8 ludging the Quality of Fourth Generation Evaluation

If we accept the definition of disciplined inquiry as set forth by Cronbach and Suppes (1969), it seems clear that standards for judging the quality of such inquiry are essential. The Cronbach and Suppes definition (1969, pp. 15-16) suggests that disciplined inquiry "has a texture that displays the raw materials entering into the arguments and the local processes by which they were compressed and rearranged to make the conclusions credible." Thus a disciplined inquiry process must be publicly acceptable and open to iudgments about the "compression and rearrangement" processes involved.

We have, in another context, argued that evaluation is properly construed as one form of disciplined inquiry (Lincoln & Guba, 1985, 1986b) along with two other forms of such inquiry: research and policy analysis. Neither of the two latter, however, should be confused with evaluation, which has its own intended products, audiences, and outcomes (Lincoln 8 Guba, 1985, 1986b).

It is, as a result, incumbent on us to deal with the question of the nature of quality criteria that may be appropriate primarily to evaluation, particularly in view of the fact that we have proposed a form of evaluation that differs in such dramatic ways from the first three generations' predecessors. A useful beginning is to consider those standards and criteria that have been devised for conventional evaluation.

Standards and Criteria for Conventional Evaluation

Not surprisingly, the first such criteria took the form of *test* standards, which exist currently in the form of the 1974 revision of the *Standards for Educational and Psychological Tests.* These standards were developed by a joint committee of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (American Psychological Association, 1974).

This committee early on recognized the importance of tests used in program evaluations and intended to devote a section of their standards to this topic. For a variety of reasons-for instance, lack of time and necessity to limit the scope of the test standards document (see Joint Committee, 1981, p. 142)—the Joint Committee on Test Standards decided not to proceed along these lines, but instead recommended to the parent groups that a new joint committee devoted especially to this task be appointed. Responsive action was taken in 1975, with the establishment of the Joint Committee on Standards for Educational Evaluation, under the directorship of Dr. Daniel Stufflebeam.

The Joint Committee on Standards for Educational Evaluation undertook its task with enthusiasm, believing that there would be

> several benefits from the development of sound standards: a common language to facilitate communication and collaboration in evaluation; a set of general rules for dealing with a variety of specific evaluation problems; a conceptual framework by which to study the oft-confusing world of evaluation; a set of working definitions to guide research and development on the evaluation process; a public statement of the state of the art in educational evaluation; a basis for self-regulation and

accountability by professional evaluators; and an aid to developing public credibility for the educational evaluation field. (1981, p. 5)

The Joint Committee did not intend to devise criteria standards that would in any way inhibit the growth of evaluation as a field of professional activity. The committee, for example, disclaimed having a particular view of what constitutes good education. They claimed that they attempted to recognize in the standards all types of studies used in evaluation. They wanted to encourage the sound use of a variety of evaluation methods (both quantitative and qualitative, whenever and wherever appropriate). And because the members of the Joint Committee were themselves experienced evaluators, they wrote the standards in ways that would help evaluators identify and confront political realities in and around the projects that they evaluated (Joint Committee, 1981, pp. 5-7).

It is our best judgment that, while the *Standards*[1] devised by the Joint Committee are not especially congenial to the posture of fourth generation evaluation, neither are they destructive of its aims and processes. The interested reader is invited to study the ***Standards*** from this perspective. Our position is that we can live with these standards although they are by no means very powerful for judging the quality of a given evaluation on those matters that are of central importance to fourth generation evaluation. Quality criteria are, of course, the central focus of this chapter, and we shall shortly take up the question of how one judges the processes and products of a fourth generation evaluation. But another word is in order regarding standards.

The ***Standards for Evaluation Practice*** (Rossi, 1982), developed by a committee of the Evaluation Research Society, on the other hand, are absolutely unacceptable from the standpoint of the fourth generation evaluator. (In 1987, the Evaluation Research Society joined with the Evaluation Network to form the American Evaluation Association. Currently, the American Evaluation Association operates with two sets of standards; the organization has been quite open about accepting either set as a guide to professional practice. For reasons that will become clear, we believe it is impossible to move in the direction some have suggested to merge the two sets of standards.) The *ERS* ***Standards*** are divided into six sections, felt to be "roughly in order of typical occurrence" (p. 11): formulation and negotiation, structure and design, data collection and preparation, data analysis and interpretation, communication and disclosure, and utilization. The ERS ***Standards*** embody a series of assumptions—especially with respect to the evaluator role, and to some extent, the relative power of the client-that are not acceptable to fourth generation evaluators, and that, furthermore, make their merger with the Joint Committee's ***Standards*** unlikely to impossible (Lincoln, 1985). Some analysis of why that is so will make the issue clearer.

First, it is assumed that interaction between client and evaluator is likely to be limited to those contacts needed to "formulate and negotiate" (Rossi, 1982, p.12) and to "communicate and disclose" (p. 15). But fourth generation evaluators argue that negotiation occurs continuously throughout an evaluation, as does data analysis and the interpretation process. (Experienced evaluators who are not themselves fourth generation evaluators intuitively understand this to be the case in many instances.) Evaluation activities, they say, are cyclic, feedback-fcedforward in nature, and iterative, whereas the ***ERS Standards*** paint them as linear and highly sequenced, implying that there are cutoff times for each activity beyond which no more of that activity actually occurs.

Second, strong emphasis is placed on the quantitative and experimental aspects of evaluation. Words like "treatment," "sampling," "reliability," "validity," "generalizability," "replicability," and "cause-effect" relations leave little doubt about which methods are believed to possess the most power. In fact, the preface to the "Structure and Design" standards itself specifies that evaluation ***case*** studies are "as subject to specification as the design of an experimental study" (Rossi, 1982, p. 13). Nowhere are qualitative methods (in the service of *any* paradigm) given direct approval. In the discussion of the adequacy of methodologies; no attempt is made to specify criteria appropriate to more constructivist/responsive evaluation efforts. Case studies are treated simply as looser variants of scientific technical reports.

Third, the emphasis on uncovering cause-effect dimensions or relationships flies in the face of constructivism's denial of the efficacy of that concept. No mention is made of the fact that often no satisfactory "cause" *can* be isolated for a given "effect." The mandated search for statistically significant cause-effect relationships often blinds evaluators and clients alike to more diffuse but also more powerful social forces operating within a given context, program, project, or site. And, most certainly, ascertaining what people think exists and why they think so is at least as important as verification of some a priori postulate about cause-effect relationships *that the evaluator thinks exists.*

Fourth, the *ERS Standards* call, both explicitly and implicitly, for the evaluation to be "shaped" in such a way as to meet the information and decision making needs of the "client," who is, typically, the person or agency that both has the power to commission the evaluation (the legal authority) and is the agent who is to contract and pay for it. There is, of course, nothing inherently wrong in ensuring that an evaluation meets a client's information or decision-making needs. In fact, that is why many of them are mounted. But the fourth generation evaluator has a much expanded idea of who ought to have access to information, who ought to have to power to withhold it (certainly not the "client" or funder), and who ought to be involved in decision making. Servicing decision-making or information needs, particularly for a single person or agency, serves only to concentrate power in the hands of those who already possess inordinate power relative to program participants, targets, or stakeholders. Having such limited foci for evaluations typically is disenfranchising and disempowering to the many other stakeholders who are invariably involved. The fourth generation evaluator typically refuses to accept a contract in which information is released only at the discretion of the "client" (funder). To try to fulfill such a criterion, given constructivist methodology, would be impossible in any event. The fourth generation evaluator considers decision making only one of the many objectives to be served in any given evaluation effort. Thus this emphasis on shaping the evaluation with the client's information and decision-making needs in mind would not only unnaturally but also unnecessarily and unethically limit the range of activity of the fourth generation evaluator.

Finally, the ERS *Stundurds* also fail to recognize the role that values play not only in evaluation (a process based, after all, on the root premise of values, which might properly be *expected* to pay attention to values) but in inquiry more broadly and generally. The continuing emphasis in this set of standards on objectivity and freedom from bias-criteria that are, after all, grounded in positivist ontological and epistemological assumptions-ignores mounting recognition by even the conventional scientific community that all science, and certainly social science, is value-bound (Bahm, 1971; Baumrind, 1979, 1985). In retrospect, the possibility of acting to "value" a project (program, curriculum, and so on) while acting as though values were unimportant or corrupting to the valuing (evaluation) effort should have struck us long ago as bizarre, if not contradictory, behavior. Hindsight is always 20-20.

Our overall conclusion, then, is that, while certain existing standards may be usefully applied to fourth generation evaluations, others are not only not useful but are actually destructive of the aims of fourth generation evaluation. Specifically, the Joint Committee's *Stundurds* may be applied, and would do no harm, but those of the ERS are contradictory to and exclusive of the goals of the fourth generation. Even those that are useful, however, are probably not as powerful as others that we will shortly suggest.

Criteria for Judging the Adequacy of Fourth Generation Evaluation

Is it possible to identify standards that seem more appropriate to fourth generation evaluation? Are there standards that are also more powerful than the all-purpose professional standards currently available, such as the Joint Committee *Stundurds?* We believe that there are three different approaches to considering the quality of goodness of a fourth generation evaluation (or, for that matter, any constructivist inquiry): invoking the so-called *parallel* or quasi-foundational criteria, which we have typically termed the *trustworthiness* criteria; considering the unique contribution made to goodness or quality by the *nature of the hermeneutic process itself;* and invoking a new set of non-foundational criteria-but criteria embedded in the basic belief system of constructivism itself-which we have termed the *authenticity* criteria. We shall take up each in turn.

The Parallel Criteria (Trustworthiness)

These criteria for judging adequacy (goodness, quality) are called the *parallel,* or *foundational,* criteria because they are intended to parallel the rigor criteria that have been used within the conventional paradigm for many years. Typically, conventional criteria for Judging the rigor of inquiries include internal validity, external validity, reliability, and objectivity.

Internal validity is defined *conventionally* within the *positivist* paradigm as the extent to which variations in an outcome or dependent variable can be attributed to controlled variation in an independent variable (Lincoln & Guba, 1985, p. 290); or, as Cook and Campbell (1979, p. 37) put it, the "approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause." Assessing internal validity is the central means for ascertaining the "truth value" of a given inquiry, that is, the extent to which it establishes how things really are and really work. Establishing truth value involves asking the question, "How can one establish confidence in the 'truth' of the findings of a particular inquiry for the subjects [sic] with whom and the context in which the inquiry was carried out?" There are a number of putative threats to the internal validity of any inquiry-including history, maturation, testing, instrumentation, statistical regression, differential selection, experimental mortality, and selection-for which the inquiry design must compensate, either by controlling and/or randomizing processes.

External validity can be defined (positivistically) as "the approximate validity with which we infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect and across different types of persons, settings and times" (Coo& Campbell, 1979, p. 37). External validity has as its purpose a response to the applicability (or generalizability) question: "How can one determine the extent to which the findings of a particular inquiry have applicability in other contexts or with other subjects [*sic*]?" Just as there are threats to internal validity, there are, in conventional inquiry, threats to external validity, including selection effects, setting effects, history effects and construct effects (Guba 8 Lincoln, 1985, pp. 291-92; Lecompte Goetz, 1982). When these threats are taken care of, then a given study should have applicability to the larger population from which the smaller sample was drawn.

Reliability (in positivist terms) responds to questions about the consistency of a given inquiry and is typically a precondition for validity, because a study that is unreliable cannot possess validity (Lincoln & Guba, 1985, p. 292). Reliability refers to a given study's (or instrument's) consistency, predictability, dependability, stability, and/or accuracy, and the establishment of reliability for a given study typically rests on replication, assuming that every repetition of the same, or equivalent, instruments to the same phenomena will yield similar measurements. In conventional inquiry, reliability can be threatened by several factors, including any careless act in the measurement or assessment process, by instrumental decay (including "decay" of the human instrument), by assessments that are insufficiently long (or intense), by ambiguities of various sorts, and by others. The question that determines consistency is usually some variation of this one: "How can an inquirer decide whether the findings of a given inquiry would be repeated if the inquiry were replicated with the same (or similar) subjects [sic] in the same (or a similar) setting or context?"

Objectivity responds to the positivist demand for neutrality, and requires a demonstration that a given inquiry is free of bias, valuesand/or prejudice. The guiding question is this: "How can an inquirer establish the degree to which the findings of a given inquiry are determined only by the subjects [sic] of the inquiry and the conditions of the inquiry, and not by the biases, motivations, interests, values, prejudiceand/or perspectives of the inquirer (or his/her client)?" Typically, freedom from the contamination of values or bias in a study is warranted in either of two ways: intersubjective agreement, or the utilization of a methodology and a set of methods that are thought to render the study impervious to human bias or distortion. The experiment is believed to be such a method by adherents of positivism. Other, lesser methods threaten objectivity by permitting inquirer (and others') values to reflect or distort the "natural" data, by creating the possibility that openly ideological inquiry (e.g., feminist or neo-Marxist) may be pursued, or by legitimating the data and information generated by a single "subjective" observer.

Within the framework of logical positivism, the foregoing criteria are perfectly reasonable and appropriate. This is the case because internal validity, external validity, reliability, and objectivity are grounded-that is, have their foundational assumptions rooted in-the ontological and epistemological framework of that paradigm (model, worldview) for inquiry. But the traditional criteria are unworkable for constructivist, responsive approaches on axiomatic grounds (Guba & Lincoln, 1'981; Lincoln 8 Guba, 1985).

It is clear that internal validity, which is nothing more than an assessment of the degree of isomorphism between a study's findings and the "real" world, cannot have meaning as a criterion in a paradigm that rejects a realist ontology. If realities are instead assumed to exist only in mentally constructed form, what sense could it make to look fisomorphisms? External validity, a concept that embodies the very essence of generalizability, likewise can have little meaning if the "realities" to which one might wish to generalize exist in different forms in different minds, depending on different encountered circumstances and history, based on different experiences, interpreted within different value systems. Reliability is essentially an assessment of stability-of the phenomena being assessed and of the instruments used to assess them. Ordinarily it is assumed that phenomena are unchanging (at least in the short haul), so that any instrument that assesses them ought, on replicated readings, provide essentially the same assessment (otherwise it is judged unreliable). But if the phenomenon can also change-and change is central to the growth and refinement of constructions-then reliability is useless as a goodness criterion. Finally, objectivity clearly reflects the positivist epistemological position thatsubject/object dualism is possible, but if a rival paradigm asserts that interaction (monism) is inevitable, what can objectivity mean? As Morgan (1983) has noted so well, goodness criteria are themselves rooted in the assumptions of the paradigm for which they are designed; one cannot expect positivist criteria to apply*in any sense* to constructivist studies, including fourth generation evaluation.

What then might be criteria that *are* meaningful within a constructivist inquiry? As a first approximation to answering this question we set about to develop a set parallel to those conventional four, staying as close as possible to them conceptually while adjusting for the changed requirements posed by substituting constructivist for positivist ontology and epistemology. This process-trying to understand what might be criteria appropriate to the axioms themselves-gave rise to the following criteria (Lincoln 8 Guba, 1986a).

Credibility. The credibility criterion is parallel to internal validity in that the idea of isomorphism between findings and an objective reality is replaced by isomorphism between constructed realities of respondents and the reconstructions attributed to them. That is, instead of focusing on a presumed "real" reality, "out there," the focus has moved to establishing the match between the constructed realities of respondents (or stakeholders) and those realities as represented by the evaluator and attributed to various stakeholders. There are several techniques for increasing the probability that such isomorphism will be verified, or for actually verifying it. Included among those techniques (widely recognized by anthropologists, sociologists, and others who engage in fieldwork) are the following:

(1) *Prolonged engagement:* Substantial involvement at the site of the inquiry, in order to overcome the effects of misinformation, distortion, or presented "fronts," to establish the rapport and build the trust necessary to uncover constructions, and to facilitate immersing oneself in and understanding the context's culture (Lincoln 8 Guba, 1986a, pp. 303-304).

(2) Persistent observation: Sufficient observation to enable the evaluator to "identify those characteristics and elements in the situation that are most relevant to the problem or issue being pursued and [to focus] on them in detail" (Lincoln & Guba, 1986a, p. 304). The object of persistent observation is to add depth to the scope which prolonged engagement affords.

(3) Peer *debriefing:* The process of engaging, with a disinterested peer, in extended and extensive discussions of one's findings, conclusions, tentative analyses, and, occasionally, field stresses, the purpose of which is both "testing out" the findings with someone who has no contractual interest in the situation and also helping to make propositional that tacit and implicit information that the evaluator might possess. The disinterested peer poses searching questions in order to help the evaluator understand his or her own posture and values and their role in the inquiry; to facilitate testing working hypotheses outside the context; to provide an opportunity to search out and try next methodological steps in an emergent design; and as a means of reducing the psychological stress that normally comes with fieldwork-a means of catharsis within a confidential, professsional relationship.

(4) *Negative case analysis:* The process of revising working hypotheses in the light of hindsight, with an eye toward developing and refining a

given hypothesis (or set of them) until it accounts for all known cases. Negative case analysis may be thought of as parallel or analogous to statistical tests for quantitative data (Kidder, 1981) and should be treated in the same way. That is, just as no one achieves statistical significance at the .000 level, so probably the qualitative data analyst ought not to expect that *all* cases would fit into appropriate categories. But when some reasonable number do, then negative case analysis provides confidence that the evaluator has tried and rejected all rival hypotheses save the appropriate one.

(5) *Progressive subjectivity:* The process of monitoring the evaluator's (or any inquirer's) own developing construction. It is obvious that no inquirer engages in an inquiry with a blank mind, a tabula rasa. It is precisely because the inquirer's mind is not blank that we find him or her engaged in the particular investigation. But it is equally obvious that any construction that emerges from an inquiry must, to be true constructivist principles, be a *joint* one. The inquirer's construction cannot be given privilege over that of anyone else (except insofar as he or she may be able to introduce a wider range of information and a higher level of sophistication than may any other single respondent). The technique of progressive subjectivism is designed to provide a check on the degree of privilege. And it is simple to execute. Prior to engaging i*any* activity at the site or in the context in which the investigation is to proceed, the inquirer records his or her a priori construction-what he or she expects to find once the study is under way-and archives that record. A most useful archivist is the debriefer, whom we have already discussed. At regular intervals throughout the study the inquirer *again* records his or her developing construction. If the inquirer affords too much privilege to the original constructions (or to earlier constructions as time progresses), it is safe to assume that he or she is not paying as much attention to the constructions offered by other participants as they deserve. The debriefer is in a sensitive position to note such a tendency and to challenge the inquirer about it. If the inquirer "finds" only what he or she expected to find, initially, or seems to become "stuck" or "frozen" on some intermediate construction, credibility suffers.

(6) ***Member checks:*** The process of testing hypotheses, data, preliminary categories, and interpretations with members of the stakeholding groups from whom the original constructions were collected. This is the single most crucial technique for establishing credibility. If the evaluator wants to establish that the multiple realities he or she presents are those that stakeholders have provided, the most certain test is verifying those multiple constructions with those who provided them. This process occurs continuously, both during the data collection and analysis stage, and, again, when (and if) a narrative case study is prepared. Member checks can be formal and informal, and with individuals (for instance, after interviews, in order to verify that what was written down is what was intended to be communicated) or with groups (for instance, as portions of the case study are written, members of stakeholding groups are asked to react to what has been presented as representing their construction).

Member checking serves a number of functions, including the following:

- It allows the evaluator to assess the intent of given action-what it is that a given respondent intended by acting in a certain way or by proffering certain information;
- it gives the respondent *(member* of stakeholding group) the chance to correct errors of fact or errors of interpretation;
- it provides interviewees (informants, respondents) the chance to offer additional information, especially by allowing them to "understand" the situation as a stranger understands it; this often stimulates a respondent to think about information, which further illuminates a given construction and can bring out information that might have been forgotten if the opportunity to review the interview had not occurred;
- it puts the respondent "on record" as having said certain things and as having agreed that the interviewer "got it right";
- it allows a chance for the inquirer to summarize, not only for the respondent but as a first step toward analysis of a given interview; and
- it gives the respondent a chance to judge overall adequacy of the interview itself in addition to providing the opportunity to confirm individual data items (Lincoln & Guba, 1985, p. 314).

Claims to adequacy of the overall inquiry most often are made by means of a formal member check, usually just prior to submitting a fmal agenda for negotiation of the case study (the purpose of which is to lay out the contextual particulars relevant to this negotiation). This member check session, involving knowledgeable and articulate individuals from each stakeholding group, has as its focus an inspection of the case study, the purpose of which is to correct errors of fact and/or interpretation. Of course, there are sometimes problems with the member-check process. We should note, however, that upon completion of five extensive case studies of five widely disparate sites across the United States-in the Special Education in Rural America Project (Skrtic et al., 1985)—the final formal member-check process failed to turn up a single suggestion for correction *of interpretation*. Several errors of fact were noted on several of the sites' case studies, but of the hundreds of persons interviewed, not one single person felt compelled to challenge the interpretations finally written into case study form. This is a powerful example of the kind of trust the hermeneutic process, carried out with integrity, can engender. No person, no matter how powerful or remote from power, at any site, felt that her or his construction had been misrepresented.

Sometimes, stakeholding groups brought together for the final member check may be adversarial. Some among the groups may feel that deliberate confrontation is in their best interests (much as the decades of the 1960s and 1970s brought forth a particular kind of social confrontation-called "mau-mauing"—with a definitive social end: recognition and funding for local minority groups). Member-checking processes can also be misleading, in the event that all members of a stakeholding group share a common myth, decide that they will maintain an organizational front, or even deliberately conspire to withhold information. If a conspiracy is afoot, being naive is no help, and nothing short of wide experience (and occasionally being "taken" by shrewd clients) can overcome naivete. But the process itself is an enormous help to avoiding conspiracies, because the openness of the process, and the free flow of information, serve to counteract the secrecy needed to maintain myths, fronts, and deliberate deception.

The reader who is familiar with our earlier work will notice that we have avoided a discussion of triangulation as a credibility check. In part, we have done so because triangulation itself carries too positivist an implication, to wit, that there exist unchanging phenomena so that triangulation can logically be a check. For those readers who have found the idea of triangulation a useful one, we would offer the following caveat: Member-checking processes ought to be dedicated to verifying that the *constructions* collected are those that have been offered by respondents, while triangulation should be thought of as referring to cross-checking specific data items of a factual nature (number of target persons served, number of children enrolled in a school-lunch program, number of handicapped elementary children in Foster School District who are in self-contained classrooms, number of fourth-grade mathematics textbooks purchased by the district for the 1987-1988 school year, number of high school English teachers employed, and the like).

Transferability. Transferability may be thought of as parallel to external validity or generalizability. Rigorously speaking, the positivist paradigm requires both sending and receiving contexts to be at least random samples from the same population. In the constructivist paradigm, external validity is replaced by an empirical process for checking the degree of similarity between sending and receiving contexts. Further, the burden of proof for claimed generalizability is on the *inquirer*, while the burden of proof for claimed transferability is on the *receiver*.

Generalization, in the conventional paradigm, is absolute, at least when conditions for randomization and sampling are met. But transferability is always relative and depends entirely on the degree to which salient conditions overlap or match. The major technique for establishing the degree of transferability is thick description, a term first attributed to anthropologist Gilbert Ryle and elaborated by Clifford Geertz (1973). Just what constitutes" 'proper' thick description is still not completely resolved" (Lincoln & Guba, 1985, p. 316). Furthermore, it may never be, since the conditions that need to obtain to declare transferability between Context A and Context B may in fact change with the nature of the inquiry; the "criteria that separate relevant from irrelevant descriptors are still largely undefined" (Lincoln 8 Guba, 1985, p. 316). But the constructivist works with very different types of "confidence limits"; the hypotheses relevant to naturalistic inquiries are only "working" and, therefore, are liable to disconfirmation or to assessments of nonutility, even in the same context, at a later period of time. The object of the game in making transferability judgments is to set out all the working hypotheses for *this* study, and to provide an extensive and careful description of the time, the place, the context, the culture in which those hypotheses were foundto be salient. The constructivist does not provide the confidence limits of the study. Rather, what he or she does is to provide as complete a data base as humanly possible in order to facilitate transferability judgments on the part of others who may wish to apply the study to their own situations (or situations in which they have an interest).

Dependability. Dependability is parallel to the conventional criterion of reliability, in that it is concerned with the stability of the data over time. Often such instability occurs because inquirers are bored, are exhausted, or are under considerable psychological stress from the intensity of the process. But dependability specifically excludes changes that occur because of overt methodological decisions by the evaluator or because of maturing reconstructions. In conventional inquiry, ofcourse, alterations in methodology (design) of the study would render reliability greatly suspect, if not totally meaningless (unstable). Likewise, shifts in hypotheses, constructs, and the like are thought to expose studies to unreliability.

But methodological changes and shifts in constructions are expected products of an emergent design dedicated to increasingly sophisticated constructions. Far from being threats to dependability, such changes and shifts are hallmarks of a maturing-and successful-inquiry. But such changes and shifts need to be both tracked and trackable (publicly inspectable), so that outside reviewers of such an evaluation can explore the process, judge the decisions that were made, and understand what salient factors in the context led the evaluator to the decisions and interpretations made. The technique for documenting the logic of process and method decisions is the dependability audit.

The inquiry audit is a procedure based on the metaphor (and actual process) of the fiscal audit. In afiscal audit, two kinds of issues are explored: First, to what extent is the*process* an established, trackable, and documentable process, and, second, to what extent are various data in the bookkeeping system actually confirmable? The dependability audit relies on the first, or process, judgment. The other half of the auditing process rests in the fourth trustworthiness criterion, confirmability. We shall discuss both of these forms of auditing together in thefollowing section.

Confirmability. Confirmability may be thought of as parallel to the conventional criterion of objectivity. Like objectivity, confirmability is concerned with assuring that data, interpretations, and outcomes of inquiries are rooted in contexts and persons apart from the evaluator and are not simply figments of the evaluator's imagination. Unlike the conventional paradigm, which roots its assurances of objectivity in *method*—that is, follow the process correctly and you will have findings that are divorced from the values, motives, biases, or political persuasions ofthe inquirer-the constructivist paradigm's assurances of integrityof the findings are rooted in the data themselves. This means that data (constructions, assertions, facts, and so on) can be tracked to their sources, and that the logic used to assemble the interpretations into structurally coherent and corrorborating wholes is both explicit and implicit in the narrative of a case study. Thus both the "raw products" and the "processes used to compress them," as Cronbach and Suppes (1969) put it, are available to be inspected and confirmed by outside reviewers of the study. The usual technique for confirming the data and interpretations of a given study is the confirmability audit.

This audit and the dependability audit alluded to above can be carried out together (and probably should be). As we mentioned earlier, the concept of an inquiry audit is rooted in the metaphor of the fiscal audit. The fiscal auditor is concerned with attesting to the quality and appropriateness ofthe *accounting process,* and in similar fashion the inquiry auditor is concerned with attesting to the quality and appropriateness of the *inquiry process.* That examination is effectively the *dependability audit.* But a fiscal auditor is also concerned with the "bottom line," that is, can entries in the accounts ledgers be verified and do the numbers add up right? In similar fashion the inquiry auditor attests to the fact that the "data" (facts, figures, and constructions) can all be traced to original sources, and the process by which they were converted to the "bottom line" ("compressed and rearranged to make the conclusions credible," to use Cronbach & Suppes's terminology) can be confirmed. This effectively is the confirmability audit to which we alluded. Algorithms for setting up an "audit trail" and for carrying out an inquiry audit are described in detail in Schwandt andHalpern (1988). An abbreviated discussion of the process can also be found in Lincoln and Guba (1985).

The Hermeneutic Process as Its Own Quality Control

Another way of judging the quality of evaluations conducted as fourth generational is to look within the process itself. Conventional evaluation, for example, is dependent virtually entirely on external, objective assessments of its quality for confirmation of goodness. But fourth generation evaluation is conducted via a hermeneutic, dialectic process. Data inputs are analyzed immediately on receipt. They may be "fed back" for comment, elaboration, correction, revision, expansion, or whatever to the very respondents who provided them only a moment ago. But those data inputs will also surely be incorporated into the emerging joint, collaborative reconstruction that emerges as the process continues. The opportunities for error to go undetected and/or unchallenged are very small in such a process. It is the immediate and continuing interplay of information that militates against the possibility of noncredible outcomes. It is difficult to maintain false fronts, or support deliberate deception when information is subject to continuous and multiple challenges from a variety of stakeholders. The publicly inspectable and inspected nature of the hermeneutic process itself prevents much of the kinds of secrecy and information poverty that have characterized client-focused evaluations of other generations.

Further, the possibility that the so-called biases or prejudices of the evaluator can shape the results is virtually zero, provided only that the evaluation is conducted in accordance with hermeneutic dialectic principles. (The argument that not all evaluators will "play the game" honorably and honestly is unconvincing. The same observation can be made of inquiry conducted within *any* paradigm, as recent experience so well attests.) So long as the evaluator's constructions (to which she or he is entitled as is any other constructor; calling them biases may have persuasive value but is hardly compelling) are laid on the table along with all the others and are made to withstand the same barrage of challenge, criticism, and counterexample as any others, there is no basis for according them any special influence, for better or worse.

The Authenticity Criteria

The above two approaches to the problem of criteria of goodness of fourth generation evaluations, while useful, are not entirely satisfying (either to us or to our critics). The first are, after all, *parallel* criteria. They have their roots and origins in positivist assumptions, and while adjustments have been made for the different assumptions of the naturalist paradigm, there remains a feeling of constraint, a feeling of continuing to play "in the friendly confines" of the opposition's home court.

In addition to their positivist ring, they share a second characteristic that leaves an uncomfortable feeling: they are primarily *methodological* criteria. That is, they speak to *methods* that can ensure one has carried out the process correctly. In the positivist paradigm, method has primacy. Method is critical for ensuring that the results are trustworthy. But method is only one consideration in constructivist inquiry or fourth generation evaluation. Outcome, product, and negotiation criteria are equally important in judging a given inquiry. Relying solely on criteria that speak to methods, as do the parallel criteria, leaves an inquiry vulnerable to questions regarding whether stakeholder rights were in fact honored. To put the point more bluntly, prolonged engagement and persistent observation (or any other *methods* one might choose) do not ensure that stakeholder constructions have been collected and faithfully represented. So reliance on pure or pristine method alone is insufficient to guarantee that the intent of the inquiry effort was achieved.

The second approach, while rooted in constructivism, suffers from being implicit to the process, and hence is not very persuasive to those who wish to see explicit evidence. We were moved as a result (and at the gentle critical prompting of our caring critic, John K. Smith) to devise what we have now called "authenticity criteria," which spring directly from constructivism's own basic assumptions. That is, they could have been invented by someone who had never heard of positivism or its claims for rigor. These criteria can be explicitly confirmed and would be addressed in any case study emerging from a constructivist evaluation. The authenticity criteria include the following (Lincoln & Guba, 1986a):

Fairness. Fairness refers to the extent to which different constructions and their underlying value structures are solicited and honored within the evaluation process. These different constructions must be presented, clarified, checked (as in the member-checking process), and taken into account in a balanced and evenhanded way. Since inquiry (and evaluations) are value-bound and value-situated, and evaluators inevitably confront a situation of value pluralism, then multiple constructions resting on differing value systems will emerge from stakeholders in and around the evaluation effort. The role of the evaluator is to seek out, and communicate, all such constructions and to explicate the ways in which such constructions-and their underlying value systems-are in conflict.

There are two techniques for achieving fairness. The first involves stakeholder identification and the solicitation of within-group constructions. The process of identifying all potential stakeholders and seeking out their constructions should become a part of the permanent audit trail completed for each evaluation case study. The presentation of constructions will be most clearly displayed in the identification of conflict over claims, concerns, and issues. Explicating the differences between belief and value systems is "not always an easy task, but exploration of values when clear conflict is evident should be a part of the data-gathering and data-analysis processes (especially during, for instance, the content analysis of individual interviews)" (Lincoln 8 Guba1986a, p. 79).

The second step in achieving this criterion is the open negotiation of recommendations and of the agenda for subsequent action. This process is especially visible in the methodological steps of prioritizing unresolved claims, concerns, and issues, collecting information relevant to them as well as adding a level of sophistication that may be required, preparing the agenda for negotiation, and carrying out the negotiation itself, as carried out by equally skilled bargainers, from approximately equal positions of power, and with the same (equal) information available to all. The open negotiation is modeled on labor negotiation and arbitration (and, indeed, our rules were devised from a study of that literature). Negotiations that are true to fourth generation evaluations have the following characteristics:

(1) They must be open, carried out in full view of the parties or the parties' representatives; closed sessions, secret codicils, or the like are not permitted.

(2) The negotiations must be carried out with equally skilled bargainers. While it is hardly ever the case that all stakeholders will be equally skilled bargainers, all sides shouldhave access to skilled bargainers. When it is necessary, the evaluator will act as adviser and educator to the less skilled. We are aware that this appears to be an advocacy role, which some will resist, but we have already argued earlier that the proper and appropriate province of the evaluator is the empowerment of previously disenfranchised stakeholders, so this does not breach the assumptions or goals of fourth generation evaluation.

(3) The negotiation must be carried out from approximately equal positions of power, not just in principle but also in practice.

(4) The negotiation must be carried out under circumstances where all parties are in possession of the same level of information; in some instances, this may mean that stakeholders may require assistance in understanding what the information means for their interests, but providing such assistance is also a legitimate role of the evaluator.

(5) The negotiation itself must focus on matters that are known to be relevant.

(6) Finally, the negotiation must be conducted in accordance with rules that the stakeholders themselves devised and to which all have assented.

Fairness also requires the creation of an appellate mechanism should any negotiating party feel that the rules are not observed. It also mandates fully informed consent with respect to any process that is part of the evaluation procedures. Consent is obtained not only prior to opening the evaluation effort but as information is uncovered and shared; as power relationships shift, this consent must be renegotiated continuously. And last, fairness requires the constant use of the member-check process, not only for the purpose of commenting on whether the constructions have been received "as sent" but for the purpose of commenting on the fairness process (adapted, Lincoln &Guba, 1986a).

Since discussions of fairness are fairly straightforward in other literatures, it is reasonably clear what this criterion might mean if achieved, and it is reasonably documentable when it has been achieved. The next criteria are more ambiguous, although, clearly, documentation as to their achievement needs to be provided.

Ontological authenticity. This criterion refers to the extent to which individual respondents' ownemic constructions are improved, matured, expanded, and elaborated, in that they now possess more information and have become more sophisticated in its use. It is, literally, "improvement in the individual's (or group's) conscious experiencing of the world" (Lincoln & Guba, 1986a, p. 81).

Ontological authenticity can be enhanced through the provision of vicarious experience, which enhances the opportunity for individual respondents (stakeholders and others) to apprehend their own "worlds" in more informed and sophisticated ways. Insofar as the evaluator can make available examples, cases, or other material that aids participants to reassess their own experience-seeing how it is the same as or different from the experience of others-it may serve to enhance their own awareness of the context in which theyfind themselves. While vicarious experience may not be enough, it is nevertheless a powerful tool for expanding respondents' own awareness or consciousness, particularly of structural aspects of a given context or community.

There are two techniques for demonstrating that the criterion of ontological authenticity has been achieved. First, there is the testimony of selected respondents. When individual stakeholders can attest to the fact that they now understand a broader range of issues, or that they can appreciate (understand, comprehend) issues that they previously failed to understand-that is evidence of ontological authenticity. Second, the audit trail for the case study should have entries of individual constructions recorded at different points in the evaluation process. Those entries ought to include those of the evaluator as well, in order to document "progressive subjectivity."

Educative authenticity. Educative authenticity represents the extent to which individual respondents' understanding of and appreciation for the constructions of *others* outside their stakeholding group are enhanced.

> It is not enough that the actors in some contexts achieve, individually, more sophisticated or mature constructions, or those that are more ontologically authentic. It is also*essential* that they come to appreciate (apprehend, discern, understand)-not necessarily like or agree with— the constructions that are made by others and to understand how those constructions are rootedin the different values systems of those others. (Lincoln 8Guba, 1986a, p. 81)

Stakeholders should at least have the opportunity to be confronted with the constructions of others very different from themselves, for, among other things, the chance to see how different value systems evoke very different solutions to issues surrounding the evaluand.

There are two techniques for establishing whether or not educative authenticity has been achieved. First, testimony of selected participants in the process will attest to the fact that they have comprehended and understood the constructions of others different from themselves. This testimony will often emerge in the negotiation process, and so will be not only documentable but publicly available. Second, at the end of the process, the audit trail should contain entries related to the developing understanding or appreciation as seen through exchanges during the hermeneutic circles process.

Catalytic authenticity. This criterion may be defined as the extent to which action is stimulated and facilitated by the evaluation processes. Reaching new and more sophisticated constructions, and achieving some appreciation of the positions of others, even achieved within a system of consummate fairness, is simply not enough. The purpose of evaluation is some form of action and/or decision making. Thus no fourth generation evaluation is complete without action being prompted on the part of participants.

Any number of clues lead us to observe that *action* is singularly lacking in most evaluations: the call for getting "theory into action"; the preoccupation in recent decades with "dissemination" at the national level; the creation and maintenance of federal laboratories, centers, and dissemination networks; the non-utilization of evaluations, and the general disenchantment with evaluation efforts at the federal level, together with the concomitant lowering of funding levels for such activity. This form of evaluation, with its heavy involvement of stakeholders, participants, and targets promises to stimulate action in a manner and at a level unheard of in the first three generations.

There are three techniques for assuring that this criterion has been met. First, there should be available testimony of participants from all stakeholding groups, including not only testimony of their interest in acting on the evaluation but their willingness to become involved in doing so. Second, we can rely on resolutions issuing from the negotiating sessions themselves. When action is jointly negotiated, it should follow that action is "owned" by participants and, therefore, as the research has shown, more willingly carried out. And third, there is, of course, systematic follow-up within some given time period to assess the extent of action and change revolving about the evaluation effort.

Tactical authenticity. It is not enough to be stimulated to action. It is quite possible to want, and even to need, to act, but to lack the power to do so in any meaningful way. Thus tactical authenticity refers to the degree to which stakeholders and participants are empowered to act. The first step in empowerment, of course, is taken when all stakeholders and others at risk are provided with the opportunity to contribute inputs to the evaluation and to have a hand in shaping its focus and its strategies. But this process of empowerment must be continued throughout the process for participants to be *fully* empowered to act at the consummation of the negotiation process.

There are three ways in which tactical authenticity may be demonstrated. First, testimony of selected participants and stakeholders from all groups is solicited. (It is clearly not enough simply to survey the clients and funders.) Second, some follow-up has to be undertaken in order to determine which groups do in fact participate and to examine the ways in which they participate. And, finally, some judgment can be rendered, usually by participants and evaluator alike, as to the degree of empowerment during the evaluation process itself Was it participatory? Have all stakeholders felt that they or their representatives have had a significant role in the process? Are all participants more skilled than previously in understanding and utilizing power and negotiation techniques? If the answers to those questions are uniformly yes, then tactical authenticity has probably been achieved.

Summary

It is apparent that there are many ways to assess the goodness of a fourth generation evaluation. It is not and need not be the case that such evaluations are sloppy, corner-cutting, or unmindful of standards. Quite the opposite. It ought to be evident that the most basic question is this: "*What* standards ought apply?" We have described several ways to respond to this question in this chapter, and have tried to indicate where the proposed standards come from and/or how they have been derived. Each set has utility for certain purposes. The trick is not to confuse the purposes. It is also important to keep in mind that goodness criteria, like paradigms, are rooted in certain assumptions. Thus it is not appropriate to judge constructivist evaluations by positivistic criteria or standards, or vice versa. To each its proper and appropriate set.

Note

1. The *Standards* devised by the Joint Committee in 1981 are intended to apply to evaluations of educational programs, projects, and materials. A reconstituted Joint Committee has recently published (JointComittee, 1988) a second set of standards that apply to the evaluation of educational personnel; these standards are not covered in the present discussion.