CHAPTER 14

□ Validity in Educational Research

*Margaret A. Eisenhart and
Kenneth R. Howe*

The Handbook of Qualitative Research in Education

Introduction

Validity—generally defined as the trustworthiness of inferences drawn from data—has always been a concern in educational research. Questions about validity historically arose in the context of experimentalist research and, accordingly, so did their answers. The emergence of nonexperimental, so-called "qualitative," methods in educational research over the past two decades, however, poses new questions. In particular, should experimentalist conceptions be applied to alternative research designs? If so, how? If not, what conceptions should be applied instead?

These are the kinds of questions we entertain in this paper. We begin with the conventional conception of validity as defined by Campbell and Stanley in the early 1960s and used by a generation of educational researchers working in the 1960s and 1970s. Next, we discuss several kinds of alternative conceptions that grew out of and responded to the special features of qualitative research as used in education. With these alternative conceptions of validity as our point of departure, we then develop our own position. A crucial feature of the position we develop is our distinction between *general* and *design-specific* standards of validity. The import of this distinction will become clear as the paper unfolds. Its basic thrust is that all educational research is subject to the same general criteria of validity even though quite distinct and specialized criteria are required to conduct and evaluate specific kinds of research studies. We end the paper with a discussion of how general validity and design-specific instances fit together.

The Conventional Approach

Conventional conceptions of validity in educational research derive most directly from Campbell and Stanley (1963).[1] Focusing on experimental and quasi-experimental designs, they divided validity into two kinds, internal and external. Internal validity, referred to as the *sine qua non* of good experimental design, pertains to the credibility of inferences that experimental treatments (factors) cause effects under certain well-defined circumstances. To meet the requirements of internal validity, other factors that reside in the way the study was conducted (i.e., "internal" to the design of the study) and that may have caused the effect must be ruled out. External validity pertains to generalizing the effects observed under experimental conditions to other populations and contexts. To meet the requirements of external validity (or to define its limits), factors that limit the study's application to other situations—factors such as the characteristics of the people, settings, or variables investigated (i.e., that are "external" to the conduct of the study)—must be ruled out.

Campbell and Stanley's basic approach was thus to require evidence that each type of validity has been met or approached. In the case of internal validity, the researcher must show that various "threats" (to be described later) that might affect the interpretation of results are controlled for in the research design. For external validity, the researcher must show that the characteristics of the people, settings, and variables that define the experimental conditions are unlikely to matter when the treatment is applied to other targeted populations and situations. It should be noted that despite Campbell and Stanley's claim that internal validity is the *sine qua non* of good research design, they emphasized the importance of external validity, particularly for educational research. According to them, because of the practical nature of educational research, "generalizations to applied settings of known character is the desideratum" (Campbell and Stanley, 1963:5).

As we indicated earlier, the emergence and subsequently acknowledged legitimacy of alternative, so-called qualitative, methods in educational research over the past two decades posed a challenge to the conventional conception of validity. Questions were raised about the appropriateness of using the conventional conception as a guide or standard for qualitative research. Some suggested that the conventional approach is epistemologically unsuited to be a general standard for most educational research. Others wondered, "If not this standard, then what?" There have been three major responses to this challenge: adaptations of the conventional approach, alternatives to the conventional approach, and eclecticism.

Adaptations of the Conventional Approach

Norman Denzin, a sociologist and author of *The Research Act: A Theoretical Introduction to Sociological Methods*, a textbook now in its third edition (1970, 1978, 1989), is frequently cited in papers about qualitative educational research. Denzin's book is devoted to descriptions and comparisons of research designs—Denzin calls them "methodologies"—commonly used by sociologists (specifically symbolic interactionists). In the 1989 edition, Denzin compared seven methodologies: experiments, surveys, participant observations, unobtrusive methods (techniques used by a researcher who is physically removed from the events being studied), life histories, interviewing, and filming.

Validity is a basis for Denzin's comparison of these research methodologies. He relied on Campbell's (1963a,b) list of eight factors that threaten internal validity and four factors that threaten external validity as criteria for assessing the methodologies. The eight threats to internal validity are historical events occurring between measurements, maturation of subjects between measurements, subject selection effects on results, interaction of maturation and selection effects on results, loss of subjects, testing effects, changes in instrumentation, and statistical regression toward the mean of groups originally chosen for their extreme scores. The four threats to external validity are differences in likely response to testing (reactive effects of testing), differences in likely response to treatment (reactive effects of experiment), multiple treatment interference, and subject selection efffects on applications. With only minor alterations to his definitions of the threats, Denzin compares each of the seven methodologies in terms of their ability to minimize the 12 threats.

Denzin explained, however, that minimizing the threats is not addressed in the same way in each research design. Threats may be approached experimentally, in which case the researcher can manipulate subjects, variables, and conditions, and threats can be controlled through this manipulation. When such direct manipulation is not possible, threats may be treated with multivariate analysis, whereby the researcher uses statistical manipulations to approximate the controls possible in a true experiment. In both experiments and multivariate analysis, causal inferences are accepted as valid when the probability that observed correlations are spurious or accidental is very low. A third and quite different approach to handling threats to validity is analytic induction. Using this approach, the researcher aims to identify directly time order, covariance, and other threats. This approach is used in the most qualitative of the designs Denzin discusses—participant observation and life history. Analytic induction requires that a researcher consider every piece of data before inferring causality. Support for (read: the validity of) the inferences advanced is strengthened by demonstrations that researchers have searched for data expected to severely test or negate inferences and that the researchers' emerging inferences accommodate, explain, or account for the variations discovered.

Denzin showed that each design has strengths and weaknesses with respect to minimizing threats, and he found experiments *and* participant observation especially strong overall (see the table on p. 30, 1989 edition). Because designs differ as to which threats they address most strongly, Denzin argued that when experimental and nonexperimental designs are compared, threat-by-threat, they tend to compensate for one another's weaknesses. Because different designs are more or less vulnerable to different threats, Denzin claimed (as did Campbell) that the most valid studies are those that rely on several research designs, thus reducing threats in as many strong ways as possible.

Judith Goetz and Margaret LeCompte, an educational anthropologist and an educational sociologist, respectively, adopted a strategy similar to Denzin's in their discussion of ethnographic analogues for Campbell and Stanley's internal and external validity. In their book, *Ethnography and Qualitative Design in Educational Research* (1984), they outlined threats to internal and external validity following Campbell and Stanley and illustrated ways of dealing with these threats within the context of ethnographic design (see especially pp. 220–232).

Like Campbell and Stanley [and also following Hansen (1979) and Pelto and Pelto (1978)], Goetz and LeCompte (1984) wrote:

> Establishing validity requires (1) determining the extent to which conclusions effectively represent empirical reality and (2) assessing whether constructs devised by researchers represent or measure the categories of human experience that occur. . . . *Internal validity* refers to the extent to which scientific observations and measurements are authentic representations of some reality; *external validity* refers to the degree to which such representations can be compared legitimately across groups [italics in original]. (p. 210)

Goetz and LeCompte imply that the spirit of Campbell and Stanley's threats can be translated into terms applicable to nonexperimental designs like ethnographies. For example, they suggest that the internal validity of ethnographic studies can be addressed in part by adopting procedures that increase the likelihood that an authentic picture of the participants' reality is elicited. Thus, the internal validity of ethnographic research is judged to be strong when researchers spend long periods of time in the field so as to get to know participants, their views, and situations; when the researchers' actions and interviews are conducted in the idiom of participants; and when the researcher is directly involved in the lives of those being studied. Goetz and LeCompte's approach to external validity is the same. They suggest, for example, that careful and extensive descriptions of the settings and people being studied, of the social conditions of study, and of the constructs being used give other researchers the information necessary to assess the typicality of a situation, to identify appropriate comparison groups and translation issues, and thereby to meet the requirements of external validity in the context of ethnographic research (see also Eisenhart, 1988).

Relying on Cook and Campbell (1979), Goetz and LeCompte also translate a third kind of conventional validity: construct validity, or the extent to which abstract ideas ("constructs") used in research studies (e.g., self-esteem, culture) match the empirical evidence used to indicate or measure the abstraction. Although construct validity has conventionally been used to assess the correspondence of test constructs to test items, Goetz and LeCompte (1984:225) argue that construct validity can be straightforwardly translated into ethnographic research practice when ethnographers demonstrate that the categories they are using are meaningful to participants or reflect the way participants actually experience reality.

Goetz and LeCompte also include in their discussion more global considerations that figure into the determination that an ethnographic study is valid. (Their global considerations foreshadow our general standards of validity.) They stress, for example, that the theoretical orientation guiding the research project—in their case, the structural–functional perspective of cultural anthropology—influences the way the general meaning of validity is translated into a research design, the factors that threaten validity, and the means of minimizing such threats. They emphasize that because the *primary* criterion for selection of a research design must be whether the design allows the researcher to address the research questions posed, the answer to *this* question may lead to an amalgamation of two or more of what they call the "ideal-typical abstractions" (Goetz and LeCompte, 1984:47–48) of research designs such as those set forth by Denzin.[2] Furthermore, they add to their list of attributes of a good study: completeness (does the report of the study contain all the elements considered necessary for a research report of this kind?), appropriateness (are the approach and design used effective and suitable for the research questions posed?), clarity (is it easy and straightforward to figure out what the study is about and why it is approached and designed in the way it is?), comprehensiveness (is the scope of the study large enough to address convincingly the questions posed?), credibility (are the conduct and results of the study believable?), and significance (does the study make an important contribution?) (pp. 233–245).

In summary, Goetz and LeCompte's adaptation of the conventional conception of validity entails a translation of the tenets of the conventional approach into criteria that make sense for ethnographic design. They also add more global considerations to their criteria for determining a study's validity. Their basic approach—to translate the conventional meanings of validity into ethnographic terms—differs from Denzin's approach, which is to consider how various designs, when used together, meet the requirements of the conventional approach. Yet both adapt Campbell and Stanley's definitions of internal and external validity so as to encompass nonexperimental as well as experimental research designs.

Alternatives to the Conventional Conception

A second response to the challenge posed by the emergence of alternative research methods is deep skepticism toward (e.g., Erickson, 1986) or outright rejection of (e.g., Lincoln and Guba, 1985; and, for different reasons, Roman, 1989) the notion that the conventional conception of validity may be fruitfully applied to alternative methods.[3] The ultimate basis for these more radical forms of divergence from the conventional conception is to be found in the various facets of the positivist–interpretivist–criticalist controversy, a discussion of which is beyond the scope of this paper (but see e.g., Bredo and Feinberg, 1982; Howe, 1985, 1988; Howe and Eisenhart, 1990; Roman, 1989; Roman and Apple, 1990). For present purposes, it is sufficient to note the emphasis that our first two exemplars of this more radical form of divergence—Erickson (1986) and Lincoln and Guba (1985)—place on the so-called "insider's perspective" and the emphasis that Roman (1989) places on exposing and transforming the power relations constituting research practice.

According to Erickson, the "basic validity criterion" of alternative methods is "*the immediate and local meanings of actions*, as defined from the actors' point of view [italics in original]" (Erickson, 1986:119). This criterion applies to the audience as well as the subjects of research.

Erickson discusses many of the same issues as Goetz and LeCompte, but in a way more commonly used for discussing literature. In Erickson's view, the crucial piece of "ethnographic validity" is the way the "story" is told and evidence for its authenticity provided [see also Van Maanen (1988) who argues for more "narrative ingenuity" in the way ethnogaphic accounts are written]. Erickson (1986) points out that the presentation of text-based data, most often in some kind of story form, has rhetorical, analytic, and evidentiary functions:

> The [story] persuades the reader that things were in the setting as the author claims they were, because the sense of immediate presence captures the reader's attention, and because the concrete particulars of the events reported in the [story] instantiate the general analytic concepts (patterns of culture and social organization) the author is using to organize the research report . . . . In sum, richness of detail in and of itself does not make a [story] ethnographically valid. Rather, it is the combination of richness and interpretive perspective that makes the account valid. Such a valid account is not simply a description; it is an analysis. Within the details of the story, selected carefully, is contained a statement of a theory of organization and meaning of the events described. (p. 150)

Erickson also emphasizes the need to meet criteria of quality with reference to how the results will be understood and used by various audiences. In his version of validity, concerns about clarity, appropriateness, and so forth take on the added burden of being clear, appropriate, and *useful* to potential audiences [e.g., teachers (see especially Erickson, 1986:153–156)]. This is a point we will return to later.

Lincoln and Guba (1985) have taken a more extreme position with regard to standards for nonexperimental educational research designs. Because of the special character of what they call "naturalistic studies," they advocate developing an entirely different set of standards by which to judge the soundness of naturalistic research. In their view, the two prime directives of naturalistic research, or "inquiry," are that the researcher does not influence or manipulate (or does so to a very limited degree) the conditions of study and that the researcher imposes no *a priori* categories on the results of the study (p. 8). They view naturalistic research as an "alternative research paradigm," an approach with a distinctly different ontological and epistemological basis from that underlying experimental research. As such, they propose that distinctly different research designs and different standards for validity must be used when conducting naturalistic research.

To refer to the overall quality of a piece of research, Lincoln and Guba (1985) use the term "trustworthiness" of research. They write, "The basic issue in relation to trustworthiness is simple: How can an inquirer persuade his or her audiences . . . that the findings of an inquiry are worth paying attention to, worth taking account of? What arguments can be mounted, what criteria invoked, what questions asked, that would be persuasive on this issue?" (p. 290).

Interestingly, despite the different labels and their contention that the different paradigm of naturalistic inquiry demands that standards be developed specifically for it, Lincoln and Guba begin their discussion with standards that are close analogues to those of Campbell and

Stanley. In particular, they list four kinds of trustworthiness, two of which, "truth value" and "applicability," are analogous to internal validity and external validity, respectively. Truth value refers to the accuracy (or "truth") of the findings for those beings studied. Applicabilitv refers to the likelihood that the findings will pertain to other groups in other situations. The other two kinds of trustworthiness are consistency (or reliability in experimentalist terminology) and neutrality (or objectivity). Lincoln and Guba argue that all inquirers are concerned about these general standards of trustworthiness, but the meaning of each standard, the nature of threats, and the means of minimizing them will be distinctly different within experimental (what they call "positivist") and naturalist paradigms. Thus, each paradigm will need its own ways to handle the threats.

Lincoln and Guba argue that naturalistic inquiry is fundamentally *not* about determining causes and, thus, that it is inappropriate to pursue truth value (internal validity) by demonstrating that causes and their effects have been isolated. They propose that the analogous standard for naturalistic inquiry—where the major aim is to reconstruct the perspectives of those being studied—is the demonstration that the researcher's interpretations of data (the findings) are credible to those who provided the data. Meeting this standard has two parts: first, carrying out research in such a way as to increase the chances that respondent categories rather than researcher categories will dominate the findings and, second, having respondents approve the researchers' interpretations (Lincoln and Guba, 1985:296). Specific "techniques" for meeting the standard of credibility are described by Lincoln and Guba (1985:301–316). These techniques are examples, meant to illustrate ways in which the naturalist's special form of validity may be operationalized. They include techniques for prolonged involvement with those being studied, techniques for systematically considering many sources of data, techniques for obtaining and analyzing data so as to be able to consider them from different angles and perspectives, techniques for refining working assertions or themes pertaining to the data, and techniques for respondents' review of researchers' findings.

Lincoln and Guba argue that the experimentalist's procedures for external validity—assuring the representativeness of treatment conditions to application conditions, particularly through randomization—*prime facie* do not apply to naturalistic inquiry. They believe that naturalistic inquirers are responsible only for clearly and comprehensively describing the contextual conditions of their studies. They argue that the establishment of external validity in naturalistic inquiry is an empirical matter and must be determined by those who wish to apply the findings somewhere else. Potential audiences for research findings must themselves determine whether the context in which they are interested is sufficiently similar to the context from which research findings derive to make their transfer possible and reasonable (Lincoln and Guba, 1985:298). Because the determination of external validity is made by potential users, no techniques are provided by Lincoln and Guba to meet this standard (see p. 316). Rather, they contend that the researcher is obligated to provide the "data base" or "thick descriptions" necessary to make judgments about application possible.

Leslie Roman has taken a very different extreme position on the validity of nonexperimental research (Roman, 1989; Roman and Apple, 1990). From her perspective as a feminist materialist, she contends that *both* experimental and naturalistic researchers have incorrectly assumed that they could achieve validity primarily by limiting the researcher's involvement ("subjectivity" or "bias") in the collection, analysis, and reporting of data. Experimentalists use various procedures, such as random assignment, double-blind controls, and statistical manipulations in an attempt to limit the researcher's influence and to constrain the generalizations drawn from specific results; naturalistic researchers attempt to hold their own views in abeyance to permit the emergence of the insiders' perspective and the inductive development of theory to explain and extend the results from a specific case or group. Drawing on a criticalist approach to educational research (see also Anderson, 1989), Roman argues that experimental and naturalistic researchers do not question the ways in which power relations of the wider society are perpetuated in research practice. Neither group takes seriously the possibility that research constructs, procedures, and results (be they in the form of variables or people's conscious models) sustain historically specific power relations and material interests. From the alternative perspective of the criticalist, control over who to study, what to study, how to conduct the study, and the relationship of the researcher to participants or subjects is always worked out in terms of the power relations governing the wider society, unless steps are taken to ensure that research studies are democratically designed and results are democratically produced. Democratization of educational research is the goal of critical education research (see also Lather, 1986). From Roman's standpoint as a feminist materialist (within the criticalist tradition), she argues that valid research must use a methodology that (1) resonates with the lived experiences of the group being researched, (2) enables members of the group to comprehend and transform their experiences of subordination, (3) reduces the divide between the researcher's intellectual work and group members' ordinary ways of describing and understanding their experiences, and (4) allows the researcher's prior theoretical and political commitments to be informed and transformed by understandings derived from the group's experiences (Roman and Apple, 1990:63–64). Needless to say, these features of a valid study are quite different from those suggested by either the conventional approach or the alternative approaches exemplified by Erickson or Lincoln and Guba.

Eclecticism

Many educational researchers who perceive important differences between experimental and alternative research designs nonetheless continue to have respect for and to be significantly influenced by Campbell and Stanley's two kinds of validity. Thus, a third response to the challenge to conventional validity posed by the emergence of alternative methodologies is a form of eclecticism in which criteria for validity accommodate ideas emanating from both experimental and alternative methodologies.

For example, Mary Lee Smith and Gene Glass, educational researchers, begin their book, *Research and Evaluation in Education and the Social Sciences* (1987), with a set of criteria for interpreting and judging the merits, that is, the validity, of educational research studies (pp. 2–6). They indicate that the criteria are generally applicable to any research design. However, as they proceed, they find that they must exempt one kind of research—naturalistic studies—from these general criteria.

Extending the tradition of identifying different kinds of validity, Smith and Glass (1987) list four: logical, construct, internal, and external. They write, "If the study has *logical validity*, the reader should be able to follow the argument and assess whether the hypothesis follows logically from the problem, whether the methods follow logically and consistently from the hypothesis, the findings from the methods, and the conclusions from the findings [italics in original]" (p. 2). A study has *construct validity* when the measures used by the researchers can be shown to correspond to the abstract "construct" under investigation (p. 4). Consistent with Campbell and Stanley, Smith and Glass add *internal validity*, which depends on ruling out alternative causes for the results of the study (p. 5), and *external validity*, which depends on demonstrating the generalizability of the results to other groups or situations (p. 6). Smith and Glass also say with reference to external validity, "In research, the people involved directly, the *sample*, are only of interest to the extent to which they inform us about similar groups of people not directly involved in the study [italics in original]" (p. 6). All but the first of these criteria have a decidedly experimental bent, by which control, randomization, and statistical manipulation are the primary means for establishing validity. However, the first standard, logical validity is a more extensive albeit inexact standard—one that focuses on the logic of decisions made in the design and course of the research project, rather than on the use of orthodox technical procedures. (This is yet another idea we will return to when we develop our general standards of validity.)

Smith and Glass's book also includes a discussion of "naturalistic studies," in which they find themselves unable to use three of the four criteria for validity outlined earlier. In striking contrast to their preceding statement about the role of the sample, Smith and Glass (1987) define the purposes of naturalistic studies to be "to understand the persons involved, their behavior and perceptions, and the influence of the physical, social, and psychological environment or *context* on them." They define the researcher's job to be "to describe [the persons involved] and interpret their actions for persons who have not been there and seen them directly—that is, for the readers [italics in original]" (p. 253). In fact, they go on to exempt naturalistic studies from all but the first criterion—logical validity—by arguing that the idiosyncracies of naturalistic studies defy the application of the kind of uniform standards that can be applied to the other research designs they treat, namely, experimental, quasi-experimental, causal-comparative, correlational, and survey studies. In lieu of standards, they propose some "issues" to be considered in assessing the quality of naturalistic studies. These issues include length of time in the field; the researcher's access to data from various sources; the researcher's subjectivity and biases; the clarity, completeness, and logic of the researcher's reasoning about the study; and the demonstration that final results have been obtained through triangulation (p. 278).

Their approach of listing "issues" for consideration as a guide for the conduct and assessment of naturalistic research is similar to the position described by J. K. Smith (1990, from Feyerabend). Smith suggests that the standards for some kinds of research are best thought of in terms of open-ended "lists" of general concerns that a researcher should address in some way in the research, rather than in terms of rules for admitting evidence or extending conclusions (the conventional approach).

Appraisal

Each of these three general responses to the challenge posed by the advent of alternative methods to the conventional conception of validity makes a significant contribution: adaptations of the conventional conception illustrate substantial commonalities that exist between experimental and alternative methods; alternatives to the conventional conception illustrate substantial differences; and eclecticism suggests that educational research can (perhaps) accommodate both.

In our judgment, however, more needs to be said, particularly with regard to explicating a general approach to validity that accommodates both "quantitative" and "qualitative" research designs. Because we deny that quantitative methods can be separated off and justified by appeal to a peculiar scientific (read: positivist) epistemology (Howe, 1985, 1988; Howe and Eisenhart, 1990), we think the pursuit of some general standards is appropriate and useful and, given a proper understanding of validity, unavoidable. On the other hand, we recognize that specific research designs have their own logic and coherence. Thus, a general approach to validity must accommodate differences among specific research designs.

Our aim is not to refute or dismiss the conceptions we have considered so far but, rather, to distill a more comprehensive account of validity. Our approach is to identify research studies with arguments (Dunn, 1982; House, 1977) and to define a valid argument as one that is credible in a general as well as a design-specific way.

The metaphor of research study as argument is useful in educational research for three reasons (following Dunn, 1982). First, the metaphor of argument discourages "facile distinctions between 'science' and 'ordinary knowledge,'" and the "patently false conclusion that knowledge derived from one source is inherently superior." Second, the metaphor "provides a conceptual framework that not only accommodates the experimental metaphor—including 'threats to validity' and their philosophic justification—but also permits a radical enlargement of standards for assessing and challenging knowledge claims." And third, the metaphor encourages the idea of public debate and scrutiny of research processes and results (Dunn, 1982:295).

Characterizing all educational research studies in terms of the general concept of an argument leads rather straightforwardly to a general concept of validity that can be applied across all such arguments regardless of their particular contents (for the application of this conception of validity to testing practice, see Messick, 1989). On the other hand, judgments regarding the validity of a particular argument also turn on whether the argument is credible to relevant audiences, allowing that the kinds of evidence and associated principles employed in particular arguments vary substantially. (This is a general point, not confined to validity in educational research: Consider valid argumentation in law versus physics.) Viewed in this light, the three approaches to validity in nonexperimental research described earlier (adaptation of the experimental or conventional approach, alternatives, eclecticism) fall somewhat short insofar as they encourage the view either that all arguments must be evaluated in terms of precisely the same criteria (adaptation) or that there must be different *kinds* of validity (alternatives to the conventional conception and eclecticism). In our view, it is more fruitful to think in terms of one *kind* of validity with different *design-specific instances*. Such a general conception of validity helps vitiate methodological imperialism and, at the same time, is consistent with the different kinds of knowledge and technical skills that go into marshalling and evaluating research-based arguments.

The position we will advance in the remainder of this paper has three parts. First, we argue that the field of educational research as a whole has certain concerns that transcend or are separate from those of specific disciplines or designs for research and that, for this reason, some *general standards* for the conduct of educational research that cut across all forms of educational research can and should be articulated. In our view, general standards should require that research studies be cogently developed, competently produced, coherent with respect to previous work, important, ethical, and comprehensive. We describe each of these features in more detail later.

Second, we think that although general standards of validity establish broad boundaries, they do not thereby dictate the specific strategies and techniques that researchers use when employing specific research designs. Instead, *design-specific standards*—which are subsumed by the general standards and which articulate the particular evidence, knowledge, principles, and technical skills that differentiate alternative designs—are required. Moreover, and as we have argued elsewhere (Howe and Eisenhart, 1990), such design-specific standards necessarily undergo revision and reconceptualization as scholars within various traditions conduct their work over time.

Finally, we consider how issues of substance and methodology peculiar to specific research designs may be construed as instances of variously interpreting and applying our general standards of validity. We will illustrate this relationship in the case of one specific design: educational ethnography.

Before turning to the articulation of our conception of validity, one further introductory point is in order. It has been suggested by some [e.g., J. K. Smith (1990)], that the emergence of alternative research designs (notably what he labels "constructivism," including the recent work of Lincoln and Guba; cf. Lincoln, 1990) may lead to the conclusion that standards of validity are ephemeral at best and can be no more precise than the everyday norms governing social interactions and negotiations. In our estimation, this view is far too extreme. Although neither static nor mechanically applicable, articulated standards of validity serve at least three important functions: They allow economy of thought in designing and evaluating educational studies; they provide the starting point for reflection on and improvement of the educational research enterprise; and they serve as the vehicle both for communicating within and across research traditions and for orienting newcomers (indeed, facilitating such forms of communication presumably is one of the major aims of this handbook).

Five General Standards for Validity in Educational Research

The five general standards we are about to advance can be usefully employed as guides for making valid arguments in educational research and can encompass, without undue constraint, distinct disciplinary and methodological arguments associated with specific research designs. Our first three standards are rules of thumb for systematic consideration of research studies *qua* arguments; they may be appropriately invoked across substantially different arguments, even though their precise application in a given study requires sophisticated and specialized knowledge. The fourth and fifth standards address more global requirements, whose application is not necessarily dictated in ways peculiar to specific designs.

Standard 1: The Fit between Research Questions, Data Collection Procedures, and Analysis Techniques[4]

Hilary Putnam remarks, "If you want to know why a square peg doesn't fit into a round hole, you had better *not* describe the peg in terms of its constituent elementary particles" (Rorty, 1982:201; attributed by Rorty.) Although Putnam's target is reductionism in scientific explanation, his remark also has a more prosaic meaning: The data collection techniques employed should fit, or be suitable for answering, the research question entertained. A corollary of this standard is that research questions should drive data collection techniques and analysis rather than vice versa.

We were certainly not, as graduate students or newly minted professors, the first to realize that the research designs presented to us in our courses and textbooks did not always fit the questions we most wanted to answer. As a graduate student, the expedient thing to do may be to whittle down the question so that a conventional research design could be used to address it. As researchers in a field in which major problems confront us, where significant debates about educational practice rage, and where person power and money to conduct research are limited, such an expedient solution is not justified. Instead, we think that methods sometimes must be modified, combined, and even created to address the research questions that need study.[5]

Correctly ordering research questions and methods, and developing their fit, is of course a complex issue. We do not mean to suggest that researchers can proceed as if they are blank slates—free of prior interests, commitments, and methodological expertise. Neither can they behave as if they have super intellects—capable of competently choosing from all of the relevant questions and methodologies. Nor, finally, can they operate as if they had available infinite time and resources. In some sense, then, research methodology will indeed drive research. On the other hand, the degree to which this occurs should be minimized. Research studies *qua* arguments have questionable validity when methodological preferences or matters of convenience, rather than research questions, drive the study design. Valid studies require cogently developed designs.

Standard 2: The Effective Application of Specific Data Collection and Analysis Techniques

In addition to deriving coherently from research questions, data collection and analysis techniques also must be competently applied, in a more-or-less technical sense. Research studies *qua* arguments cannot be valid without credible reasons for a specific choice of subjects, data-gathering procedures, and analysis techniques. Various principles guide how interviews should be conducted, how instruments should be designed, how sampling should proceed, how data should be reduced, and so forth, such that rather immediate "low-inference" conclusions are rendered credible. If credibility is not achieved at this level, then the more general (and more important)

conclusions that ultimately rest on these low-inference conclusions will be suspect.

It is not the case that educational researchers must create brand new principles and procedures for competently conducting their work. Principles and systematic procedures for the conduct and assessment of numerous qualitative (as well as quantitative) research designs have been formulated and debated for years within the social science disciplines. Although some modification of technical standards from the social sciences may be necessary for educational research purposes, it is incumbent on educational researchers, who wish to demonstrate that their techniques have been competently applied, to locate their work in the historical, disciplinary, or traditional contexts in which the methods used have been developed.

Standard 3: Alertness to and Coherence of Prior Knowledge

Linking research questions with data collection and analysis techniques and competently applying the latter do not assure that a study will render credible conclusions, because studies also must be judged against a background of existing theoretical, substantive, or explicit practical knowledge. For arguments to satisfy this standard, they must be built on some theoretical tradition or contribute to some substantive area or practical arena. In other words, the assumptions and goals embedded in the development and conduct of the study must be exposed and considered. Only if this is done can the arguments derived from a new study be placed in their appropriate context and the arguments of one study appropriately compared to those of other studies.

Perhaps less obvious is the researcher's own prior knowledge, or "subjectivity" (Peshkin, 1988). Peshkin has argued that subjectivity is the basis for the researcher's distinctive contribution, which comes from joining personal interpretations with the data that have been collected and analyzed. As with assumptions derived from the literature, subjectivities must be made explicit if they are to advance, rather than obscure, the validity of research *qua* argument.

Standard 4: Value Constraints

Gone are the days when it was philosophically respectable to believe it possible (and desirable) to bracket values in the design and conduct of social research, particularly in "applied" areas such as education. The conduct of educational research is subject to both "external" and "internal" value constraints (Howe, 1985). Valid research studies *qua* arguments must include discussion of values, that is, of the worth in importance or usefulness of the study and of its risks.

External Value Constraints

External value constraints concern whether the research is valuable for informing and improving educational practice—the "so what?" question. Research might be well designed and conducted in a *technical* sense, but that alone is an insufficient criterion of worth. Valid studies must be worthwhile. The concern with important issues, when considered in the context of educational practice, has several implications. One is that research investigations be comprehensive enough to convey and expose the important and profound problems and issues that arise for practitioners. This is not primarily a matter of increasing the scope of research projects so that more data can be collected and analyzed, or of developing sophisticated technical means for more rapidly and precisely handling data. Rather, it means committing the educational research community to multifaceted investigations of major educational issues—whether they be at the level of pedagogy, policy, or social theory—and then demanding that researchers ground their methodology in the nature of these issues.

Admittedly, judgments of the worth of research projects can be very difficult to make. They have the potential to be exceedingly biased, as anyone who has served on a human subjects committee can attest. However, these are not judgments from which researchers can (or do) forever run and hide [witness the recent exchange in *Educational Researcher* between Finn (1988) and Shavelson and Berliner (1988) in which they debate whether educational research has or has not made an important contribution to the improvement of educational practice; also see the more recent discussion by Philip Jackson (1990), also in *Educational Researcher*]. Researchers are best advised to put questions about the worth of research immediately on the table, lest implicit judgments about worth or lack of it operate behind the scenes, as a kind of hidden agenda. Clearly, even if others might be puzzled about the study's worth, educational researchers themselves should be able to communicate what value their research has (if only potentially) for educational practice.

The conclusions of educational research also should be accessible to the general education community. That is, the language of the results and implications must be cast in a form that is understandable to and debatable by various audiences (those who might read accounts of the research) or stakeholders (those who have a material interest in the results or uses of the research) in a particular setting—teachers, administrators, and parents, as well as educational researchers with varying perspectives and expertise. Accordingly, researchers must give attention to the social, political, and cultural features of the contexts and individuals they investigate and to which the results might be applied (Erickson, 1986; House, 1980:Chap. 12; Weiss, 1983). Researchers must also be sensitive to the inevitably value-laden language that they employ—terms such as "at risk," "developed," or "culturally different"—to avoid mystifying their findings and cloaking them in a false "scientific objectivity."

Valid research studies *qua* arguments, then, should explicitly address, in language that is generally accessible to the community of interested parties, the importance of the research and its (potential) usefulness. This requirement facilitates and encourages public debate of educational issues and of the implications of research results.

Internal Value Constraints

Internal value constraints refer to research ethics. We call research ethics "internal" because they concern the *way* research is conducted vis-à-vis research subjects, not with the (external) value of results. For example, Stanley Milgram's (1974) research on obedience to authority rendered valuable insights regarding the power of researchers to elicit compliance from subjects to perform ethically objectionable actions. The way Milgram treated his subjects was highly objectionable, however—so much so that he would not be permitted to do his research today. (Ironically, Milgram's findings, at least indirectly, underpin current requirements for informed consent, especially those that require researchers to communicate clearly to subjects that they are free to withdraw from research at any time and without penalty.)

Internal value constraints are distinguishable from other concerns about validity insofar as observing them sometimes requires reducing the precision and certainty of findings. For instance, randomized double-blind experiments are notorious for the kind of trade-off they engender between the risk : benefit ratio that applies to the subjects of such research and the value of the knowledge that can be obtained for guiding future action. These concerns are especially relevant to "qualitative" researchers because they must weigh the quality of the data they can gather (and whether they can gather any data at all) against principles such as confidentiality, privacy, and truth-telling. Although internal value constraints, or research ethics, can be distinguished from more conventional issues of research credibility, they are nonetheless crucial to evaluating the legitimacy of research designs and procedures, and thus we believe to the validity of a research study.

Standard 5: Comprehensiveness[6]

Our fifth standard—comprehensiveness—encompasses responding in a holistic way to and balancing the first four standards as well as going beyond them. We mean "comprehensiveness" in three senses. First, with regard to standards 1–3, standard 5 demands a judgment about the overall clarity, coherence, and competence—what might also be called "overall theoretical and technical quality"—of the study. Second, with respect to standards 1–4, standard 5 requires a balancing of the overall technical quality, the value and importance of the study, and the risks involved in the study. As indicated earlier, meeting one standard, such as protection of human subjects, may require tradeoffs against other standards. This second aspect of standard 5 calls for thoughtful consideration and explanation of such tradeoffs.

Third, standard 5 requires comprehensiveness in the sense of being alert to and able to employ knowledge from outside the particular perspective and tradition within which one is working and being able to apply general principles for evaluating arguments. For example, Denzin (1989), Goetz and LeCompte (1984), and Shulman (1988) argue that "triangulation by theory"—or application of various explanations to the data at hand and selection of the most plausible one to "explain" the research results—is a powerful strategy for establishing the validity of a theoretical explanation. It may also be considered a strategy for comprehensiveness by demonstrating that a study, competently and ethically conceived and conducted, can stand up to the challenge posed by other approaches or different results. When researchers demonstrate that, or explain the reasons why, other relevant approaches should be rejected or disconfirming data should be questioned, their studies are more comprehensive than when they do not.

Our discussion of general standards in the context of educational research can be summarized and concluded as follows: All instances of valid research-based arguments in educational research, regardless of design-specific peculiarities, take the same general form—that is, important educational issues must serve as the basis for formulating important research questions and an appropriate and ethical research design; research questions and methods must be competently linked, methods must be competently applied, prior commitments must be exposed; the potential worth of the results must be weighed against the risks associated with the study; and, overall, a comprehensiveness must be achieved that balances design quality and importance against risks and permits the robustness of conclusions to be assessed. As these requirements were discussed, it should have become clear that the understanding of validity we are proposing is a unitary construct. The five standards are not independent of each other; they cannot be applied separately. They are interrelated and must be considered together. The import of a unitary and holistic construct of validity is clarified in the next section.

Design-Specific Standards

Our five general standards are designed to encompass, without undue constraint, the specific standards and norms of particular research designs (e.g., ethnographic research, quasi-experimental research, survey research). In this way, valid arguments in educational ethnography or test construction, for example, become instances (not kinds) of our general conception of validity. At the specific level, building a case for validity requires meeting the requirements of the general standards with reference to the underlying assumptions, topics, and methodological techniques associated with a given design. However, design-specific tenets may have little to do with investigations in education, because often the designs have been developed in the social or natural sciences for inquiry into other phenomena. Thus, the requirements of our five standards must be sufficiently general to accommmodate considerable variation among specific designs that might be used in educational research, yet be discriminating enough to differentiate the validity of various studies *qua* arguments for educational research. What would it mean in practice, then, to apply the five general standards to a particular research design or an individual study?

In the next three sections we will focus first on one specific research design—educational ethnography—and then on one ethnographic study—reported by Holland and Eisenhart in *Educated in Romance: Women, Achievement, and College Culture* (1990)—to illustrate how the five standards might be applied and how the fit between general and design-specific standards might be achieved.

Assessing Educational Ethnography

One major assumption guides our discussion of the validity of specific research designs: What counts as a valid argument in the context of a specific research design and what steps are sensible to take to establish that an argument is valid will depend on the tenets of the specific design tradition. In other words, the design must be informed by the work and workers within that tradition (even if only to indicate how a study will depart or diverge from that tradition). Furthermore, within traditions, what constitutes a valid study will change over time (Howe and Eisenhart, 1990). Thus, we expect that the manner of addressing our five general standards will be affected by the history, norms, and ongoing debates of the tradition in which a particular study is conducted.

Using the case of educational ethnography, we illustrate that standards 1–3 are not meaningful as criteria for validity unless considered together; they are not independent criteria that can be separately applied and met in some studies but not others. Furthermore, in the case of educational ethnography, meeting the requirements of standard 3 (identification of the relevant body of previous work and the researcher's commitments) is prior and crucial to determinations of whether standard 1 (cogent development) or 2 (competent application) can be met. Although statements about the ethnographic logic of standard 1 or the ethnographic criteria for standard 2 are sometimes made in the abstract, we argue that the application of either one of the first two standards cannot stand without standard 3. Similarly, standard 3 is a hollow component of research validity unless tied to the requirements of standards 1 and 2.

Second, meeting the requirements of standards 1–3 can be pointless, costly, or even harmful without also satisfying the requirements of standards 4 (value constraints) and 5 (comprehensiveness). In other words, the validity of ethnographic research, or any other specific research design, for educational research depends on all five standards taken together.

We begin our discussion of ethnographic research with reference to the first three standards. We find that the first three general standards (cogent design, competent application, and connection to previous work) can be addressed largely from within the ethnographic tradition. That is, ethnographers and others can rely on traditions of scholarship and established norms in cultural anthropology and fieldwork sociology to locate and appropriately design their ethnographic research studies. When we then turn to standards 4 and 5, however, we find that they require consideration of matters not customarily treated within the ethnographic tradition.[7] This is really no surprise, because standards 4 and 5 help define research as relevant to the practice of education.

Our Standards 1–3 and the Ethnographic Research Tradition

To determine whether our first three standards of general validity for educational research are met in the case of an educational ethnography, we must ask "Is there credible evidence, pursuant to the disciplinary tenets underlying ethnography, that data collection and analysis procedures were cogently developed from research questions, and that these procedures were competently applied?" To answer these questions for an ethnography, we would begin by trying to identify the disciplinary context in which the study and its methodology were conceived. For purposes of illustration, we will focus on the context of cultural anthropology or, more specifically, on one of its subareas—educational anthropology.

Identification of the appropriate disciplinary context is not necessarily a simple matter. Because disciplinary traditions of scholarship are multifaceted and often divided into distinct subareas, it is of paramount importance to identify the specific subarea of work in which a study is located. For example, although the general purposes and assumptions of educational anthropology can be identified (Eisenhart, 1988), many subareas, in which small groups of researchers pursue particular topics in specialized ways, also exist [see, for example, the authors writing about systematic ethnography, microethnography, feminist materialist ethnography, and discourse analysis in this volume, and, for a general discussion of these subareas, see Jacob (1987, 1988) and the rejoinder by Atkinson, Delamont, and Hammersley (1988)]. The subareas share some general orientations, such as a commitment to identify the sociocultural processes that constitute education in a particular setting, and general assumptions, such as that human behavior and human learning are responsive to a context that is pervaded by patterns of culture and social relations that are, as well, interpreted and reconstructed by participants. However, within subareas, educational anthropologists make different decisions about the topics of major importance, the primary assumptions, and methodological preferences. It is these subareas to which particular studies are addressed and in which research designs including procedures are actually worked out.

The importance of clearly defining or identifying with certainty the scholarship tradition before attempting to assess the research design or methods used can be elucidated with a simple illustration of the implications of using one or another definition of culture when conceiving an ethnographic study. Many cultural anthropologists take the theoretical position that culture consists of the meanings that society, by partitioning the world through its institutions, language, and the collective activities of groups, encourages members to hold. But cultural meanings might take several empirical forms. Meanings can be represented in the organization of social life; that is, in the way institutions (schools, families, occupations, religions, etc.) allocate and represent roles, responsibilities, and rewards (Geertz, 1987). Meanings also can be represented in the words that people use to describe the world and their place in it (Quinn and Holland, 1987). (There are many other ways meanings might be represented; we use these two for illustrative purposes only.) When studying any kind of meaning, anthropologists may consider insider perspectives (those meanings recognized by members of the group), outsider perspectives (usually those meanings identified by a researcher), or interactive perspectives (those meanings that arise when insiders and outsiders communicate with each other).

To anticipate the research designs and procedures necessary for a particular study of meaning then, it is first necessary to identify the kind(s) of meaning of interest [i.e., to identify the tradition of scholarship and/or the commitments of the researcher (standard 3)]. In the first case (institutional meanings), the research design and procedures must address at least two research questions to meet the standard 1 requirement for cogent development: What is the evidence that the meanings attributed to the institution are understood (through compliance, resistance, or opposition) by those who participate in or observe it? And what is the evidence that meanings attributed to an institution are pervasive in the society where it exists? Given these research questions, methodological procedures must be devoted to competently collecting relevant evidence and triangulating evidence from numerous participants in and observers of institutions, as well as across institutions, of the society. Where evidence of similar interpretations by insiders and outsiders is provided, the findings are stronger than if evidence were provided from only one source. Where patterns across institutions can be provided, confidence is increased that the findings (meanings) are pervasive.

In the case of the second type of meaning (cultural models), first-person accounts of events and actions, such as those given by participants or those given by "observers" about their own experiences (cf. Kirkup, 1986; Van Maanen, 1988), are especially necessary to provide a basis for researcher inferences about collective meanings. Accounts made by ethnographers who try to become insiders could be considered useful in this sense too, although less so than true insider accounts. Similarly, insider corroboration of outsider accounts is weaker evidence for a finding—in this case—than insiders' own accounts. When meanings are provided by researchers who infer them from first-hand accounts, evidence is needed to demonstrate that inferred meanings can encompass or predict actions of the people to whom the models are attributed (cf. Eisenhart and Holland, in press). Finally, when interactive perspectives are of interest, evidence is needed that people from different positions or backgrounds come to take the same or similar meaning from observed actions or during their activities together (cf. Tobin, Wu, and Davidson, 1989).

Thus, for the validity of an educational ethnography to be judged in terms of our first three general standards, the study's place in a subarea tradition must be identified first. From there, the criteria used within the subarea to identify a good ethnography can be provided, thus establishing the design-specific norms by which standard 1 (cogent development of the research design from the research questions), standard 2 (competent application of procedures), and the remainder of the standard 3 requirement (to make clear the researcher's prior commitments or subjectivities) can be meaningfully assessed.

Our position on the interdependence of the first three standards differs from the position of those who would use salient characteristics of ethnographic methodology alone to develop a good (valid) ethnographic study. Spindler (1982), perhaps taking for granted a set of theoretical commitments, proposed such a list, which he called "criteria for a good ethnography." His criteria included observations must be contextualized, prolonged, and repetitive; hypotheses, questions, and instruments for the study should emerge as the study proceeds; judgments about what is most significant to study should be deferred until the orienting phase of the field study has been completed; participants' views of reality are revealed by inferences drawn primarily from direct observation and various forms of ethnographic interviewing; sociocultural knowledge—both implicit and explicit—that participants bring to and generate in social settings should be revealed and understood (Spindler, 1982:6–7). Although these criteria can be taken as features of many ethnographic studies, they cannot serve well as guides to cogent research designs or the competent application of techniques unless researchers can show that the features make sense, given specific research purposes.

In the context of this discussion of validity in ethnographic research, orthodox ethnographic techniques (e.g., participant observation, spending a long period of time in the field, learning the customs and language of the group, triangulating data sources and methods)—about which a great deal has been written both by educational anthropologists and educational researchers—may be ways to achieve validity, but their presence in a study is not sufficient to demonstrate that validity has been achieved. Evidence that these procedural steps were taken cannot stand in place of answers to questions about why the research topic was conceived as it was, the nature of the assumptions or commitments made, and a rationale for the research questions asked. The mere presence of familiar procedural steps cannot, by themselves, provide convincing evidence for the validity of anthropological or ethnographic arguments, nor should the steps be constraints on efforts to meet demands for evidence in other relevant or innovative ways. The test of a valid argument in light of our first three standards, in educational anthropology as elsewhere, lies fundamentally in the appropriateness of methods used given purposes selected. And scholars well-versed in the purposes of a subarea are in the best position to make good selections and to pass judgments on the appropriateness of methods.

The application of our fourth and fifth general standards (regarding value constraints and comprehensiveness) to educational ethnography takes a different form. To meet these requirements, ethnographers must take additional steps beyond those normally considered sufficient by the community of educational anthropologists.

Meeting the Requirements of Standard 4, or Establishing the External and Internal Value of Ethnographic Research

Regarding external value constraints, cultural anthropologists usually assume that their research questions and findings will be of interest primarily to other anthropologists or other students of human behavior. These uses are thought to be informational or advisory—as food for thought—by others interested in explaining or understanding sociocultural phenomena. For educational anthropologists who participate in the educational research community, another use (whether intended or not) is to interpret, affect, or change educational practice. Although some educational anthropologists have argued that they do not intend their research to be used in "applied" contexts, we argue that the general standards of validity proposed earlier do and should call upon educational anthropologists to put their purposes, interests, and insights pertaining to educational practice into plain language for public debate, even if ethnographic convention does not. Educational anthropologists should be asked to clarify their claims for the worth or power of sociocultural theories in contrast to, for example, claims by psychologists or economists. Anthropologists' claims about the worth of their studies for the improvement of educational practice must, we think, be made explicit and face the challenges brought by other educational specialists and practitioners. By and large in educational anthropology/ethnography, this has not been a major concern, although we know of no good reasons why ethnographers should not engage in such debates.

Turning to our standards pertaining to internal value constraints, the issues here include: Whose privacy is threatened, or peace-of-mind disrupted, by the research? For what or for whom will knowledge gained from the research be used? What are the personal and social implications of eliciting such knowledge, knowing it, and using it? In other words, who is privileged or disadvantaged, who receives the benefit, and who pays the price? In summary, is there evidence indicating that the study's purposes and results outweigh any risks?

Considered in this light, it is startling to realize how infrequently educational ethnographers have discussed the internal value of their work, at least publicly. In fact, the lore of educational anthropology includes the recommendation that researchers not divulge their ideas, plans, or worries to those being studied until after the work has been completed. Although this approach is consistent with ethnographic convention, it is not consistent with our general commitment to internal value constraints as outlined in standard 4. To meet our general requirements for validity in educational research is, we think, to make these internal value considerations explicit and thorough-going in the entire design and conduct of educational ethnography. This requirement places a new limit on what ethnographers can study—covert studies or studies in which informed consent cannot reasonably be obtained would be prohibited. The requirement also extends the ethnographer's obligation to apprise research participants of what the study is about and what its likely outcomes will be throughout the entire period of the study.

Meeting Standard 5, or Establishing the Comprehensiveness of Ethnographic Research

To meet the additional requirements (beyond what is required to meet standards 1–4) of comprehensiveness in the conduct of educational anthropology/ethnography is to balance the requirements of the first four standards and to place one study's arguments and evidence alongside alternatives, both from within educational anthropology and without it, that is, from the educational research community. Educational anthropologists must make their design standards and conventions, as well as their decisions about ethical and other tradeoffs, clear to others outside their own community of scholars. Then, they must enter into debates about the most compelling explanations, the most convincing evidence, and the most useful and least harmful ways of thinking about educational reform and taking action pursuant to it. To some extent, educational anthropologists already do this, as do research specialists in other fields. However, many of these so-called debates occur among close associates or specialists within a subarea and never reach a level where divergent perspectives clash and must be reconciled if reform of practice or policy is to follow. As we said about external value constraints, we do not think educational ethnographers have good reasons to remain outside the fray.

In the next section, we discuss the validity for educational research of a specific educational ethnography. In using this example, we suggest what must be debated and decided by educational researchers with the assistance of educational ethnographers. We emphasize that a determination of the validity of an individual study for educational research purposes depends on two things: (1) an application of our five general standards that is sensitive to the research conventions of the tradition in which the study was conceived and (2) the researcher's ability to clearly, cogently, and comprehensively describe the study with respect to the general standards.

Assessing One Educational Ethnography

The book *Educated in Romance: Women, Achievement, and College Culture* (1990), by Dorothy Holland and Margaret Eisenhart, analyzes the college experiences and career commitments of a small number of academically talented black and white women who began college in 1979. The research reported in the book included an ethnographic study of the women during their first 2 years of college, follow-up interviews with the women in 1983 and 1987, a survey, and a series of ethnosemantic studies of a larger number of college women and some men. Here we discuss the ethnographic study only.

If we were to assess the validity of this ethnographic study according to the position we have taken earlier, we would have to ask: "Does this study make a valid argument in a general way (for educational research) as well as a design-specific way (for educational anthropology)?" In other words, does this study measure up to our standards for

general validity as well as to design-specific standards for ethnography?

It should be noted at the outset that at the general level we are applying standards to this work that Holland and Eisenhart did not anticipate. They did not write their book primarily for an educational research audience. Anthropologists themselves, they wrote for their colleagues in anthropology. One question before us is whether the argument developed in this book meets design-specific standards for a good ethnography. A second question, and more important for this paper, is whether the argument in the book also meets the requirements of general validity for educational research as described in our five general standards. Because the two of us are not in a good position to make a conclusive judgment about the validity of this study—we do not, after all, constitute either the community of educational anthropologists or educational researchers—we use the example primarily to illustrate the kinds of questions that come up and must be decided when the validity of a specific study for educational research is being considered.

If we begin with the standard 3 requirement to locate the ethnography within a subarea tradition of educational anthropology, we find that the ethnographic study reported in *Educated in Romance* was originally designed within the context of one subarea (devoted to explorations based on theories of symbolic interactionism), but the ethnographic data were eventually used to address research questions derived from another subarea (devoted to explorations based on theories of social reproduction). In the book, the authors devote considerable space—two chapters—to locating their final work in a tradition of scholarship. Part of a third chapter is devoted to an explanation of the two authors' own interests, attitudes, and biases pertinent to the study and its evolution. In general, we found extensive coverage of information relevant to assessing how well the study measures up to standard 3.

But given the standard 1 requirement for cogent development of the research design, what are we to make of the switch from one scholarship tradition to another? On the one hand, the authors provide considerable information about the fit between both their original and final research questions and design. They discuss in some detail how they came to realize that their original ideas about how college life would influence the career-related decisions of college women—the ideas that led to their research questions and design—were not borne out by their data. Based on previous research including some of their own, the authors originally anticipated that student peer groups would exert a direct influence on women's thinking and actions related to majors, careers, and other plans for adulthood. The ethnographic study was designed to investigate this influence by addressing the following four research questions (Holland and Eisenhart, 1979:16–17): (1) What is the content of male versus female roles and identities as promoted by the peer groups of college-age females? (2) What is the process through which college-age women are affected by their peer group in choice of college major? (3) What variation is there among peer groups of college-age women with respect to content of male versus female roles and identities and what seem to be factors promoting these inter-peer-group differences? (4) To what extent can peer-group characteristics "explain" differences in choice of majors by college-age women?

However, the authors found that the women's peers had very little information about each other's career-related plans and cared very little about them. The authors were forced to ask themselves a different question: What were the women and their peers interested in and what did they spend their time doing on campus? The need for the second question became obvious only after most of the ethnographic data had been collected and a preliminary analysis of some of the data had been completed. At about the same time, new theoretical perspectives and debates were emerging in educational anthropology and affected the authors' thinking about their study and data. Thus, the data were ultimately used to address a different set of research questions than originally intended (Holland and Eisenhart, 1990:59–60). These questions were the following: (1) What were the women's responses to the university? (2) How did their responses oppose, if they did, the patriarchal conditions that they faced? (3) How did their everyday experiences, their "lived culture," enter into the "choices" and "decisions" that they were making about their future careers and domestic arrangements? (4) What role did the peer group play in affecting university women's "choices" and "decisions" about their future lives? (5) What were the important divisions within the peer group and the important issues of "gender politics" within the student body? The final research questions were derived from the researchers' experiences with their data and new ways of thinking; these questions would not have occurred to them in 1979.

Did method drive research in an invalid way here? Yes and no. In one sense, the data obtained from the ethnographic study determined the future course of the study. On the other hand, ethnographic research is well known for just the sort of flexibility illustrated in this study. According to ethnographic research tradition, flexibility is valuable to the extent that it permits the researcher to adjust her or his original research questions or procedures to fit the special characteristics of those being studied (see Goetz and LeCompte, 1984; Spradley, 1980; also Spindler, 1982, cited earlier). Often, changes in research design, questions, and procedures are considered *necessary* to produce valid ethnographic results, that is, to demonstrate that the participants' culture, not that of the researchers, is being described and analyzed. In this light, unexpected evidence, not the method *per se*, made the change necessary and served to validate it, at least in the eyes of ethnographers. However, with respect to our standard 1 for educational research, a question remains about the fit between the original methods and the later research questions. Were the methods appropriate for the new questions?

Additional light can be shed on this question by referring again to ethnographic convention. At the time Holland and Eisenhart (1979) formulated their study, subareas of educational anthropology devoted to studies of symbolic interaction and social reproduction shared some commitments to ethnographic research design [compare for example the symbolic interactionist purposes and ethnographic designs described in Spindler's *Doing the Ethnography of Schooling* (1982) or Erickson and Shultz' *The Counselor as Gatekeeper* (1982) with the social reproduction purposes and ethnographic designs in Everhart's *Reading, Writing, and Resistance* (1983) or Willis' *Learning to Labor* (1977)]. In both subareas, the criteria for a good ethnography as outlined by Spindler (1982) would have applied in 1979. Holland and Eisenhart's study met all these criteria. However, by the mid-1980s, when Holland and Eisenhart were analyzing and writing up their findings, ethnographic research design criteria for studies based on theories of social reproduction (and their various revisions) were being reconceptualized along lines similar to Roman's position discussed earlier in this paper (Roman and Apple, 1990). Thus, it seems that the research design used in *Educated in Romance* was cogent, for its time and place in ethnographic tradition, but it might not be were the study conceived today.

Regarding our general standard 2 (requiring competent use of procedures), the authors of *Educated in Romance* took the approach of describing their procedures in the appendix of their book. Like most ethnographers, they did not comment directly on their reasons for selecting the procedures they used or on the limitations of their procedures. They relied primarily on the conventions and shorthand descriptions in which they had been trained as cultural anthropologists and on the power of the data revealed in the book to establish the appropriateness and quality of their techniques. This is standard operating procedure for cultural anthropologists (Geertz, 1988) and, thus, may be considered adequate, among ethnographers, to establish their competence in using ethnographic procedures. However, although enough information is provided to know *what* procedures the authors used, we cannot learn enough about the authors' reasons for using certain procedures to meet the spirit of standard 2. If educational researchers are to translate among diverse studies of similar topics, they must be told about the reasons for as well as the conduct of their methodological procedures.

Not surprisingly, Holland and Eisenhart also followed ethnographic convention regarding value constraints in their study. Before the study began, they apprised potential study participants of the nature of their involvement (e.g., that researchers would be spending large amounts of time with participants, that researchers would try to get to know participants as friends and to understand their worlds as they did). They explained the topic of the study at the time. They promised confidentiality and obtained written consent. However, they felt no special compunction to alert participants to later decisions to change the topic of the research or to have participants review or approve the researchers' interpretations. In fact, their silence was so complete that during the final follow-up interviews in 1987, some of the participants told the interviewers that they did not want to give any more information until they could read what had been written about them.

Again, this approach is consistent with ethnographic convention, but it is not consistent with our general commitment to internal value constraints in educational research as outlined in standard 4. To meet our general requirements for validity in educational research is, we think, to make these internal value considerations explicit throughout the design, conduct, revision, and interpretation of a research study.

With respect to standard 5 (comprehensiveness), the overall quality, balance of tradeoffs, and durability of *Educated in Romance* remain to be debated within the ethnographic and educational research communities. Our brief review of the book indicates some of the questions that must be addressed in the debate. Because the book is so new, the relevant communities are just now beginning to read it and assess it.

Summary

Returning to our general standards with the illustrations from educational anthropology and *Educated in Romance* in mind, we find that adhering to the first three general standards means identifying the subarea of scholarship to which a particular anthropological study of education is intended to contribute, formulating timely research questions (for the subarea), and choosing research methods that will permit the questions to be addressed, in that order. Although a commitment to meet our general standards 1–3 may require more explanation of background assumptions and methodological strategies than would be necessary within anthropology, meeting the general standards does not necessarily require steps different or additional to those ordinarily required in educational anthropology. In other words, the first three standards do not constrain or change the focus of work within the subarea. This is not the case for the other two standards. Standards 4 and 5 require researchers to address more general but serious questions about the significance of the research, the use and manipulation of human subjects as a part or consequence of the work, the researchers' commitment to and success at explaining and using the study's results and its implications for constructive change, and the ability of the research to stand up to public debate of its merits and worth.

In general, the treatment we have given above to educational anthropology and ethnography serves as an illustration of how we could assess the validity of arguments based on other specific research designs (both qualitative and quantitative) for educational research. We think it likely, however, that other specific designs will measure up to our standards for general validity in different ways. For example, those who conduct experimental studies seem to have very well-developed conventions for handling and describing their methodological competence (standard 2), yet their articles may include very little about the scholarship traditions and commitments that underlie or motivate their work (standard 3). Naturalistic inquirers, on the other hand, seem to have well-developed ideas about handling value constraints but lack clear or agreed-upon standards for research design or procedures. Naturalistic inquirers may be able to rely on the standards developed within their own tradition to meet the requirements of standard 4 but may have to look elsewhere for help to meet the requirements of standard 2. In another contrast, educational ethnographers, who have well-developed standards for research design but only limited conventions for handling value constraints, may rely on their subarea tradition to meet the requirements of standard 2 and look elsewhere for advice about how to meet standard 4.

Conclusion

We observed at the outset of this chapter that the appearance and subsequent growth of the use of qualitative methods in educational research spurred interest in developing formal standards for assessing

Campbell, D. T. (Overman, E. S., Ed.). (1988). *Methodology and epistemology for social science: Selected papers.* Chicago: University of Chicago Press.

Campbell, D. T., and Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cook, T. D., and Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Denzin, N. (1970, 1978). *The research act: A theoretical introduction to sociological methods* (1st and 2nd eds.). New York: McGraw-Hill.

Denzin, N. (1989). *The research act: A theoretical introduction to sociological methods* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Dunn, W. N. (1982). Reforms as arguments. *Knowledge: Creation, Diffusion, Utilization,* 3(3), 293–326.

Eisenhart, M. (1988). The ethnographic research tradition and mathematics education research. *Journal for Research in Mathematics Education,* 19(2), 99–114.

Eisenhart, M., and Holland, D. C. (in press). Gender constructs and career commitment: The influence of peer culture on women in college. *In* T. L. Whitehead and B. Reid (Eds.), *Gender constructs and social issues.* Champaign, IL: University of Illinois Press.

Erickson, F. (1986). Qualitative methods of research on teaching. *In* M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 119–161). New York: Macmillan.

Erickson, F. (1989). Validity in qualitative research. Paper presented at meeting of the American Educational Research Association, San Francisco, March.

Erickson, F., and Shultz, J. (1982). *The counselor as gatekeeper: Social interaction in interviews.* New York: Academic Press.

Everhart, R. B. (1983). *Reading, writing, and resistance: Adolescence and labor in a junior high school.* Boston: Routledge and Kegan Paul.

Finn, C. E. (1988). What ails education research? *Educational Researcher,* 17(1), 5–8.

Geertz, C. (1987). Interpretive anthropology. *In* H. Applebaum (Ed.), *Perspectives in cultural anthropology* (pp. 520–524). Albany: State University of New York Press.

Geertz, C. (1988). *Works and lives: The anthropologist as author.* Stanford, CA: Stanford University Press.

Goetz, J. P., and LeCompte, M. D. (1984). *Ethnography and qualitative design in educational research.* New York: Academic Press.

Hansen, J. F. (1979). *Sociocultural perspectives on human learning: An introduction to educational anthropology.* Englewood Cliffs, NJ: Prentice-Hall.

Holland, D. C., and Eisenhart, M. A. (1979). Women's peer groups and choice of career. Proposal for research project. Washington, DC: National Institute of Education.

Holland, D. C., and Eisenhart, M. A. (1990). *Educated in romance: Women, achievement, and college culture.* Chicago: University of Chicago Press.

House, E. R. (1977). *The logic of evaluative argument.* CSE Monograph Series in Evaluation. Los Angeles: Center for the Study of Evaluation, University of California.

House, E. R. (1980). *Evaluating with validity.* Beverly Hills, CA: Sage Publications.

Howe, K. (1985). Two dogmas of educational research. *Educational Researcher,* 14(8), 10–18.

Howe, K. (1988). Against the quantitative–qualitative incompatibility thesis (or, dogmas die hard). *Educational Researcher,* 17(8), 10–16.

Howe, K., and Eisenhart, M. (1990). Standards for qualitative (and quantitative) research: A prolegomenon. *Educational Researcher,* 19(4), 2–9.

Jackson, P. W. (1990). The functions of educational research. *Educational Researcher,* 19(7), 3–9.

Jacob, E. (1987). Qualitative research traditions: A review. *Review of Educational Research,* 57(1), 1–50.

Jacob, E. (1988). Clarifying qualitative research: A focus on traditions. *Educational Researcher,* 17(1), 16–24.

Kaplan, A. (1964). *The conduct of inquiry.* San Francisco: Chandler.

Kirkup, G. (1986). The feminist evaluator. *In* E. R. House (Ed.), *New directions in educational evaluation* (pp. 68–84). London: Falmer Press.

Lather, P. (1986). Research as praxis. *Harvard Educational Review,* 56(3), 257–277.

LeCompte, M. D., and Preissle, J. (in press). *Ethnography and qualitative design in educational research* (2nd ed.). San Diego: Academic Press.

Lincoln, Y. (1990). The making of a constructivist: A remembrance of transformations past. In E. Guba (Ed.), *The paradigm dialog* (pp. 67–87). Newbury Park, CA: Sage Publications.

Lincoln, Y., and Guba, E. (1985). *Naturalistic inquiry.* Beverly Hills, CA: Sage Publications.

Messick, S. (1989). Validity. *In* R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan Publishing.

Milgram, S. (1974). *Obedience to authority: An experimental view.* New York: Harper & Row.

Pelto, P. J., and Pelto, G. H. (1978). *Anthropological research: The structure of inquiry* (2nd ed.). Cambridge: Cambridge University Press.

Peshkin, A. (1988). In search of subjectivity—One's own. *Educational Researcher,* 17(7), 17–22.

Phillips, D. (1987). Validity in qualitative research: Why the worry with warrant will not wane. *Education and Urban Society,* 20(1), 9–24.

Preissle-Goetz, J. P. (1989). Validity in qualitative research. Paper presented at meeting of the American Educational Research Association, San Francisco, March.

Quinn, N., and Holland, D. (1987). *Cultural models in language and thought.* Cambridge: Cambridge University Press.

Roman, L. (1989). Double exposure: Politics of feminist research. Paper presented at the Qualitative Research in Education Conference, University of Georgia, Athens, GA, January.

Roman, L., and Apple, M. (1990). Is naturalism a move away from positivism? Materialist and feminist approaches to subjectivity in ethnographic research. *In* E. Eisner and A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 38–73). New York: Teachers College Press.

Rorty, R. (1982). Method, social science, social hope. *In* R. Rorty (Ed.), *Consequences of pragmatism* (pp. 191–210). Minneapolis: University of Minnesota Press.

Shavelson, R. J., and Berliner, D. C. (1988). Erosion of the educational research infrastructure. *Educational Researcher,* 17(1), 9–12.

Shulman, L. (1988). Disciplines of inquiry in education: An overview. *In* R. M. Jaeger (Ed.), *Complementary methods for research in education* (pp. 3–17). Washington, DC: American Educational Research Association.

Smith, J. K. (1990). Alternative research paradigms and the problem of criteria. *In* E. Guba (Ed.), *The paradigm dialog* (pp. 167–187). Newbury Park, CA: Sage Publications.

Smith, M. L., and Glass, G. V. (1987). *Research and evaluation in education and the social sciences.* Englewood Cliffs, NJ: Prentice-Hall.

Spindler, G. D. (1982). General introduction. *In* G. D. Spindler (Ed.), *Doing the ethnography of schooling: Educational anthropology in action* (pp. 1–13). New York: Holt, Rinehart & Winston.

Spradley, J. P. (1980). *Participant observation.* New York: Holt, Rinehart & Winston.

Tobin, J., Wu, W., and Davidson, D. (1989). *Preschool in three cultures: Japan, China, and the United States.* New Haven: Yale University Press.

Van Maanen, J. (1988). *Tales of the field.* Chicago: University of Chicago Press.

Weiss, C. H. (1983). The stakeholder approach to evaluation: Origins and promise. *In* A. S. Bryk (Ed.), *Stakeholder-based evaluation.* New Directions for Program Evaluation (no. 17, pp. 3–14). San Francisco: Jossey-Bass.

Willis, P. (1977). *Learning to labor: How working class kids get working class jobs.* New York: Columbia University Press.

Wolcott, H. (1990). On seeking—and rejecting—validity in qualitative research. *In* E. Eisner and A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 121–152). New York: Teachers College Press.