Evaluating Honors Programs: An Outcomes Approach

NCHC Monographs in Honors Education

Jacqueline Reihman
Associate Professor of Psychology
SUNY College at Oswego

Sara Varhus
Associate Dean of Arts and Sciences
SUNY College at Oswego

William R. Whipple
Director of the Honors Program
University of Maine

A Publication of the National Collegiate Honors Council

Table of Contents

INTRODUCTION

The evaluation of academic programs has always been a complex and sensitive issue. Evaluations are undertaken to determine which programs will survive in an era of straitened economic circumstances, to gain or maintain accreditation, or to tell us how our programs can be improved. They may apply some normative standard of quality, or address an academic program's unique situation and mission. They may include the following: review of budget, evaluation of staff, description of the program's operation, demonstration of faculty and student satisfaction, or measures of what students have learned. They may use standardized or locally developed tests of achievement; tests of ethical and cognitive development; analysis of demographic data; transcript analysis; course evaluations; exit examinations; or faculty, student, and alumni surveys.

In the seven years since NCHC published *Handbook for the Evaluation of an Honors Program*, demonstrating the effectiveness of academic programs has become a widespread obligation, and educators have approached this task with increased sophistication. The most commonly cited causes of this trend are recent indictments of higher education in widely publicized national reports, and the determination of governments and governing boards (and provosts and deans) to ensure that support for education is money well spent (Warren 3). Whatever the cause, evaluation has become a central academic responsibility.

It is probably useful to clarify some of the terminology of evaluation. "Evaluation" and "assessment" are general terms for the examination of the quality or effectiveness of an academic program, although in some contexts "assessment" refers more strictly to the measurement of what students learn or how they change, while "evaluation" is taken to include a broader look at how well a program functions (its "process"). "Program review" often refers to the process of deciding a program's status (whether to retain or close it). "Formative" evaluation is done for the purpose of showing the effectiveness of various dimensions of a program for the purpose of internal review and improvement; "summative" evaluation is intended to provide the basis for a global decision about a program or course. "Outcomes measurements," in the strictest sense, refer to the examination of student

achievements after completion of a program—keeping track of students who go on to graduate school is an example of this—while "value-added" assessment refers to the attempt to show the way in which a course of study changes, educates, or influences students—usually this involves tracking a group of students during the time they are involved in a program. Sometimes, "student outcomes" refers to both of these kinds of data.

Furthermore, it is clear that among the many ways in which academic programs can be evaluated, showing that a program has clear educational goals and objectives and that its students achieve those goals and objectives is the most highly regarded and persuasive approach. In its published standards for accreditation, the Middle States Association of Colleges and Schools cites a "persistent concern as to the relation between goals and outcomes" as a characteristic of excellence in institutions of all types, and specifies the assessment of student outcomes as the key to program review: "They do not depend wholly on any indices, but they search and weigh them all for evidence of progress or success or of results expected but which did not occur" (*Characteristics* 11). In this climate, it is of paramount importance to develop and adopt procedures—surveys of alumni achievements, measures of intellectual and academic development, or standardized tests of achievement—to determine and describe learning that has occurred in our programs. Other questions—whether to use an outside evaluator and whether to invoke normative standards—are important but secondary.

In light of this increased concern with evaluation, and in particular the evaluation of student outcomes, the paucity of evaluations of honors programs is surprising. In *The Best for the Best: A Composite Profile of Honors Programs in American Colleges and Universities* (1985), Catherine Randall and S. Nicole Collier observe that "examples of efforts to evaluate the effect of honors programs on the college career and/or life after college of honors students, are extremely rare. And most of those that exist are fundamentally anecdotal" (n. pag.). The "Evaluation Handbook Questionnaire" sent to all NCHC member programs in preparation for this book corroborates this. Half of the respondents to this survey indicated that their programs had undergone an evaluation in the past five years, but that student satisfaction was the feature most frequently examined. (10% of these programs use tests of achievement in specific areas or require students completing the program to pass an exit examination.). Although the

majority of the programs do require faculty to use course evaluations, more than half of those have no systematic procedures for using that information. We believe that because honors programs offer what are by definition special educational opportunities to superior students, it is important to examine what our programs are achieving.

Stated more positively, because honors programs usually offer special, innovative academic programs within larger institutions, they represent an excellent opportunity to compare the results of differing curricula and to explore ways to realize in our students our most cherished educational aspirations. Most honors programs have lofty goals. Randall and Collier report some representative objectives: "to stretch, strengthen, and stimulate superior students"; or "to motivate. . . challenge. . . enrich" exceptional students. They summarize their findings in this way: "Honors programs are seen as a mechanism whereby gifted students may expand the *breadth* and *depth* of their learning by, in many cases, the provision of greater *flexibility* in their curricula, and in all cases by the provision of a *different* educational experience." But the nagging question for most honors programs is whether we are succeeding in producing "stretched, strengthened, and stimulated" students—or, for that matter, whether we know what a "stretched, strengthened and stimulated student" is? If we can examine our degree of success in achieving such goals, we will have information critically useful in the modification of our own and other honors programs, as well as powerful arguments for continued institutional support. (See also Catherine Cater 17.)

In addition to this general need for honors programs to examine their unique services to their students, other characteristics which form an important context for the evaluation of honors programs were revealed in the "Evaluation Handbook Questionnaire." While these concerns are not necessarily universal, they will influence the questions posed in many honors evaluations:

1. Causes of Attrition. Although respondents were evenly divided in citing as causes of attrition in their programs the difficulty of requirements, transfers to other colleges, and the failure of students to maintain good academic standing, a problem in students' commitment to or attitude

toward their education was frequently specified. In other words, the quality and nature of students' academic involvement is a rather common issue in honors programs. It also poses a challenge to the evaluator: student commitment or involvement is not synonymous with student satisfaction, and traditional academic measurements do not address this general quality or attitude.

2. Honors Curricula. The overwhelming proportion of respondents (73%) described their curricula as liberal studies, with 17% indicating an interdisciplinary focus. We can conclude that, because most honors programs are not responsible for in-depth mastery of specific academic disciplines, assessment of student achievement in honors programs must usually address the broad ethical and intellectual goals of general education.

3. Cultural and Community Activities. Most (78%) of the programs responding encourage or require students to participate in extracurricular cultural activities, and nearly half of the programs encourage/require students to participate in community service projects.

4. Administrative Structure and Budget. When it comes to designing and executing an evaluation, most honors programs do not have lavish resources to draw on. One fourth of those who responded to our survey indicated that they have no administrative positions, and an additional half of the programs have only one administrative line; our data also corroborate the findings of Randall and Collier that honors budgets are usually modest (Randall and Collier n. pag.).

5. Advisement. Honors advisement is usually at least in part the responsibility of honors personnel, and academic advisement (as opposed to career or personal advisement) is cited as the most important advisement function. What is the best indicator of effective advisement?

Especially in light of the fact that, in the "Evaluation Handbook Questionnaire," most of those who indicated that they had undergone a program evaluation in the past five years did so for the purpose of self-assessment (only 2% needed to report directly to accrediting agencies), it seems crucial to present an overview of methods for determining the effect of academic programs on our students. The first chapter surveys evaluation strategies

and defines some terms and concepts in program evaluation. The second through the fifth chapters reflect the typical sequence of evaluation activities. The second chapter identifies the first issues to consider—the context in which evaluation takes place, the identification of programmatic goals for evaluation, and the possible foci of an evaluation. In the third chapter, we present an overview of evaluation design, and in the fourth, strategies for gathering data and sampling. The fifth chapter discusses analysis and interpretation of the data collected. And finally, the sixth chapter and the appendix offer additional information which we think will be useful to honors personnel: the sixth chapter takes up questions dealing with the evaluation of special projects, with an emphasis on non-quantitative methods of evaluation, and in the appendix we suggest sources for further reading. This survey of the process of evaluation introduces some of the technical aspects of evaluation. While program evaluation often requires special expertise—in many cases, evaluation cannot be a do-it-yourself project—we believe that honors personnel can make decisions about evaluation only if they are acquainted with the many aspects of program evaluation.

Works Cited in Introduction

Cater, Catherine. "A Brief Consideration of Honors Research." *The National Honors Report* 7.4 (1986): 17-18.

Commission on Higher Education: Middle States Association of Colleges and Schools. *Characteristics of Excellence in Higher Education: Standards for Accreditation.* 1982.

Randall, Catherine J. and S. Nicole Collier. "The Best for the Best: a Composite Profile of Honors Programs in American Colleges and Universities." Unpublished.

Warren, Jonathan. "Assessment at the Source: What is Assessment and Why are We So Obsessed with It?" *Liberal Education* 73 (1987): 2-6.

CHAPTER ONE: The Nature of Evaluation

Two assumptions underly this handbook: First, we assume that honors program evaluation must be tailored to the individual program and institutional context. It must address the explicit goals and objectives of the program in question (or, if those are not explicit, must begin with the articulation of educational goals and objectives), and it must be appropriate to the audience and purpose of the evaluation. This is not to say that honors evaluation will not involve the use of standardized data or normative comparisons—it may be that comparisons to other honors programs or academic programs in general are appropriate for an evaluation. (However, let us say again that there is an almost total absence to date of data about honors student outcomes.) Second, we assume that, whether evaluation is undertaken in response to external pressures like accreditation review, or for the purpose of understanding and improving the program, it is important that honors evaluation use objective methods in assessing outcomes. Therefore, because we expect honors evaluation to be eclectic, this handbook is intended to be a sourcebook of evaluation tools and strategies—including assessment strategies—and their uses, strengths, and weaknesses.

I. Approaches to Evaluation. There are, broadly speaking, three approaches which may be taken in evaluating an honors program. We will call these approaches operational, process, and outcome oriented evaluations, and we note in passing that there is no standard terminology in this area. Some writers use the term "process" to refer to what we call "operational" while others define "outcome" so broadly that any approach can be called outcome evaluation. We have chosen our terms in the belief that they make the distinctions among approaches clear, especially for those who are not professional evaluators.

Operational evaluation has to do with the program's operations and how these fit into the ongoing mission of the larger educational institution. In operational evaluation one examines the structure of a program, the nature of its client population, and its use of resources; it can involve such things as recording the numbers of students and faculty involved with the program, monitoring costs associated with the program's operations, and attending to

curricular issues such as the number and nature of courses offered, student enrollment, and so on. Operational evaluation frequently includes inviting a consultant who is experienced in honors work to visit the campus and make comparisons with similar programs at other comparable schools. What is gained through this arrangement is a comparative picture of the honors program within the context of other programs whose missions and problems resemble those of the program under study. Comparative operational analysis may also result in cross-fertilization of ideas among institutions. Operational evaluation produces information which is particularly valuable to college or university administrators, and it is often the basis for budget requests. It is not the primary focus of this book; however, the *Handbook for the Evaluation of an Honors Program*, by C. Grey Austin et. al., includes discussions of operational evaluation. Process evaluation concerns itself with a program's ongoing activities and the ways in which these interact with the concerns of the populations which it affects — often called stakeholder populations. In the case of an honors program, the stakeholders might include students, faculty, the university administration and trustees, funding organizations, alumni, and the community in general.

Process evaluation usually requires an external evaluator, who conducts extensive interviews with representatives of the various stakeholder populations and refines the perceptions of these groups into a report on the program's process. The effectiveness of such an evaluation depends heavily on the skill of the evaluator in framing questions and consolidating the results. Process evaluation provides valuable insight into the ways in which the program is perceived by those whom it affects, and it is especially useful in evaluating special projects (see Chapter Six).

One particularly important type of process evaluation is known as naturalistic evaluation. Naturalistic evaluation is distinguished from the positivistic approach emphasized in this book in two important ways. In the first place, while positivistic evaluation begins with a statement of the goals of the program and proceeds to assess how effectively those goals are being met, the naturalistic approach considers a program's objectives as an emergent feature of its process; rather than asking whether specific aims are being accomplished, naturalistic evaluation studies the program's processes and seeks to extract from these the actual impact of the program. Second, the naturalistic approach regards the evaluative process to be a part

of the ongoing process of the program; there is no sharp distinction between program activities and evaluation activities. From a naturalistic perspective, evaluation is a continuing aspect of a program, whereas the positivistic approach tends to regard evaluation as an activity outside the boundaries of the program itself.

Naturalistic evaluation has been applied to educational settings with good effect; however, a detailed consideration of this approach lies outside the province of this book. Readers interested in learning more about the naturalistic school of evaluation should consult the work of E.G. Guba and Y. Lincoln (see Appendix).

Outcome-oriented evaluation focuses on the consequences of a program. In the case of an educational program, this means that its particular concern will be how the education of its students is affected by participation in the program. Underlying the evaluation is the question: "Has the program effectively met its goals and objectives?" To answer this question, an outcome evaluation applies the methods of quasi-experimental, survey, and interview research. Because we believe that the most useful result of an evaluation is an understanding of the degree to which its intended aims are being met, we have directed this book primarily toward outcome evaluation.

We have emphasized earlier the importance of objective methods. By "objective" we mean that the evaluation is undertaken without preconceptions about the results and that the methods used will yield findings which would be convincing to a disinterested outsider. It is important to emphasize that "objective" is not the same thing as "quantitative." Objective methods need not necessarily be quantitative, and numerical methods are not always objective.

II. The Tools of Evaluation. Program evaluation is a branch of the social sciences, and like most technical fields it has developed a specialized language to refer to basic concepts. For the most part these concepts are easy to understand, but the terminology may disorient those who are unfamiliar with this type of research. In this handbook we try to avoid unnecessary jargon, but some terminology must be introduced if honors directors and faculty are to take an intelligent part in an evaluation (as,

indeed, they should.) In order to evaluate a program it is necessary to study individuals involved with that program; in the case of honors programs these will be students, faculty, alumni, and occasionally other groups such as administrators or members of the community. In evaluation research these individuals are referred to as "subjects" — a term borrowed from behavioral psychology and one which carries the inappropriate connotation of passivity on the part of the "subject." Despite this connotation, the term is in universal use, and an alternative is not easy to find. Therefore, we reluctantly use "subject" when we need a word to describe those individuals who are being studied as part of an evaluation. "Variables" are, in social science research, dimensions along which subjects differ from each other. In the more experimental branches of the social sciences, variables are divided into "independent variables" (factors, deliberate or accidental, which are thought to influence a subject's thoughts or actions) and "dependent variables" (changes in the subject's ideas or behavior as a result, presumably, of independent variables). For our purposes, however, the unqualified term "variables" will usually refer to what would be termed independent variables in experimental research; the equivalent of dependent variables will be termed "outcomes."

Consider the following highly simplified evaluation. We are examining an honors program, and we want to know whether honors students receive better grades than those received by non-honors students. We are also interested in whether honors students are more likely to apply to graduate programs. Since we know that both grades and tendency to apply for graduate study are influenced by academic aptitude, we wish also to take into account the aptitude of both our honors and non-honors students, and we decide to use SAT scores (combined verbal and math) as a measure of aptitude. In this example we have defined two variables and two outcomes. The variables are (1) whether or not a particular subject is an honors student; and (2) the subject's combined SAT score. The outcomes are (1) the subject's grade point average (GPA); and (2) whether or not the subject applies to graduate school.

We will return to this example later, but for the moment notice that the two variables defined above differ in one important respect. The subjects' SAT scores are a quantified variable; each subject will have a score in the range from 400 (the lowest possible combined SAT) to 1600 (the highest

possible). The presumption is that subjects with higher scores have more aptitude than those with lower scores, and subjects may have any score within the range of possible scores. It is quite possible, if we are dealing with a small number of subjects, that no two of them will have the same score. The other variable concerns participation in the honors program. There are no degrees of this; either a particular subject is in honors or she or he isn't. There are no quantities to be measured in this case. This is what is known as a categorical variable. The same distinction can be made with regard to the two outcomes: GPA's are quantitative measures, and whether or not a subject applies to graduate school is a categorical measure. Techniques of data analysis differ for the two kinds of data (see Chapter Five on Analyzing and Interpreting Results), but both kinds are important, and a good evaluation will usually include both quantifiable and categorical data.

The process by which we select and define variables and outcome measures to reflect what we want to know is called operationalization; it is a critically important aspect of evaluation, and one which honors staff must understand and take part in. Even when an external evaluator is brought in to perform the evaluation, she or he will require intelligent assistance from the honors director and/or faculty in operationalizing the measures and variables which will be involved in the evaluation. In the simple case described above it was assumed that SAT scores could be used as an indicator of academic aptitude. Is this appropriate? There is no simple answer: educators disagree on this point. Similarly, it was decided to use as one outcome measure applications to graduate schools. Is this the best measure to employ? Perhaps we should look at acceptance into graduate programs rather than application. Perhaps we want more than two categories: (1) students who do not apply for graduate study; (2) students who apply but do not enroll; (3) students who enroll but do not complete their graduate program; (4) students who receive a graduate degree. (Note that categorical data may involve more than two categories; they are still categorical, rather than quantitative, data.) Decisions of this nature should be made with the knowledgeable assistance of honors staff: no one else is in a position to decide whether a particular operationalization is appropriate to the program's goals and structure. Good operationalization of both quantified and non-quantified data is vital for effective evaluation; while it does not, in

itself, guarantee a successful result, poor operationalization invariably leads to results which are difficult to interpret or misleading.

Once the variables and outcome measures to be used have been selected and operationalized, it is necessary to plan how they will be used to answer the questions which we need to have answered. This process is known as design and it is discussed in detail in Chapter Three. Consider the example described above. We stated that we wanted to compare grades and graduate school applications for honors and non-honors students, taking into account academic aptitude (operationalized as combined SAT scores). How, exactly, will we take account of SAT scores? We might decide to compare the honors students with a control group of non-honors students whose averaged SAT score was identical to that of the honors students. Or we might use a technique known as matching, where for each honors student being studied we find a non-honors student with the same SAT score. (In the latter case we will have to decide whether we will consider two students matched when one has a verbal SAT of 450 and a math SAT of 700 and the other has a verbal score of 720 and a math score of 430—another problem of operationalization) Still another way to take SAT score into account would be to use statistical techniques such as analysis of variance which allow us to "factor out" the effect of a variable such as SAT scores. Such techniques are very powerful, especially when outcome measures are quantifiable; but their usefulness is limited when one is dealing with categorical outcome measures.

Decisions about design, like those regarding operationalization, require the collaborative efforts of the evaluation team and the honors program's faculty and staff. Even if an external evaluator is not involved, honors directors usually require assistance from a consultant trained in research methods and statistical analysis at this stage. (Such assistance is available on many campuses in departments of mathematics, psychology, sociology, or economics.) However, it must be emphasized that, as with operationalization, honors staff must take an active and intelligent role in the decisions made in the process of design. The validity of an evaluation is compromised by a design which specifies inappropriate controls or dubious statistical procedures, but it is similarly compromised if the design fails to address the questions of concern to those responsible for the program. The best designs result when honors directors work actively and knowledgeably with specialists in research methods. Only when both methodological and programmatic concerns are carefully addressed can a design be fully effective.

III. Validity and Quantifiability. It was noted above that good evaluations typically include both quantifiable and non-quantifiable variables and outcome measures. It is worth spending a moment to consider the two types of data. Categorical data divide subjects into two or more categories: honors/non-honors, male/female, first year/second year/third year/fourth year, and so on. Quantifiable data reflect varying degrees of whatever attribute is measured; in addition to SAT's and GPA's, quantifiable data would include outcomes such as number of papers written or starting salary after graduation, and variables such as number of cultural activities attended, age, or number of siblings. (These last two items are examples of what are known as demographic variables, meaning that they refer to a subject's background rather than to a way in which the program influences the subject.) Some quantifiable measures have been studied systematically within large populations so that from an individual's score one can make comparisons between the individual and a population of his/her peers; these are known as standardized measures. Most standardized measures are quantified (like the SAT or the ACT Comp), although there exist some standardized categorical measures (such as the Myers-Briggs Type Inventory or the Kolb Learning Style Inventory).

A crucial concern regarding any measure to be used is the question of validity. Validity means that the instrument chosen actually measures what we want to have measured. An accurate clock is a valid instrument for measuring time; one which runs slow or fast is invalid. Validity is so important to evaluation that we will refer to it in most of the chapters which follow; for the present we will discuss the two principal ways in which the validity of a measure may be established. Empirical validation arises from comparisons between the instrument in question and other instruments. In the case of a clock, we may establish its validity by comparing it to a standard, such as the time-keeping devices maintained by the National Bureau of Standards. In this case, the validity rests upon the credibility of the standard, which is why official organizations exist to maintain dependable standards for such basic dimensions as time, length, and weight. If comparison with an official standard is inconvenient or impossible we may instead compare our instrument with others which measure the same

quantity or a related one: thus we might compare our clock to a number of other clocks, or to the movements of the stars (one of the earliest known validating standards). If, when used in combination with appropriate tables, our clock accurately predicts high and low tides, the clock has empirical validity (in this case, a particular form of empirical validity known as predictive validity). Many standardized variables and measures used in evaluation have been validated in this way. Empirical validation requires that the instrument in question be compared to some other measure, in either a predictive or a correlative way. What if no other instrument exists, or is conveniently accessible? We then turn to what is known as face validity. This means that the instrument we are using has an obvious natural validity which would be apparent to most people. To measure the amount a student has learned by using grades is an example; we take it more or less for granted that students with higher grades have learned more—we do not need to perform correlational or predictive studies in most cases. Note that face validity is more subjective than validity by comparison with a standard. To examine another measure commonly used in universities, the quality of faculty scholarship is frequently assessed by the number of publications authored by the faculty member. To many academics (including the majority of deans and department chairs) this measure of scholarly productivity has obvious face validity. To others (especially faculty coming up for tenure) its face validity is less obvious. The acrimony with which the "publish or perish" philosophy is debated illustrates the problems which arise when people disagree about face validity. But face validity does not always lead to controversy. Frequently the outcomes of a program can be assessed easily and effectively by using measures which have clear face validity. If, for example, one goal of an honors program is to encourage students to exhibit good citizenship, a good measure might be whether the students are registered to vote. Since voting is widely accepted as one of the responsibilities of a good citizen, this measure has face validity as part of an assessment of citizenship. (It would be more controversial if we used this measure as an exclusive test of good citizenship; voter registration has a more obvious validity as a component of good citizenship than as the sole defining feature.) Many highly useful measures rely on face validity in this way. In particular, most locally-developed instruments need to exhibit face validity, since it would usually be prohibitively costly and time-consuming to establish empirical validity for them.

Each type of data which can be used in an evaluation has specific advantages. Quantifiable data permit powerful statistical analyses, but they are useful only where an underlying dimension can be identified and a means of identifying differing degrees can be operationalized; otherwise, qualitative measures are more appropriate. Standardized measures typically have been empirically validated and permit comparisons between the local sample and more global populations, but they will provide information only on the scales for which they have been validated, and this may not be the information needed. One of the most common errors made in program evaluation is selection of instruments simply because they are quantifiable and standardized. The first criterion in selection of instruments is whether or not they address the questions which need to be answered. If there are quantifiable and/or standardized tests which provide the necessary information, it will be advantageous to use them; but they are of no use unless they answer a relevant question. A crude, home-grown instrument which measures what we want to know is preferable to a sophisticated standardized test which measures something else.

Suggestions for Further Reading

Judd, Charles M. "Combining Process and Outcome Evaluation." In *Multiple Methods in Program Evaluation.* Ed. Melvin M. Mark and R. Lance Shotland. New Directions for Program Evaluation 35. 1987.

CHAPTER TWO: Focusing an Evaluation

The purpose of this chapter is to pose questions which will clarify the context in which an evaluation takes place and define the focus of the evaluation. The initial stage in program evaluation is to consider and answer these questions. First are the questions relating to the purpose of the evaluation and the process through which it is conducted. Second, we address the matter of identifying the goals and objectives of a program. And, third, we suggest ways to define, limit, and structure the evaluation. All of these matters should be addressed in the preliminary stages of the evaluation.

Of course, the makeup of the group consulted in the planning of an evaluation will vary from program to program. Grey Austin suggests that evaluation be "an open process, one that is as systematic as possible, but in which there is time for the unrestricted conversation that can generate new ideas and in which the assembled data are reviewed not simply to answer prescriptive questions but to discover unanticipated areas of significance" (Austin 6). It is additionally important that the "constituencies" of an evaluation be consulted. For example, it may be important that the overall focus of the evaluation be approved by the person or group requesting the evaluation. Or, if the goals and objectives of a program are not explicit and need to be defined, the director should invite the honors committee, faculty, and students to participate in outlining the purposes of the program. Evaluation is a sensitive matter. Specific evaluative instruments as well as general questions about evaluation can be controversial, and it is important to confront these matters early and in an open manner.

I. The Context for Evaluation. We hope that the following questions about organizing an evaluation will stimulate thinking among those involved with the evaluative task so that it may be approached with a clarity of purpose:

A. Who wants the evaluation and why has the evaluation been requested? A variety of persons or groups may request an evaluation of your honors program: accrediting agencies, college or university administrators, campus governance bodies, honors boards, and others. And there may be

a variety of explicit and implicit motivations for evaluating: determining the impact of a set of courses, granting the program permanent status within the institution, or improving the program's staffing and funding level. Evaluation is conducted not in a vacuum, but rather within a political context. Understanding the motivations of those requesting the evaluation is important. For example, an evaluation that rates students' satisfaction with honors courses and advisement might not offer compelling support for a program's continued survival in an institution where students in general are satisfied with their courses and advisement. Or, a quasi-experimental administration of a test of ethical development will not satisfy an administrator concerned with honors students' success after graduation.

B. Who "owns" the data? Often, an evaluation uncovers some unfavorable, or at least unanticipated consequences. It is important to know before an evaluation begins how the findings will be used and treated. What has been requested of you? An evaluation, or a report on the program in question? Although it is probably most persuasive to make an evaluation, with its overall plan and results as described by an evaluator, available to all concerned people, it is possible to make a distinction between the evaluation results and a report, which might make selective use of data generated in the evaluation. Furthermore, consideration must be given to publication rights. We hope that some evaluation studies will be interesting enough from the perspective of either methodology or programmatic concerns that publication or presentation of findings is warranted. To avoid potential conflicts, these issues should be discussed, particularly if an "outside" evaluator is charged with the task of conducting the evaluation.

C. Is there a commitment to using the results of the evaluation? Once again, this is a politically sensitive question. Ideally, evaluation involves a continuing, cyclic process in which new information is integrated into program modifications and then the evaluation process begins again. One way to increase the likelihood that evaluation results will be used is to involve persons with the power to execute change in planning and design sessions. A related concern here is the quality and integrity of the evaluation itself. Because the evaluation process should result in better academic programs, it is also important that the evaluation be well done: an evaluation study which is poorly planned or executed, or where conclusions are ambiguous or subject to multiple interpretations, is likely to result in

significant problems later on. Certainly, future evaluation efforts will be thwarted, sabotaged, or resisted altogether.

D. Who is responsible for conducting the evaluation? Is the evaluation going to be conducted by someone within the honors program itself (perhaps even the director), or is it going to be done by some neutral, third party external to the program? (We define the evaluator as the person who oversees the design, execution, and summarizing of the results of the study.) There are advantages and disadvantages in each of these choices.

The most important advantage of an "internal" evaluator is knowledge of the program—its goals and objectives, its day-to-day operations, and its students and faculty. It is also usually true that the internal evaluator will be already trusted by program personnel. Thus, valuable time need not be wasted.

There are, however, disadvantages associated with the internal evaluator. First, it may be the case that there is no one on the honors faculty skilled in the area of measurement. Second, there is a problem of validity when the evaluator is involved in the program being evaluated. Since the person asking the questions is internal to the program, students or faculty may be reluctant to respond honestly to a faculty member or director. This would be particularly true in longitudinal designs where student tracking (i.e., requiring names) is essential.

Third-party evaluators, on the other hand, may have difficulty securing valid information for precisely the opposite reasons. Because they are "outsiders," it is possible that students may feel a need to defend "their" program and consequently be less than candid. In addition, while third-party evaluators might do a fine job of conducting the evaluation in the short term, they may not have the long-term commitment to insure that findings are ultimately used. And a third-party evaluator is likely to cost more than an internal evaluator.

Some things, though, are likely to run more smoothly with an "outside" evaluator. If well chosen, he or she is likely to be skilled in the area of design and methodology, and will be familiar with and have access to measurement tools. And, ultimately, the study may be more rigorous. A useful source of referrals is the American Evaluation Association, which is a professional

organization to which many well-qualified evaluation specialists belong. (You can request a membership list from the American Evaluation Association, 9555 Persimmon Tree Road, Potomac, Maryland, 20854.) Perhaps the wisest option is to combine the two strategies. Close co-operation with a third-party evaluator whom participants perceive as sanctioned by honors personnel will probably result in an evaluation study that is both appropriate to the aims of the program and rigorous in design and methodology.

E. What resources are available for conducting the evaluation? Rigorous evaluation requires resources—money to compensate an outside evaluator, if one is used, and time to gather data from records and to administer the evaluation. Obviously, extensive studies cost more than limited ones, but it is important in any case to recognize in the beginning what will be needed in the way of computer resources and clerical support. Since the majority of honors programs have small budgets, it is important to note that program evaluation can be supported by the resources of an academic institution. A skilled evaluator might be found in your psychology or education department and compensated by released time. Through work study, research assistantships, or even independent study arrangements, students (perhaps not honors students) can help with the collection and management of data. And many institutions have research offices which can assist in evaluation.

F. What is the "climate" of the college or university? Timing can be crucial to the success of an evaluation effort. While you may have no control over when you evaluate, it is best to conduct evaluation within an environment which is not emotionally charged. For example, it is not unusual for new honors programs to experience resentment on the part of some faculty who regard honors as an "elitist" undertaking, while older, more established programs enjoy a wider acceptance.

II. Identifying Goals and Purposes. Whatever the reasons for evaluating an honors program, it is important that such a study be shaped by the purposes of the program itself. The preliminary stage in evaluation should include their identification and/or formulation. Often, the philosophy, goals, and objectives of a program are enumerated in some official document, and, often, the faculty and administrators of an honors program will have articulated additional objectives which should be considered as well.

(You might also find it useful to consult other sources like institutional mission statements, college-wide curricula, etc.) In effect, the evaluator and honors personnel engage in a dialogue which is essential to the evaluation and also serves to define and interpret the direction of the program. Program objectives fall into four categories:

A. Student objectives. What do you expect a student who has completed your program to possess or be able to do? Especially in a time when the role of education in the modern world is a vexed issue, answering this question is not an easy matter. Certainly, in some instances, the purpose of the program may be to enable students to achieve a specific goal—enter graduate school, or enter a special program at another institution. Or the goals might be described in terms of completing a specific and easily certifiable field of study—completing a course of study in foreign languages, or demonstrating proficiency in mathematics or computer science. But it is more likely that there will be objectives like "appreciation for the arts," or an "understanding of the interrelatedness of knowledge" that elude "certification"; or enumerations of intellectual skills like "critical thinking" which are not identified with any one academic discipline. And it is also not uncommon to find our hopes for our students expressed in models—the "civic ideal," the "specialist," or the "scribe," for example (Mayville 18-30)—models which need to be characterized more specifically for the purposes of evaluation.

B. Institutional objectives. It will be no surprise to many harried honors directors that honors programs often are responsible not only for serving a group of students, but also for benefitting their institutions through this process. Some honors programs are explicitly designed to attract able students to an institution; whether you are expected to do this or not, it is useful to be able to demonstrate that the honors program does do this. In the same vein, honors programs have been expected to promote institutional development, attract faculty, or generally improve the image of the institution. Within the institution, honors programs are often charged with creating an intellectual atmosphere that will benefit all faculty and students. In the case of both student and institutional goals, it is obvious that while there is immediate evidence of success where some of these goals are concerned—numbers of students entering graduate school or the level of outside funding for the honors program—others are more difficult to demonstrate.

C. Program responsibilities. Although the ultimate goals of honors programs, or any academic program for that matter, must be to educate students, sometimes the means to these goals—functions of the honors program like curriculum and advisement—are enumerated as programmatic goals. And, as before, these goals range from specific to sweeping. The program may be responsible for designing an interdisciplinary curriculum or intensive courses in the basic skills. While there may be questions about what an interdisciplinary course is or whether an intensive composition course is a good one, the goal has at least been addressed if the course exists. Other goals like creating an environment that "will encourage the aspirations and achievements" of honors students or providing "enriched" advisement need clarification.

D. Formative evaluation and goals. If evaluation is undertaken as a step in program revision, it should address those questions which arise in the working of the program, as well as its objectives. Even if the evaluation is primarily summative in nature, the evaluation process offers a good opportunity for obtaining information useful for improving the program. Why do talented students avoid science courses? Do students who complete the honors curriculum have a distinct attitude toward their college work? Does the interdisciplinary seminar make a difference? Such questions can be global—are honors students more sophisticated when it comes to ethical dilemmas?—or narrow—how do honors students compare to the general student at your institution in terms of demographic data?

III. Identifying Outcomes. What follows is a list of approaches to evaluating the success of an honors program. It includes suggestions for identifying student, faculty, programmatic, and institutional outcomes, but is by no means comprehensive. We expect that the unique circumstances of individual programs will generate other approaches, too. Of course, the purpose of the evaluation and the resources available will determine the variety of elements touched upon in an honors evaluation.

A. Evaluating student outcomes.

• **Transcript analysis:** Although a simple technique, comparing the transcripts of honors students to those of non-honors students can demonstrate

that because they are involved in an honors program, honors students take a different range of subjects. Of course, the limitation here is that you have no way of demonstrating the success of the courses.

• **Analysis of other student data:** Useful information and perhaps evidence of success can be found in existing data relating to students: their records as incoming students and their grades, for example. Perhaps honors students get better grades in equivalent courses than equivalent non-honors students. Or you may find that there are non-honors students who did not have the qualifications for honors as incoming students, but now appear to be doing as well in terms of grades—what should the honors program be doing for them?

• **Demographic data:** While not strictly a matter of student outcomes, you might also find it useful to solicit demographic data from them. For example, if you are concerned with strengthening students' commitment to academic excellence, it would be helpful to know that the students who stay in the honors program tend to come from families with differing social and economic status than those who withdraw from the program. It would then be possible to direct retention efforts toward the demographic groups with high attrition.

• **Achievement tests in disciplines:** It might be useful to compare how much honors students have learned in specific subject areas in comparison with non-honors students, particularly if specialization in a discipline is a part of the honors curriculum. Standardized tests like the GREs are available in many areas, although there is a problem in using a test designed to be an entrance examination as a test of achievement. As John Harris warns, these tests are intended to "spread individuals out to maximize individual differences for comparison purposes. . . . The selection-test approach works well when the purpose is to spread individuals over a continuum to select the most able. But it is awkward, to say the least, when the purpose is to certify a level of competence. It is also questionable when the purpose is to assess the impact of instruction on a group of students. Its difficulty lies in its emphasis on ability difference among individuals in the instructional group rather than difference between an instructed group and an uninstructed one" (11-13). Achievement tests can also be developed locally.

• **Assessment of general education.** Because most honors programs are responsible for liberal studies curricula, this is an area of great concern. Probably the best-known standardized test of general education is the ACT-COMP. It has the advantages of standardized tests described above, and, unlike the GREs and the specific ACTs, it is designed to measure how much students have learned. But you should ascertain whether it addresses what you set out to accomplish in your curriculum.

In light of the fact that many honors programs stress the importance of developing students' curiosity, analytical abilities, and an awareness of ethical and civic reponsibility, it is likely that you will need to develop your own measures of the achievements of honors students. Furthermore, if you have the resources and/or personnel to design on your own campus valid tests of the unique qualities you wish your students to gain, the evaluation process will be more effective in clarifying goals and contributing to ongoing improvements. Such tests need not be comprehensive in the usual sense of the word. For example, a rated essay on an issue of public policy may demonstrate that honors students, more than other students, have gained sophistication in analysis and ethical reasoning. Oral examinations or exit interviews could serve a similar purpose. (We discuss the technicalities of designing such instruments in the next chapter.) As Jonathan Warren suggests, "When the reason for . . . assessment is evaluation of an educational program rather than evaluation of individual students, every student need not be tested, and those students need not be tested on everything they have learned" (3).

• **Assessing critical thinking and cognitive development.** Many honors programs offer courses which encourage critical thinking. A useful measure of critical thinking is the Watson-Glaser Critical Thinking Appraisal, Forms A and B (see Woehlke). This is a 40-minute paper-and-pencil measure which addresses inference, recognition of assumptions, deduction, interpretation, and evaluation of arguments. It is considered to be a rigorous measure; however, it is important to note that this test assesses critical thinking only through reading.

Because an implicit goal of many honors programs is the desire to enable talented students to reach their fullest potentiality as persons, and because there is a growing awareness of the psychological and ethical aspect of

learning, some honors programs have administered a variety of developmental scales of ethical, cognitive, and personal development. Scoring of these measures usually requires specialized training and expertise. For determining students' levels of intellectual development, the Measure of Epistemological Reflection (MER), a paper-and-pencil measure of the Perry scheme of cognitive development, is useful. The instrument itself is free, although there is a cost in time and energy in training graders in the MER protocols. You can obtain further information about this test by writing M.B. Baxter Magolda, 350 McGuffey Hall, Miami University, Oxford, Ohio, 45056.

• **Self-report:** In many cases, the results of honors work can be seen in students' behavior, and the best way to gauge this is to ask them about it. For example, students can be asked to indicate which of a list of on-campus cultural events they have attended. Or, questions could be devised to assess students' level of community involvement—voting, charitable work, and so on.

• **Grading and assessment:** Perhaps there is a lesson in the fact that grading, the oldest form of assessment in academia, is often ignored in discussions of ways of measuring what students have gained. Indeed, the fact that the only routine assessment of learning is conducted by teachers themselves is cited by some as the root of the problem of academic accountability (Harris 37). However, others argue that assessment can and should be incorporated into the teaching and grading process. Warren suggests, "A systematic program to bring the assessment procedures of most faculty to a point where they can be used with reasonable confidence to indicate the substance of students' learning, as they are now used to indicate relative accomplishment, would be neither difficult nor expensive. A specific example of ongoing, course-based assessment would be frequent mini-essays—perhaps at the end of each class—in which students ask questions about the day's work. Or a group of honors faculty could work together to devise tests which address important categories of learning, and express the results of those tests not only in letter grades, but in scores which address these categories (6).

• **Alumni surveys:** Alumni surveys can show concrete accomplishments of honors students, as well as students' retrospective satisfaction with a

program at varying intervals after graduation. While it can seem persuasive to show that many honors students go on to graduate school or good jobs, you need to demonstrate that this is because of the influence of the honors program; in other words, even if you can show that, compared to non-honors students, more honors students go on to professional school, you cannot necessarily attribute that to the influence of the honors program. (See the discussion of control samples in Chapter Three.) It is important also to decide whether the goals of your program are best demonstrated by this kind of success or other alumni activities.

B. Evaluating programmatic outcomes.
Honors programs are often responsible for specific services which can be evaluated. In addition to the fairly straightforward matter of describing courses and curricula, advisement structures, and special events and activities, the following can be done to address the nature and quality of the program:

• **Evaluation of advisement:** Evaluating advisement is a rather difficult task because students and faculty may have widely varying expectations as to what advisement should accomplish; this may be an area where the clarifying function of evaluation can be important. The most obvious approach to evaluating advisement is to gauge student satisfaction, being careful to survey an appropriate control group (see Chapter Three). This evaluation could take place after each advisement session (an unwieldy process), or periodically; a survey or interviews could be used. It might also be possible to make inferences about advisement by analyzing student transcripts.

• **Evaluation of curriculum:** Faculty and students can be interviewed or surveyed about the honors courses in which they participate. For example, students can be asked to compare the structure, process, and content of an honors interdisciplinary seminar to a standard disciplinary course, or to several other courses. Or faculty might be asked to describe what they consider to be the unique features of their honors courses. Since the value of this kind of feedback is descriptive, you may wish to ask open-ended questions rather than provide a survey with a restricted range of responses.

• **Evaluation of other institutional outcomes.**

You may be able to demonstrate the impact of your program on its institution by attributing to it increased numbers of highly qualified incoming students or increased attendance at academic functions, or by an enumeration of activities sponsored by the program. On the other hand, you may need to survey faculty, administrators, and students to determine such things as how many faculty and students are aware of the the honors program and what their attitude toward it is; or, how many honors students choose to attend your institution because of the honors program.

Works Cited in Chapter Two and Suggestions for Further Reading

Austin, C. Grey, et. al. *Handbook for the Evaluation of an Honors Program*. National Collegiate Honors Council.

Harris, John. "Assessing Outcomes in Higher Education: Practical Suggestions for Getting Started," Unpublished paper prepared for the American Association of Higher Education under contract to the National Institute of Education in preparation for the National Conference on Assessment, Higher Education at the University of South Carolinam October 13-15, 1985.

Mayville, William V. "Interdisciplinarity: The Mutable Paradigm." AAHE-ERIC/Higher Education Research Report No. 9. Washington, D.C.: American Association for Higher Education, 1978.

Warren, Jonathan. "Assessment at the Source: What is Assessment and Why are we So Obsessed with It?" *Liberal Education* 73 (1987): 2-6.

Woehlke, P.L. "Watson-Glaser Critical Thinking Appraisal." *In Test Critiques*, ed. Keyser and Sweetland. Kansas City: Test Corporation of America, 1987.

CHAPTER THREE: Designing an Evaluation

After considering the context of an evaluation and defining the goals of your program, the next concern is the design of the study. In designing an evaluation, you identify the information (anticipate a logic) that will most clearly demonstrate the effects of honors program activities. Solid design and rigorous methodology are basic to a good evaluation.

I. Validity and Invalidity. Basic to all questions of design is the concern for validity. An evaluation will be valid only to the degree to which you are measuring what you think you are measuring. Invalidity, then, is the systematic but inadvertent measurement of something else. You will need to be concerned with internal as well as external validity. Internal validity is concerned with the degree to which a study accounts for variables; in other words, it allows us to respond to the question, "May we be certain that our evaluation results are due to the activities which we are studying and not something else?" In contrast, external validity has to do with the generalizability of the results of a study: will the findings of an evaluation hold true for future participants or, moreover, participants in other equivalent settings? Campbell and Stanley outline some common "threats" to both internal and external validity which clarify the concept of validity:

• **Contemporary history.** Circumstances which are coincidental with the program may have an influence on what is being measured. For instance, imagine that you are assessing a program goal concerned with the progam's ability to foster enhanced cultural awareness and appreciation. But a new performing arts center has been constructed in your community and there has been a great deal of publicity surrounding the events occurring during the center's first year. If, upon assessment, you find that students have a substantially increased cultural appreciation, you will be pleased but you will not be able to ascribe this enhanced appreciation solely to your program's influence. Certainly, a tenable alternative explanation is that students' exposure to the new performing arts center was the partial or sole reason for their change in cultural appreciation.

• **Instrumentation.** Inconsistency in the scoring and administration of the evaluation instruments leads to invalid results. For example, assume that

you are testing students at two separate times to assess short-term gain in composition skills. The first session occurs on a Monday morning when you and the subjects are ready for peak performance, the second on a Friday afternoon when you are distracted, having just been informed that your galley proofs are due at the publisher's in three days. On Friday, you are careless and hurried in your instructions and urge the students to "hurry and complete the test." If you find no difference in the performance on the two tests, it is possible that the findings reflect inconsistent test administration and not a failure to learn. It is important to anticipate sources of inconsistency in all stages of the evaluation.

• **Testing.** Reusing a test or survey can lead to distorted results: subjects learn from the pre-test and offer altered results at the post-test.

• **Statistical regression.** When subjects are chosen on the basis of extreme scores on some measure—for example, students admitted to honors programs—the results of tests administered to them may reflect that earlier status, rather than the effect which is being measured. This is a particular problem in outcomes assessment of honors students, who are usually highly successful and motivated upon entry to honors programs.

• **Subject mortality.** Validity is threatened when subjects drop out of the study.

Threats to external validity:
• **Reactive effects of testing.** A premeasure may make subjects sensitive to the purpose of either the program or the study.

• **Multiple-treatment interference**. The effects of the program may be confused with the competing effects of other programs in which subjects may be involved. For example, you test your students for an awareness of socio-political issues, and then learn that two-thirds of the students tested have been involved in a Summer Internship Program sponsored by the State Legislature. You might be tempted to attribute your students' achievement in this area to the honors program, when in fact it is likely that the internship contributed to this awareness.

II. Evaluation Designs. Evaluation design enables you to conclude that the results of your study are valid and have not been subverted by a failure to

anticipate possible rival findings. Unless you can adequately defend against plausible alternative hypotheses—unless your design minimizes threats to validity— your results will not be convincing. We will describe several evaluation designs, categorizing them as either **cross-sectional** or **longitudinal**, with further discussion of **sequential designs** and the use of **control** or **comparison groups**.

Cross-sectional designs are those which essentially take a snap-shot of performance at a single point in time; they are, in fact, "one-shot" modes of measurement. The cross-sectional design does not involve assessment of the same subjects over a period of time, but does permit assessment of different groups of subjects at one time. It is especially useful if your evaluation must be completed within a short time. For example, if you wished to assess what your students learn about non-western cultures, you could take samples of freshmen, sophomores, juniors, and seniors and measure their knowledge and proficiency at one point in time. If their level of knowledge and proficiency increases monotonically (that is, increases from the freshman to the senior sample), you would be able to relate their increased competence to class rank.

But it is important to remember that the cross-sectional design speaks only to differences between individuals and although many of the threats to internal validity have been avoided because testing occurs at one time only, significant design problems remain. First, attrition between freshman and senior years results in a distortion of the samples: if poor students drop out early on, the overall performance of the remaining advanced students will be artificially inflated. Also, because you are measuring subjects of different ages at one time of testing, and not individuals at various points, you do not know whether the differences you observe are due to developments associated with age or to the fact that subjects were born in different years (i.e., are members of different "birth cohorts") and hence exposed to different experiences. Thus, you cannot conclude that these students developed in knowledge or proficiency from their freshman to their senior years. Furthermore, you would not be able to lay claim to the fact that your program was responsible for the observed increase in knowledge of non-western cultures because it may well be that all students at your institution (honors and non-honors students alike) commonly experience a similar increase.

The first two problems can be addressed by using a **longitudinal design.** The remaining concern—the ability to ascribe developmental gains to honors program elements—is a concern addressed by the use of **control or comparison groups.** We will discuss longitudinal designs first.

In **longitudinal or time-series designs,** data is collected from the same group of people at several different times. This design is ideal for questions about the effects of educational programs—many honors program objectives take the following form, "Upon graduation from the program, students will have achieved competency in. . . ." For example, an honors program might expect that honors students will experience growth in formal reasoning ability. To assess, perhaps through an exit exam, only seniors' reasoning abilities would not demonstrate that growth had occurred. Rather, to adequately respond to this objective, we would need to have baseline information (i.e., information about that same cohort of students as entering freshmen) regarding levels of reasoning ability and then periodic checkpoints to assess students' progress in that area. Without baseline data, you will have no basis to claim that changes have occurred within individuals; although unlikely, it is possible that honors students enter the program with mature reasoning ability.

While the cross-sectional design can confound age with birth cohort, longitudinal design can confound age with what is referred to as time of testing. This is to say that the changes you observe between, say, year one and year two of your study might reflect true intra-individual changes in formal reasoning or might be due to some particular external influence coincident with your second time of testing—for example, an accidental and widespread familiarity with sample problems presented on the test.

Because of the problems implicit in both cross-sectional and longitudinal designs, researchers have developed more complex sequential design strategies. Although a full discussion of sequential design strategies is beyond the scope of this handbook, we will briefly outline a sequential design and offer it as an ideal design, albeit expensive and cumbersome. A simple sequential design appropriate for honors evaluations would begin in the first year with the basic cross-sectional strategy—groups of freshmen, sophomores, juniors and seniors measured at a single time of testing. You then follow each class group longitudinally through their honors careers.

Thus, you have three simultaneous longitudinal studies (i.e., freshmen-through-seniors, sophomores-through-seniors, and juniors-through-seniors). The comparisons available in this design permit the teasing out of the potentially invalidating effects of age, birth cohort, and time of measurement. This design becomes increasingly complex with the possible inclusion of new cross-sectional studies each year, which are, in turn, followed longitudinally. (For further reading about sequential designs, we refer you to Baltes, Reese, and Nesselroade.)

III. Comparison Groups. The use of carefully chosen **control or comparison groups** is also essential to a well-conceived design. Although longitudinal designs allow us to speak about changes that have taken place in individual students, our conclusions are limited by the fact that these changes may or may not be attributable to participation in an honors program—quite simply, you cannot ascribe results to an honors program solely through the use of a longitudinal design. It is essential to compare honors program participants to appropriately chosen groups of non-honors students—or comparison groups. A comparison group is one whose members have not participated or have not participated fully in the program. A variety of comparison groups—both student and faculty— might be appropriate for the evaluation of an honors program and some possible groups are described below. The choice of comparison group or groups will, of course, be dictated by the questions which you are trying to answer.

A. Student comparison groups:

1. Invited but declined. This is a group of students who were originally invited to participate in the honors program but who, for some reason, decided to decline the offer. Presumably, then, these students are equivalent to program participants with respect to admissions data. Differences which emerge between the two groups may be convincingly presented as a function of the honors program experience, and not academic background or native ability, although differences in levels of motivation should be considered.

2. Matched. These are students who are not affiliated with the honors program but who are matched to honors program participants on presumably important dimensions: age, GPA, gender, majors, SAT or ACT scores. If you find differences between honors students and this matched group,

you can rule out the matching variable as an explanation of the difference. For example, if honors pre-med students are better than their peers in writing about ethically complex issues, then you can suggest that this proficiency in discussing ethical problems does not result from the pre-med curriculum.

3. Drop-outs. These are students who, at one time, were honors program participants but who have left the program can be a rich source of information. When comparing honors students to drop-outs, however, it is important to make separate comparisons with students who voluntarily withdrew and with those who were required to leave for academic reasons.

4. General students. These are students who have never been involved with the honors program.

5. Students in other honors programs. Students participating in honors programs at other institutions could provide valuable comparisons.

6. Student groups within the honors program. It is possible that comparisons among different groups of students within the program might be useful. For instance, if your program allows admission at times other than the freshman year, students admitted later could be compared to students admitted as freshmen.

B. Other comparison groups:
Faculty not associated with the honors program can provide extremely important comparison data. In addition, information about administrators and other nonteaching professionals may be useful in evaluating certain program objectives.

In short, the use of any or all of the comparison groups will vastly strengthen your design. They may be used as an adjunct to the cross-sectional, longitudinal or sequential designs: the use of comparison groups with these designs increases the the validity of the conclusions you may draw from your study. They lend an increased "control" over the environment of the study which will contribute to your ability to demonstrate the influence of the honors program.

Baltes, P.B., H.W. Reese, and J.R. Nesselroade. *Life-span Developmental Psychology: Introduction to Research Methods.* Monterey, California: Brooks-Cole, 1977.

Campbell, Donald T. and J.C. Stanley. *Experimental and Quasi-experimental Designs for Research.* Chicago: Rand McNally, 1966.

CHAPTER FOUR: Collecting Data

After designing the evaluation, you will face two crucial issues in the evaluation process: data collection and sampling. Will you collect information by administering paper-and-pencil tests, by interviewing subjects by telephone, by content analysis of written essays, by observing subjects in real or simulated situations, or by examining program and individual records or documents? Certainly, a variety of options are available to you. Similarly, will you collect information from all the students concerned (a population), or will you collect information from only a sample of that population? The nature of the questions your evaluation addresses, its design, and the resources available will determine the data collection and sampling strategies which you will use.

I. Data-collection strategies. As we enumerate strategies, bear in mind that you need not restrict the evaluation to one technique; indeed, well-conducted evaluations employ multiple methods.

A. Self-report measures: questionnaires, rating or ranking scales, and semantic differentials. Written self-report measures are a relatively inexpensive means of obtaining information about subjects' opinions, attitudes, beliefs, or perceptions. **Questionnaires** can be useful in querying subjects' (either students or faculty) perceptions of the honors program. Questions can be either objective (i.e., multiple-choice, forced choice) or subjective (e.g., What is the best/worst thing about the honors program?). One needs to remember, though, that subjectively designed questions ultimately need to be codified to allow efficient data analysis. That is, prior to analyzing the data, a content analysis of open-ended questions is necessary to facilitate the categorization of responses. (See also the discussion of analysis of data in Chapter Five.)

Rating or ranking scales can be useful in appraising persons, courses, or program components. For example, honors program activities can be rated or ranked with respect to quality, interest, long-term benefits, and so on. Generally, rating scales ask subjects to respond to a theoretical continuum from, for instance, "Strongly Agree" to "Strongly Disagree." Ranking scales, on the other hand, typically ask respondents to place a list of items

in rank order. While easy to construct and often yielding useful information, ranked lists can be invalid if they do not represent meaningful categories to the subjects, or if they do not include a complete range of items for ranking.

Yet another self-report technique is the **semantic differential.** The semantic differential is helpful in assessing attitudes by asking subjects to indicate how closely their attitudes approximate those associated with opposing anchor points. For example, you might ask subjects to rate an honors activity on a five- or seven-point scale ranging from "rigorous" to "easy"; "confusing" to "straightforward"; "elitist" to "not elitist." The construction of a semantic differential scale is quite easy. Osgood provides a dictionary of words and terms which can be used as anchor points. However, this technique is somewhat difficult to score and requires the assistance of a statistician.

All of the self-report strategies are attractive because of their relative cost and ease of construction. A couple of potentially serious problems should be addressed, though. First, in all but the semantic differential there is a "response-set bias." In other words, subjects will often respond idiosyncratically to a rating or ranking scale: some may be extreme raters (i.e., choosing either end of the scale), while others tend to respond in categories in the middle of the scale. Second, self-report strategies yield information limited by what the respondent can be presumed to report accurately and honestly. If you want to assess the ability of students to engage in formal argument, for example, a self-report measure is probably not adequate to your needs.

B. Interviewing. Interviews are an appealing data-collection strategy. They may be structured or unstructured, and may be conducted face-to-face or by telephone; they are effective as a means of obtaining information about potentially sensitive matters like students' ethical development. However, all forms of interviewing are more expensive (of time and money) than self-report strategies. And if more than one interviewer is used, inter-rater reliability (the consistency with which raters collect, score and interpret the same information) must be ensured through training. Evaluations which employ multiple raters in settings like this must assess and report the degree of inter-rater reliability.

C. Written essays. Essays, a form of assessment common to the classroom, can also be used in more general studies to assess students' ability to contend with various complex issues. They are less expensive to administer than interviews, but inter-rater reliability must also be established and documented, and systematic scoring procedures must be developed.

D. Observation. This might be appropriate for assessing some types of honors program objectives. For example, if you are interested in measuring students' appreciation of the arts, you might observe variables like enthusiasm, interest, or knowledge while the students visit an art exhibition. Obviously, clear criteria must be established before the observation, and the observation itself must be systematic. The use of multiple observers raises the issue of inter-rater reliability (see above). A more serious problem is that persons who know that they are being observed often alter their normal behavior.

E. Performance measures. A real or simulated scenario can be constructed in which subjects must perform a specific task. For example, if a you are concerned with students' ability to argue logically, you could arrange for students to debate a controversial issue and then score them on their performance. This strategy is labor intensive and, hence, costly. The pay-off is that it can yield rich data about some hard-to-measure objectives.

F. Record review. Program records, which usually include such information as students' performance on standardized aptitude and achievement tests, can yield some objective information. Although record reviews are not appropriate for all honors program objectives, they are, where appropriate, an inexpensive and unobtrusive means of data-collection. A potential drawback is that records have a tendency to be incomplete or disorganized.

II. Sampling. Having made some tentative decisions about which data-collection techniques would be most appropriate for your needs, you will need to develop a sampling strategy. Sampling involves choosing a subset of subjects from a total population and usually results in savings of time and money. It is important to note, however, that sampling is not appropriate when it is easier or cheaper to assess the entire population, or when you do not have the services of someone trained in sampling methodology. The techniques outlined below are all "probability samples": all subjects in the

population have an equal and known chance of being included in the sample.

A. Simple random sample. This is one of the easiest strategies to use. It involves assigning sequential numbers to all members of the population and then, with the aid of a random number generator, choosing the designated number of sample members. Tables of random numbers are available in the *CRC Mathematical Tables Handbook*, and are accessible through most computer software packages. A significant disadvantage of the simple random sample is that many questions, especially in education, require the sampling of subgroups (e.g.. freshmen, sophomores, juniors, and seniors). That is addressed by the stratified random sample.

B. Stratified random sample. In contrast to the simple random sample, where a subset of members is chosen from a population, in stratified random sampling, the population is first separated into groups or strata, and then a simple random sample is chosen from within each stratum. Stratified random sampling can be conducted in one of two ways. Samples within each stratum can be chosen so that resulting strata have the same number of subjects (i.e., 25 freshmen, 25 sophomores, etc.) An alternative strategy would be to choose samples so that sample sizes represent proportionally categories in the population as a whole; in other words, if 30% of the population is freshmen, the sample would reflect that percentage. Results obtained from a stratified sample can be more precise than those from a simple random sample.

C. Systematic or 1 in K samples. The systematic sample is ideal when efficiency is the central concern. If you have an already compiled list of population members, you can avoid the somewhat burdensome task of using a random number generator with this strategy. Every Kth member of a population is chosen for inclusion in the sample, after the population list is entered randomly between 1 and K. (K is determined as follows: K=population size/sample size. For example, if your population size is 500 and the desired sample size is 50, then K=500/50 or 10. This means that after a randomly chosen point in the list between 1 and 10—say 7—every tenth subject thereafter would be chosen: subjects 7, 17, 27, 37, etc., would be included in the sample. It is important that the original list not be arranged in any way which is related to your measures (e.g., by GPA).

D. Simple cluster samples. To this point the unit of analysis has been the individual. Sometimes, because it is efficient and unobtrusive, it may be desirable to focus on groups, rather than individuals. For example, if you wished to examine the impact of small class size on the perceived quality of instruction, you might compare small honors classes to non-honors classes of a specified size. After securing a list of the courses fitting these criteria, you could randomly sample entire classes rather than sampling individual students from many different classes.

This question of sample size has both practical and scientific considerations. Even though you can expect that honors students will have an interest in cooperating in an evaluation, a non-response rate of 50% is fairly typical. You should anticipate non-response, and over sample. Resources aside, the reliability of an assessment procedure will increase with increasing sample sizes. Therefore, you should sample the largest number of persons that resources will allow.

Works Cited in Chapter Four and Suggestions for Further Reading

Osgood, C.E., C.J. Suci, and P.H. Tannenbaum. *The Measurement of Meaning*. Urbana, Illinois: University of Illinois Press, 1957.

CHAPTER FIVE: Analyzing and Interpreting the Results

After the decisions regarding focus and design have been made, and after data collection is complete, it is tempting to conclude that the evaluation is finished. In fact, two essential tasks remain. First, the data must be analyzed: that is, reduced from the form in which they were collected to a more easily handled form. Second, the analyzed data must be interpreted—studied carefully with a view to providing answers to the questions about the program which were originally posed when the focus of the evaluation was determined. These aspects of evaluation are discussed briefly in this chapter, because analysis of data is a highly technical topic, a detailed discussion of which lies outside the scope of this book. Interpretation of the results will usually be straightforward if the earlier steps have been carefully performed.

I. Analysis of Data. The information contained in the questionnaires, interview reports, transcripts, and test scores must be condensed into a concise form. The steps to be taken will depend upon the nature of the data collected, but they will generally fall under the headings of coding, statistical treatment, and presentation of findings.

Coding. Some data (e.g.,. responses to multiple-choice questions) will already be coded, and nothing will be required other than to transcribe or tally the results (e.g.,. "On question #4, of the 45 persons surveyed, 12 chose option A, 25 chose option B, 6 chose option C, and 2 did not respond.) Coding and condensing such data present no conceptual problems, but the process is tedious and time-consuming. Much effort can be spared (and accuracy improved) if data of this sort are collected directly in forms which can be read by computers (the ubiquitous op-scan sheet, for example).

Other forms of data (such as grade point averages) may need to be coded into categories. A list of honors students' GPAs is often less helpful than a categorized breakdown (e.g.,. 19% had GPAs between 3.3 and 3.7, etc.) This sort of coding is, again, a simple mechanical procedure which is well performed by computers.

Most difficult to code are qualitative items such as responses to subjective questionnaires or interview reports. These items can be of great value and are sometimes the only available form in which vital data can be gathered. But if the results are to be used quantitatively, it is essential to code the data into categories. The difficulty lies in ensuring that the categories are appropriate and that the coding is consistent. Suppose, in response to the question "What is the most important feature of the Honors Program to you?" we received the following responses:

1. The small classes
2. Intellectual stimulation
3. Stimulating classes
4. Chance to get to know professors
5. Classes
6. Opportunity to meet other interesting students

We might say that three of the six cited "classes" and that two cited "social opportunities" (one specifically mentioning faculty and one students). But what of response #2? Should we code it as citing "classes" (since that is where intellectual stimulation supposedly takes place)? Or might we argue that it is the social interaction with other minds which generates the stimulation and code this response with #4 and #6? Perhaps respondent #2 was thinking of the fact that honors students were given more challenging reading assignments; in this case the response does not fit either of the two categories so far defined.

While there is no easy solution for such difficulties, a coding strategy for these cases should address the following questions: 1. How many categories are to be used? Too many categories will make analysis of the data complex and interpretation difficult; too few will obscure important trends in the data. 2. How are the limits of each category to be defined (or "operationalized")? The more unambiguous the definitions of the categories, the cleaner will be the data and the easier the interpretation. 3. Is coding to be inclusive (a particular response may be coded into more than one category if it meets the specifications of both) or exclusive (a response is coded only to one single category)? Both methods have advantages, but statistical treatment of the data will differ depending on the method chosen. Above all, it is essential to be consistent: if some responses are given multiple codings, then all responses for which multiple coding would be ap-

propriate must be so coded. 4. If more than one person will be coding responses, how will inter-rater reliability be established? One common method is to have a small, random sample of responses coded by all of the coders; only if they agree on a high percentage of cases can the codings be relied upon. (Here again, the more precise the operationalization of categories, the better will be the inter-rater reliability.)

Statistical treatment of the data. The range of statistical procedures which may be called for in an evaluation range from simple averages (of GPAs, for example) to multiple regressions or factor analyses, which require highly trained personnel and sophisticated computer programs. The latter requirement is often more easily satisfied than the former; powerful statistical software packages are now commonplace on most campuses. The ability to choose appropriate statistical techniques and to apply them correctly is less common. Unless you are extremely well-versed in statistical procedures, it will be essential to obtain skilled assistance at this stage. If an outside evaluator is used, she or he should have (or have access to others who have) the necessary training. Departments of mathematics, psychology, and economics usually include faculty members skilled in statistical procedures.

Presentation of the results. It is important to set forth your findings in a way which makes them easy to understand. Tables, charts, graphs, and figures are useful ways of presenting quantitative material. Qualitative findings will require concise narrative description, examples, illustrative quotations, and occasionally even photographs or tape recordings. Suggestions for effective presentation of findings can be found in the APA Publication Manual. Here again, it is useful to have the advice of someone accustomed to writing descriptions of scientific studies for publication.

II. Interpreting the Results. The purpose of any evaluation is to answer questions about the program being examined. In the early stages of your evaluation you identified the objectives of your program, and designed the evaluation around questions of whether these objectives were being effectively attained. In the ideal case, the results of the analysis will speak clearly, directly, and unequivocally to these questions. Interpretation is simple in this event.

More commonly, however, some or all of your data will be related to program objectives in an indirect way. Perhaps direct measures would be prohibitively expensive; or perhaps none exist. A program might aim to produce graduates with a "a broad liberal education"; but no single measure of this has achieved widespread acceptance. As we stated in the chapter on evaluation design, one must then turn to indirect measures; and such measures must be interpreted. Interpretation draws the connection between what you aim for (e.g., enhanced artistic and cultural awareness on the part of students) and what you actually measure (e.g., the number of cultural events which your students voluntarily attended over the course of a year). Why should we consider the latter a measure of the former? The interpretation process is, once again, involved with the validity of the measures used; however, here we are not concerned with selecting a valid measure, but with clarifying why the measures used are the most valid ones available within the constraints which frame this evaluation.

Often interpretation will point out patterns of results among a variety of measures. There may not be any single test of "a broad liberal education," but if your data show that the students in question are (compared to appropriate control samples) more widely read, better versed in science and mathematics, more apt to speak a foreign language, possessed of a deeper sense of history, and more sensitive to aesthetic pursuits, this combination of characteristics begins to suggest that they have indeed received something which we might call a broad liberal education. Since data about the various attributes may not be clustered together when presented, it is the task of interpretation to pull together the related strands so that the point becomes clear.

Sometimes the process of interpretation yields surprises. Careful study of the data may demonstrate an unexpected strength or weakness in your program. While we have advocated that you undertake evaluation with concrete questions in sight, such "accidental" information may be of great value. Suppose, for example, students display a sharp increase in critical thinking ability between the sophomore and junior years. It is worth examining the sophomore curriculum carefully; is there a course which (perhaps inadvertently) develops reasoning skills? If so, perhaps some of the methods from this course could be applied in other parts of the curriculum.

CHAPTER SIX: Evaluating Special Projects

From time to time honors programs undertake projects of limited duration or scope. Examples might include foreign study programs, projects to develop courses or restructure curricula, recruitment drives, peer tutoring or peer advising programs, and experiments in instructional techniques. Evaluation of such projects is sometimes required (by a granting agency, for example); even when not mandated, systematic evaluation will help the honors director to determine the degree to which the project has succeeded and whether it should be repeated, modified, or discontinued.

The steps in evaluating a special project are identical to those followed in evaluating a program as a whole: determining the focus, establishing the design, collecting the data, analyzing the results, and interpreting the findings. In this chapter we will explore the differences between full-scale program evaluation and limited project evaluation in each of these steps. We will pay particular attention to non-quantitative techniques, since these are often especially well suited to evaluation of limited-scale projects.

I. Focusing the evaluation. Special projects have, in themselves, a more restricted focus than do programs as a whole. Programs are ongoing entities; projects usually have limited time spans. Frequently they involve small numbers of participants, compared with their parent programs. Often scant resources will be available for evaluating small-scale projects; hiring an external evaluator or using a very time-consuming instrument may be luxuries too costly to consider. (However, grant budgets often provide funds for evaluation of the project; some agencies require that project directors plan, budget, and arrange systematic professionally-conducted evaluations.)

Like an honors program, special projects will have specific objectives; these are apt to be more concrete and more explicitly stated than those of the program as a whole. This simplifies the first step of the evaluation, which you will recall begins with identification of the objectives. The intended outcomes will usually be quite clear. This does not mean that they will be easy to measure. Since the objectives are likely to be somewhat idiosyncratic, ready-made instruments may not exist. More than ever the evaluator

will need to be creative in custom tailoring instruments to measure the project's outcomes. Of course, existing measures can be used where they are appropriate, but the evaluator must avoid the temptation to measure what is easily measured regardless of the project's objectives.

Often qualitative (as opposed to quantitative) measures are useful when special projects are evaluated. This is true partly because it is easier to custom design qualitative instruments than quantitative ones and partly because the small numbers of participants in special projects make numerical methods less necessary (and sometimes impossible). If 200 honors students are asked "What is the most important thing you have learned from this program?" it is not feasible to list all of the responses in the evaluation report; some form of coding and quantification is essential. But to ask a similar question of the six students participating in a special project is a different matter; coding and quantifying the responses is scarcely worthwhile. It is better in this case to quote the answers of the six students directly (perhaps with some editing if the responses are lengthy.)

In the evaluation of honors programs, standardized quantifiable measures have the advantage of permitting easy comparison with other similar programs. Instruments which are custom designed and not easily quantifiable do not permit such comparisons. However, because special projects are often crafted in response to specific local needs, there is less need for these comparisons than when evaluating a program as a whole.

II. Designing the evaluation. Evaluation of special projects often calls for a simpler design than does comprehensive program evaluation. Because of the limited time-frame, longitudinal designs are seldom appropriate, although it is sometimes possible to compare pre- and post-assessments (the simplest form of longitudinal design). In many cases a single "snapshot" evaluation taken during or after the project will serve. In cases where the project results in a programmatic change (e.g., a new course or a curriculum revision), it is helpful to plan follow-up evaluations one or two years following the original project so that long-range effects can be studied.

Selection of appropriate control samples is also relatively easy. Since the project will usually affect only a subset of the honors population (student and/or faculty), a natural control group consists of a similarly-sized sample

of non-affected students and/or faculty. Depending upon the nature of the project, control samples from outside the honors population (or at least outside the *local* honors population) may be appropriate. Suppose, for example, that your honors program sponsors a study-abroad program for honors students. You may want to evaluate its effects on participating students both in comparison to honors students who do not participate and in comparison to non-honors students who take part in other overseas study programs. When controls outside the honors population are used, the precautions regarding proper selection of control samples listed in Chapter Three must be kept in mind.

III. Collecting the data. Methods of data collection are not remarkably different in special project evaluation than in whole program evaluation. The likelihood that the number of individuals affected is small simplifies the process of sampling: often it is feasible to study the entire population. (For control groups, however, sampling techniques may be required.) For the same reason (small numbers) it may be possible to use data collection techniques which would be too time consuming in a full program evaluation; written essays or extensive interviews from a handful of participants are far easier to handle than they would be if many dozens of individuals had to be included in the evaluation. Finally, of course, the limited scope of the evaluation makes it unlikely that data will have to be collected in machine-readable form.

IV. Analyzing the results. The principal difference in this stage arises from the fact that fewer quantitative measures are likely to have been used. The major headaches of coding the data can often be avoided (or at least minimized). Statistical treatment of the data will probably be less essential. It should be noted, however, that there exist *non-parametric* statistical procedures which may be applied to qualitative data, following a limited amount of coding. Even if you have collected quantitative data, special statistical handling may be required if the size of your sample is particularly small. Consult a statistician for further details.

V. Interpreting the findings. The task here is identical to the equivalent task in a whole-program evaluation: to examine the objectives of the project and ask whether your results indicate that these have been met. As before, this process is easiest if the validity of the measures with respect to the

objectives is clear. This will probably be the case in most special project evaluation; however, there may be cases in which it is important to elaborate on the results obtained from inferential or indirect measures. Just as with program evaluations, you may also want to emphasize points supported by several converging lines of evidence.

A full-scale evaluation of an honors program is likely to take months. Evaluation of a special project may require no more than a couple of days, and will rarely extend beyond a few weeks. However, it remains important that the evaluator exercise the same care for design, validity, and data collection.

APPENDIX

The following are books on the evaluation of educational programs:

Bloom, B.S., ed. *Handbook of Formative and Summative Evaluation of Student Learning*. New York: McGraw-Hill, 1971.

Bloom, B.S., G.F. Madaus, and J.T. Hastings. *Evaluation to Improve Learning*. New York: McGraw-Hill, 1981.

Cronbach, L.J., ed. *Evaluating Educational Programs and Products*. Englewood Cliffs: Educational Technology Publications, 1974.

Fink, A. and Kosecoff, J. *An Evaluation Primer*. Beverly Hills: Sage, 1980.

Joint Committee on Standards for Educational Evaluation. *Standards for Evaluations of Educational Programs, Projects, and Materials*. New York: McGraw-Hill, 1981.

Popham, W.J., ed. *Curriculum Evaluation*. Lexington, Massachusetts: D.C. Heath, 1974.

Popham, W.J. *Evaluating Instruction*. Englewood Cliffs, New Jersey: Prentice-Hall, 1973.

Popham, W.J., and E.L. Baker. *Establishing Instructional Goals*. Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

Rossi, P.H., H.E. Freeman, and S.R. Wright. *Evaluation: A Systematic Approach*. Beverly Hills: Sage, 1979.

Stufflebeam, D.L. *Educational Evaluation and Decision Making*. Itasca, Illinois: Peacock, 1971.

Webb, E.J. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally, 1966.

Worthen, B.R., and J.R. Sanders. *Educational Evaluation: Theory and Practice*. Worthington, Ohio: Charles A. Jones, 1973.

These are recent overviews of assessment. (Additional papers on assessment are available from the AAHE Assessment Forum, One Dupont Circle, Suite 600, Washington, D.C. 20036.):

Adelman, Clifford, ed. *Performance and Judgment: Essays on Principles and Practice in the Assessment of College Student Learning*. U.S. Department of Education: Office of Educational Research and Improvement. (To obtain this, contact the U.S. Government Printing Office, Washington, D.C. 20402.)

Alexander, Joanne, and Joan Stark. *Focusing on Student Academic Outcomes: A Working Paper*. NCRIPTAL Report. (See below.)

"Contexts for Assessment." AAHE Bulletin 41.2 (1988): 3-9.

Cross, K. Patricia, and Thomas Angelo. *Classroom Assessment Techniques: A Handbook for Faculty*. NCRIPTAL Report. (To obtain these and other NCRIPTAL reports, contact: National Center for Research to Improve Postsecondary Teaching and Learning (NCRIPTAL), Suite 2400 SEB, University of Michigan, Ann Arbor, Michigan, 48109-1259.)

The following is a list of sources for naturalistic evaluation:

Guba, Egon G. "Naturalistic Evaluation." In *Evaluation Practice in Review: New Directions for Program Evaluation*, no. 34. San Francisco: Jossey Bass, 1987.

_____. "What Have We Learned about Naturalistic Evaluation?" *Evaluation Practice* 8 (February, 1987): 23-43.

Guba, Egon G., and Yvonna Lincoln. *Effective Evaluation: Improving the Usefulness of Evaluation Results Through Responsive and Naturalistic Approaches.* San Francisco: Jossey-Bass, 1981.

_____. *Naturalistic Inquiry.* Beverly Hills: Sage, 1985.

Williams, David D., ed. *Naturalistic Evaluation: New Directions for Program Evaluation,* no. 30. San Francisco: Jossey Bass, 1986.

The following are articles about assessment and evaluation in the honors context:

Cohen, Ira, Ronald Dotterer, Robert Evans, Edward Napieralski, and William R. Whipple. "The New Assessment: Five NCHC Views." *The National Honors Report* 9.3 (1988): 1-11.

Reihman, Jacqueline, and Sara Varhus, "The Yardstick and the Pyramid: Outcome Evaluation and Honors Programs?," *NCHC Newsletter* 5.2-3 (1984): 1;4-5; "Goals and Measures: The Methodology of an Honors Evaluation," *NCHC Newsletter* 5.4 (1984): 6-9; (with E. Lonky) "Measures of Cognitive, Ethical, and Personal Development: A Challenge to Honors Education," *NCHC Newsletter* 6.1 (1985):9-10; "Honors Ideals and Accountability in Honors Programs: A Challenge," *NCHC Newsletter* 6.2 (1985): 8-10.