every respect, with the exception of the one variable of interest, which is deliberately made to be different in the experimental group (in experimental jargon, this is *manipulating* or *controlling* the independent variable, which is also sometimes called the *treatment*). Through

the processes of random selection or matching, the groups of people exposed to the two situations (the experimental group and the control group) are presumed to be equivalent at the start (input). If the outcome performance of the experimental group turns out to be different from the outcome performance of the control group, the experimenter is justified in concluding that the difference was caused by the environmental variable of interest because the two groups were comparable in every other respect.

We can illustrate the control group approach with an example. Let's say that we wish to introduce a radically different approach to teaching English composition in our undergraduate curriculum, but that there is some controversy within the faculty about whether the new approach will really work. We agree to conduct an experiment. We will select 10 percent of next year's new freshmen as an experimental group who will be given the new course, and the remaining 90 percent of the students will constitute the control group. Although in reality we would probably want to use several different measures to compare the outcomes of our two groups, let us temporarily assume that we use only a single outcome measure consisting of a test of competence in English composition. If we picked our 10 percent of the freshman class by lot (i.e., randomly), and if the rest of the curriculum and freshman year experience was comparable to that of the other freshmen (the control group), then we would be justified in saying that we had a true experiment. In fact, if the experimental group was really selected by lot, we would not even need a pretest input measure because we could assume that the two groups' average levels of skill in English composition were comparable at the beginning when they first started the course.[3] Then, if the two groups' average performance on the test of composition ability (outcome measure) turned out differently, we would be justified in concluding that the two approaches to teaching English composition produce (or cause) different results.

The approach taken in this hypothetical example is, unfortunately, used all too infrequently in academe. Typically, proposed changes in the curriculum are either implemented across the board for all students or, more typically, not implemented at all because of resistance and controversy within the faculty. But note here that such decisions involve precisely the same kind of logic that one finds in a control group experiment. When the faculty decides to change the curriculum, it has (implicitly) reasoned that if it *had* done a controlled experiment, the results would have favored the new curriculum. A negative faculty decision on the new curriculum also involves similar reasoning, but with a different conclusion: they assume that the experiment would *not* have favored the new curriculum.

In recent years, policy makers have pointed to experiments with randomized assignment as the "gold standard" for "scientifically based" research and

assessment in education, citing nonexperimental and less elegant methods unfit on which to base educational policy (Shavelson and Towne, 2002). Although it is true that a classical control-group experiment generates results about which one can make causal inferences with a high degree of confidence, control-group experiments in education (or in any other social science field, for that matter) create a number of other problems which, in our experience, greatly limit their usefulness. Let us see why this is the case.

First, a fundamental limitation of randomized experiments in education is that it is virtually impossible to create the same "double-blind" conditions that are typically required in medical research in which the effects of drugs are being assessed. By *double-blind* we mean that the students are not aware of which "treatment" they are receiving and the classroom teachers are not aware of which "treatment" they are administering.

A closely related problem is that when we conduct a true experiment by assigning students at random to experimental and control groups, we create a highly artificial situation that can distort our findings. Because it is difficult (and possibly unethical) to keep the students from knowing what is being done to them, they usually know that they are participants in an experiment, and they usually know the group to which they have been assigned. This knowledge will almost certainly affect the results of the experiment, and unfortunately, such effects are frequently unpredictable.[4]

We can see how this might work in the curriculum experiment just described. Students who have been assigned to the new course in English composition might resent the fact that they are being used as "guinea pigs," a reaction that in turn could have a detrimental effect on their motivation to perform well in the course. On the other hand, they might feel that they are sort of an elite group that has been singled out for special treatment, a response that might stimulate them to work especially hard. Students in the control group might feel grateful that they have been spared the fate of being used as guinea pigs. Or conversely, they might resent having been deprived of this innovative and exciting new course. Because nobody can be sure just how students in each group are really being affected by the knowledge that they are part of an experiment, it is not possible to know how this knowledge will ultimately affect their performance on the outcome measure.

From a practical point of view, the real problem here is not so much that the experimental results can be affected by the students' knowledge of the experiment, but rather that *the environmental conditions created by the experiment cannot be reproduced in the future.* In social science parlance, the external validity of the results is poor. Suppose the outcome of the experiment clearly favors the new English composition course and that this result prompts the faculty to replace the old course with the new one. Now all students must

take the same new course, and the students no longer think they are part of an experiment; the course is now simply another part of the required core curriculum that everybody has to take. How can we be sure that the course will continue to have the same beneficial effect? How do we know that the superiority of the new course in the experiment was not just a temporary consequence of student enthusiasm generated by the knowledge that they had been singled out for special treatment? How do we know that the inferior performance of the control group was not just a consequence of their indifference or resentment?

Similar problems arise when we consider the effects of the experiment on the faculty. The real dilemma here is how to assign faculty responsibility for the experiment. The rules of classical control group experimentation require that we do one of two things: either assign the faculty to teach the two composition courses by lot or have each faculty member who teaches the traditional course also teach one section of the experimental course. Neither of these requirements is entirely satisfactory, but the second is probably preferable to the first because it simulates more closely the results that would occur if a policy decision were subsequently made to drop the old course and have all students take the new one. And again, the unique conditions under which faculty are teaching reduces the external validity of the experiment or the confidence that one has that the results will be the same absent the experimental conditions.

The point of this discussion is to stress that control-group experiments in education are no panacea. We can learn much from true experiments, but they are not necessarily preferable to *natural* experiments, which we will now consider.

Natural Experiments

The I-E-O model was developed primarily for use in what we like to call *natural* experiments. In such experiments we try to study naturally occurring variations in environmental conditions and to approximate the methodological benefits of true experiments by means of complex multivariate statistical analyses. In a sense, with natural experiments we try to study the real world rather than the artificial ones that are created by experimentation. Natural experiments have two principal advantages over true experiments. First, they avoid the artificial conditions of true experiments that are created by the establishment of experimental and control groups and the random assignment of students to these groups. Second, natural experiments make it possible to study the effects of many different environmental variables at the same time. Because natural experiments permit us to compare and contrast the great variety of educational approaches and practices (i.e., the different environments) that characterize higher education in the United States, they can help us to

understand which educational environments and practices are most effective and under what conditions.

The principal limitation of natural experiments—and it is a serious one—is that the students are not assigned at random to the various educational environments. Another way of saying this is that the input characteristics of students who are exposed to one environment are usually different from the input characteristics of students who are exposed to another (comparison) environment. Students to some extent pick their environments and environments sometimes pick their students. This inequality of inputs means that the outcome performance of students exposed to different environments will almost certainly differ, even if the actual *effect* of the different environments is the same. The main purpose of the I-E-O model is to control for the effects of initial student input differences by means of multivariate analyses (see chapter 6 and the appendix for details of these procedures). In effect, such statistical equating of initial input differences attempts to accomplish by statistical means what random assignment accomplishes in pure control group experiments. The real question about any natural experiment is this: have all of the potentially biasing input variables been adequately controlled? Chapters 4 and 6 and the appendix suggest a number of specific techniques for addressing this question.

To return to our hypothetical example involving the new method of teaching English composition, if a few members of the English department became interested in such an approach, they might well want to try it out in some of their classes and to evaluate the results using a natural experimental design rather than the classical control group experiment described previously. They might persuade some of their colleagues who would be teaching with the traditional approach to let their classes serve as natural (i.e., nonrandomized) control groups or possibly they themselves could teach different sections using the two methods. No matter who does the teaching, it would be important to obtain input as well as outcome data from students in all classes. The input data should obviously include a pretest measure of competence in English composition as well as measures of any other characteristics (e.g., sex, prior grades in English courses) that might affect students' outcome performance.

Before leaving this discussion, we would like to add a word about a topic that has been much debated among social science methodologists: correlation and causation. Graduate students in education and the social sciences are routinely told that only true control group experiments permit the investigator to make causal inferences about environmental effects on outcomes, and that "you can't make causal inferences from correlational data" (this, of course, is the equivalent of saying you can't make causal inferences from natural

experiments). The fact is that you *can* make causal inferences from correlational data; people make such inferences all the time. Indeed, it would be hard to get through an average day without making such inferences. Each of us implicitly makes a causal inference when we make choices between alternatives, such as what to eat for breakfast, how to spend our day at work, and so on. Each decision implicitly involves causal reasoning—the selected alternative is assumed to lead to a better outcome than the rejected alternatives—even though we almost never have data from a pure control group experiment to help us choose.

The real issue in making causal inferences from correlational data is not that such inferences are methodologically unsound or immoral, but rather to *minimize the chances that our inferences are wrong.* In natural experiments, the best insurance against making invalid inferences is to control as many of the potentially biasing input variables as possible. Although we can never be sure that we have controlled all such variables, the more we control, the greater confidence we can have in our causal inferences.

INCOMPLETE DESIGNS

Perhaps the best way to understand the importance of the three components in the I-E-O design is to consider what happens when one or two of the three components are missing. Because typical assessment activities in higher education more often than not leave out components from the I-E-O model, we would like to discuss these incomplete designs using examples taken from real-life experiences that we have encountered on various college campuses. Four different incomplete designs will be considered: outcome-only assessments, environment-outcome assessments, input-outcome assessments, and environment-only assessments.

Outcome-Only Assessments

With the accountability and "learning outcomes" movements gaining so much popularity during the past few years, outcome-only assessment is probably the fastest growing approach of all. This approach involves the use of some kind of end-of-program assessment designed to determine whether the learning objectives of the particular program are being achieved. The most common application of this model is the course final exam, which generally has little application beyond the course where it is administered. However, the outcomes-only model is increasingly being used in much broader contexts that allow comparisons across professors, departments, and institutions. Most

institutions require all students to demonstrate some minimal levels of competency in basic skills such as mathematics and written composition before they are permitted to take college-level courses. Similarly, some public systems now require undergraduates to demonstrate minimal competency in one or more areas before receiving an associate or bachelor's degree. A good example is the upper-division writing requirement now mandated in all nineteen campuses of the California State University system. On an even broader level, Florida's College Level Academic Skills (CLAS) program requires demonstration of competence (through testing or satisfactory coursework) in communication and mathematics as a condition for achieving status as a junior. Finally, the outcome-only model is being used even at the national level. The National Assessment of Educational Progress (NAEP) periodically examines national samples of students at various levels of educational development to determine their skill levels in a variety of areas. Similarly, the national college admissions tests (the SAT and the ACT) have been used as a kind of annual barometer to gauge the effectiveness of our elementary and secondary school systems across the country. The sharp declines in these scores that occurred between the 1960s and the early 1980s provided the principal empirical foundation for the widely discussed critical report on our educational system, *A Nation at Risk* (National Commission on Excellence in Education, 1983).

The main advantage of assessments that use this model is that they focus attention on the fundamental problems of defining and measuring those outcomes that are relevant to the goals of the educational program in question. Even the process of trying to define and measure the goals of educational programs can be a useful learning experience for faculty members and policy makers. The major drawback to this approach, however, is that it produces data that are extremely difficult, if not impossible, to interpret. In other words, the *meaning* of the data generated by this approach is unclear.

Ambiguities and interpretive difficulties occur at all levels at which the outcome-only model is applied. Let's start with classroom final exams. Without additional information, the professor who attempts to evaluate his or her teaching using the course final exam is implicitly forced to assume that *what is being tested is what has been learned.* In most academic fields, such an assumption is difficult to justify. There are few courses, for example, in which students do not begin with at least some knowledge of the course subject matter. And students usually differ in this respect; some know much more than others about the subject matter before the course ever starts. Furthermore, most course final exams test a lot more than knowledge of course content because exam performance is affected by factors such as writing skill and reasoning ability. All of us who have taught college students over the years know well that if a student is sufficiently bright and talented at the start of the course, it is

possible for that student to do quite well on a final exam without really learning much of anything in the course. On the other hand, it is possible that a student whose performance on the final exam is mediocre may, in fact, have learned a great deal in the course, especially if the student began the course with no knowledge of the subject matter and with minimal examination performance skills. Perhaps the only time a professor has a reasonable basis for assuming that what is being tested is what has been learned is when the course has such highly specialized content that it would be unreasonable to assume that the students had any knowledge of the content prior to enrolling in the course. Outside of a few courses in highly technical fields or in certain natural science fields, it seems safe to assume that such situations are quite rare.

Problems associated with the application of the outcome-only method are compounded whenever the method is applied on a broader scale beyond the classroom. Take the much-heralded *A Nation at Risk*. The steadily declining college admissions test scores were cited in this report as one of the principal bases for concluding that the nation was "at risk." Although such a conclusion may have indeed been warranted by the data, the real problem is to understand *why* the decline occurred and what can be *done* about it. Is the problem with the high schools? To answer this, one would have to know how much improvement in performance the most recent classes of students exhibited during their three or four years in high school and to compare the results with a similar longitudinal assessment done during the late 1960s. If such a study were to show that the problem was not in the high schools, we would be confronted with other questions. Was the problem at the primary or intermediate levels? Or was it at the preschool level? To answer such questions, of course, it would be necessary to have input data at the beginning of each school level as well as outcome information.

But even if the decline could be isolated in terms of school levels, we would still be confronted with the even more difficult problem of understanding why the decline has occurred and what can be done about it. If we did indeed locate the problem in our secondary schools, what are we doing in the secondary schools that has created the decline? A definitive answer to this question would require us not only to have input and outcome assessments at the beginning and end of secondary school but also to have such assessments on different types of secondary schools and school programs. Because we lack such data, a great deal of effort has been invested in speculating about the reasons for the decline. Dozens of theories have been proposed ranging from changes in the curriculum to radioactive fallout from atmospheric testing conducted during the 1950s (Turnbull, 1985; Wirtz, 1977). The commission that produced *A Nation at Risk* concluded that the decline was caused in part by changes in the school curriculum and thereby recommended substantial

increases in the number of basic academic subjects that students should be required to take in secondary school. Although such curricular changes may indeed have a beneficial effect on test scores, available test data were really of little help in assisting the commission to come to such a conclusion and, in the long run, the commission was forced to resort to hunches and guesswork in making its recommendations. For all we know, the test score declines were not caused by curriculum changes. Perhaps there are many more cost-effective ways in which these declines can be turned around.

In short, the outcome-only approach to assessment is flawed on two accounts. First, there is no way of knowing how much has actually been learned as a result of an educational program because there is no input information with which to compare the outcome assessment. Second, in the absence of information on how students performed under different environmental circumstances, there is no way to tell from the assessment data which educational programs and practices are likely to be most effective.

Environment-Outcome Assessments

The environment-outcome approach to assessment represents an improvement over the outcome-only approach in that it incorporates information on environmental differences that can aid in the interpretation of student performance on the outcome assessments. However, this improvement can well turn out to be counterproductive because it encourages causal interpretations of environmental effects when these may indeed be unwarranted. The principal limitation of this approach is that no information on student input performance is included.

There are many examples of the use of the environment-outcome approach. Some institutions, for example, compare retention rates of students across different majors or between different colleges within the university. At the multiinstitutional level, different institutions can be compared with each other in terms of their retention rates, alumni achievements, and so on. The "Ph.D. productivity" studies discussed at the beginning of the chapter represent another example of the use of this approach.

As we have already shown, the main difficulty with the environment-outcome approach is that it exercises no control over differential inputs. The only situation in which we would be justified in concluding that output differences across different environments were, in fact, caused by the environmental differences is one in which the students have been assigned at random to the different environments (i.e., when we have the conditions of a true experiment). A possible exception to this caveat is the situation in which, although the subjects are not assigned at random, we have good reason to believe that

there are no important differences in input characteristics of students entering different environments. However, without actual input data, such assumptions are usually difficult to defend.

Perhaps the most egregious application of this model occurs in the achievement testing done annually in the public school systems of our states. Typically, the students at different schools within a system are examined on some achievement test and the average results are computed on a school-by-school basis. Each school is thus regarded as a different "environment." Schools in which the students get the highest average scores are thus presumed to be the best schools whereas those whose students get the lowest scores are considered to be the weakest schools. If we had reason to believe that the students entering the different schools were comparable at the point of entry, such causal conclusions would perhaps be justified. However, as we all know, different schools recruit students from vastly different socioeconomic backgrounds; their input levels of performance are almost certainly different. Under these circumstances, we would clearly expect to find outcome differences in achievement from school to school, even if the schools had identical true effects on the students' educational development. It may well be that many of the schools whose students do well on such achievement test comparisons are doing a mediocre educational job with their students and that some of the schools whose students do relatively poorly are, in fact, doing an outstanding job. Without input information on the students' initial levels of achievement and family background, there is simply no way to know how effective the different educational programs of the different schools really are.

Such problems are compounded in U.S. higher education, what with the enormous diversity of student bodies entering a variety of institutions. Even within many institutions, there can be substantial input differences between students who pick different majors, between commuters and residents, between part-time and full-time students, and between financial aid recipients and students who receive no aid. There is no way that we can reliably assess the impact of environmental experiences such as major or place of residence without input information on the characteristics of students at the point of entry.

Input-Outcome Assessments

Perhaps the prototypical study of college impact involves the testing and retesting of students at a single institution (Feldman and Newcomb, 1969). Characteristically, students complete some kind of questionnaire or inventory when they first enter college and take it again one year later, four years later, or in a few cases, many years after graduation. Measures of change or growth are obtained by comparing the students' input scores from the initial

administration with outcome scores from the follow-up administration. In subsequently interpreting these change scores, we typically assume that any observed changes are the result of the students' experiences in the educational program. In other words, such studies equate *change* with *impact*.

When such assessment studies involve the use of achievement tests or other cognitive measures, they are sometimes referred to as *value-added* assessments. We personally prefer the term *talent development*, for at least two reasons. First, the value-added concept is basically economic rather than educational in its derivation. Second, talent development seems to come much closer to describing the fundamental educational mission of most colleges and universities. Nevertheless, it should be recognized that the terms *value-added*, *talent development*, *pretest-posttest*, and *longitudinal* basically relate to the same phenomenon: repeated assessment of the same qualities on the same students done at different points in time.

This type of design has the advantage of focusing attention on the longitudinal nature of the talent development process because it views the student's outcome performance not in isolation but rather in relation to entering input performance. Its basic weakness is that it really produces no information that bears directly on the question of environmental impact. Would the same changes have occurred if the student had been exposed to a different kind of program or to no program at all?

These inferential problems are probably not as severe at the level of the individual course or class, given that course examinations are usually more specialized and focused on a relatively short time interval. It is probably reasonable for a professor to assume that changes or improvements in student performance that occur during a quarter or semester are largely attributable to the course experience. Clearly, the availability of input information can be of significant value to those of us who teach in higher education for at least two reasons. First, it tells us about the students' strengths and weaknesses early enough to give us an opportunity to adjust our teaching during the course. Second, it provides us with a baseline for assessing how much students actually learn and how much their performance improves between the beginning and the end of the course. Even so, if the results of such pretest-posttest assessments lead us to conclude that the students are not learning as well as we would like them to learn, there is really no way for us to know for sure what needs to be changed to bring about the desired degree of improvement. For this reason it would be useful for all of us who teach in higher education to begin to experiment with different approaches to teaching (i.e., with different environments) to learn more about how best to facilitate learning. The experimentation can take several forms. We might give different students different types of assignments. Or, we might teach one section

of a course using one approach and another section using a different approach. In effect, such experiments introduce *environmental variation* into our input-outcome model.

Because of the cost and time associated with collecting longitudinal pretest and posttest data, many investigators have tried to shortcut the process by simultaneously assessing freshmen and upper classmen on some measure. In addition to the problems already mentioned, this shortcut method is so full of pitfalls that one wonders if there is the slightest justification for supposing that the observed "changes" are related in any way to the college experience. For example, such an approach forces us to assume that upperclassmen are a representative sample of the total cohort of freshmen from which they were drawn. We also need to assume that this original cohort was drawn from the same population as the current freshmen who are being compared to the upperclassmen. In other words, this shortcut approach assumes that the successive entering freshmen classes have not changed with respect to the outcome measure, and that the dropouts, persisters, and transfer students are all comparable on the outcome measure. Except under unusual circumstances, neither of these assumptions can be justified.

In short, the input-outcome model produces inferential difficulties that result from the need to assume that *change* is equivalent to *environmental impact*. This problem suggests that it would be useful to regard changes in students that occur during the course of an educational program as comprising two components: change resulting from the impact of the educational environment and change resulting from other influences (maturation, effects of other unmeasured environmental variables, and so on). Note that the program being assessed may (1) bring about changes that otherwise would not occur, (2) exaggerate or accelerate changes resulting from other sources, or (3) impede or counteract changes resulting from other sources. In other words, it is even conceivable that the true effect of the environment being assessed is the *opposite of* the observed change that occurs between pretest and posttest and that the change would actually have been greater if the student had been exposed to a different environment.

Environment-Only Assessment

When some people speak of evaluation, what they have in mind is environment-only assessments. In this type of assessment we focus our attention on the educational program itself: teaching techniques, curriculum content, course materials, course assignments, physical facilities, the qualifications of professors. When faculty members evaluate each others' courses by

examining course syllabi, they are practicing environment-only assessment. Perhaps the best-known application of this method is the regional accreditation process in higher education. Traditionally, accreditation has involved an examination of the institution's libraries, physical plant, faculty-student ratios, teaching loads, required and elective courses, and the academic qualifications of the faculty such as the percentage with doctoral degrees. In recent years regional accrediting associations have begun to request information on "outcomes," but typically this information is collected in isolation from other data about the institution. In effect, this merely adds an outcome-only component to the usual environment-only component of regional accrediting. A notable exception is the Southern Association of Colleges and Schools, which requests all of its institutions to produce "evidence of improvement" or what amounts to talent development data on student learning. Virginia has taken the mandate further, requiring all public institutions to engage in value-added outcomes assessment in six core competencies: written communication, quantitative reasoning, scientific reasoning, critical thinking, oral communication, and information technology.

Another well-known example of environment-only assessment is the periodic ratings of the "quality" of graduate programs (Cartter, 1966; Jones, Lindzey, and Coggeshall, 1982; Rose and Anderson, 1970). Although these are, in effect, reputational surveys, they appear to be primarily a reflection of the scholarly productivity and reputation of the faculty in the particular graduate department being rated (Drew and Karpf, 1981).

The problem with environment-only assessment is that it runs afoul of the same difficulties noted in our previous critique (see chapter 1) of the reputational and resources approaches to excellence (i.e., no information bearing directly on learning or the talent-development process is gathered). In other words, no matter how detailed the descriptive information made available through this method, no data concerning the actual impact or effectiveness of the educational program is provided. In the absence of such information, we are forced to *infer* it to make any evaluative judgments about the program. For example, if a particular course syllabus is regarded as deficient in some respect, it is necessary to assume that the alleged deficiency produces some unwanted result in terms of the desired educational outcomes of the course (student learning). It is also necessary to assume that the recommended remedy in the syllabus will produce a better outcome. Similarly, if a visiting accrediting team decides that the institution's library is deficient in some respect and recommends that it be changed, the team is implicitly assuming that the alleged deficiency causes some decrement in student (or faculty) talent

development, which would once again be remedied by implementing the recommended change.

Because the environment-only method is particularly popular in evaluating curricula, one should recognize that such evaluations necessarily assume that "what is taught is what is learned." There is, however, one situation in which this method can be applied with a reasonable degree of confidence. Assume, for example, that prior longitudinal research has shown that a particular kind of educational intervention, curriculum, or program produces better results (in terms of improvements from input to outcome) than other approaches. Armed with such information, the assessor can then examine the content and method of the program being evaluated to determine whether it possesses the most desirable components (as determined by the previous research). Recommendations for change under these conditions would not be based on speculation but on previously established empirical findings. Once well-designed longitudinal research has established the causal connections between environmental characteristics and particular educational outcomes, such information can provide the basis for environment-only assessments that can be carried out much more rapidly and at much lower cost than elaborate longitudinal studies. This approach was used in the design and development of the National Survey of Student Engagement (NSSE) (Kuh, 2001),[5] which includes a number of items reflecting environmental experiences that have been shown in previous I-E-O studies to be associated with favorable student outcomes.

SUMMARY

This chapter has presented a conceptual model to be used as a general guide in designing and implementing assessment activities on any campus. The I-E-O model is predicated on the assumption that the principal means by which assessment can be used to improve educational practice is by enlightening the educator about the comparative effectiveness of different educational policies and practices. The I-E-O model is specifically designed to produce information on how outcomes are affected by different educational policies and practices. Use of this model should allow those responsible for assessment activities to enhance their understanding of how student or faculty development is affected by various educational policies and practices.

The three informational components of the model—inputs, environments, and outcomes—are discussed in more detail in the next three chapters.