The emergence of a new paradigm of inquiry (naturalistic) has, unsurprisingly enough, led to a demand for rigorous criteria that meet traditional standards of inquiry. Two sets are suggested, one of which, the "trustworthiness" criteria, parallels conventional criteria, while the second, "authenticity" criteria, is implied directly by new paradigm assumptions.

But Is It Rigorous? Trustworthiness and Authenticity in Naturalistic Evaluation

Yvonna S. Lincoln,
Egon G. Guba

Until very recently, program evaluation has been conducted almost exclusively under the assumptions of the conventional, scientific inquiry paradigm using (ideally) experimentally based methodologies and methods. Under such assumptions, a central concern for evaluation, which has been considered a variant of research and therefore subject to the same rules, has been how to maintain maximum rigor while departing from laboratory control to work in the "real" world.

The real-world conditions of social action programs have led to increasing relaxation of the rules of rigor, even to the extent of devising studies looser than quasi-experiments. Threats to rigor thus abound in

We are indebted to Judy Meloy, graduate student at Indiana University, who scoured the literature for references to fairness and who developed a working paper on which many of our ideas depend.

D. D. Williams (Ed.). *Naturalistic Evaluation.*
New Directions for Program Evaluation, no. 30. San Francisco: Jossey-Bass, June 1986.

sections explaining how, when, and under what conditions the evaluation was conducted so that the extent of departure from desired levels of rigor might be judged. Maintaining true experimental or even quasi-experimental designs, meeting the requirements of internal and external validity, devising valid and reliable instrumentation, probabilistically and representatively selecting subjects and assigning them randomly to treatments, and other requirements of sound procedure have often been impossible to meet in the world of schools and social action. Design problems aside, the ethics of treatment given and treatment withheld poses formidable problems in a litigious society (Lincoln and Guba, 1985b).

Given the sheer technical difficulties of trying to maintain rigor and given the proliferation of evaluation reports that conclude with that ubiquitous finding, "no significant differences," is it not surprising that the demand for new evaluation forms has increased. What is surprising—for all the disappointment with experimental designs—is the *continued* demand that new models must demonstrate the ability to meet the same impossible criteria! Evaluators and clients both have placed on new-paradigm evaluation (Guba and Lincoln, 1981; Lincoln and Guba, 1985a) the expectation that naturalistic evaluations must be rigorous in the conventional sense, despite the fact that the basic paradigm undergirding the evaluation approach has shifted.

Under traditional standards for rigor (which have remained largely unmet in past evaluations), clients and program funders ask whether naturalistic evaluations are not so subjective that they cannot be trusted. They ask what roles values and multiple realities can legitimately play in evaluations and whether a different team of evaluators might not arrive at entirely different conclusions and recommendations, operating perhaps from a different set of values. Thus, the rigor question continues to plague evaluators and clients alike, and much space and energy is again consumed in the evaluation report explaining how different and distinct paradigms call forth different evaluative questions, different issues, and entirely separate and distinct criteria for determining the reliability and authenticity—as opposed to rigor—of findings and recommendations.

Rigor in the Conventional Sense

The criteria used to test rigor in the conventional, scientific paradigm are well known. They include exploring the truth value of the inquiry or evaluation (internal validity), its applicability (external validity or generalizability), its consistency (reliability or replicability), and its neutrality (objectivity). These four criteria, when fulfilled, obviate problems of confounding, atypicality, instability, and bias, respectively, and they do so, also respectively, by the techniques of controlling or randomizing possible sources of confounding, representative sampling, replication,

and insulation of the investigator (Guba, 1981; Lincoln and Guba, 1985a). In fact, to use a graceful old English cliché, the criteria are honored more in the breach than in the observance; evaluation is but a special and particularly public instance of the impossibility of fulfilling such methodological requirements.

Rigor in the Naturalistic Sense: Trustworthiness and Authenticity

Ontological, epistemological, and methodological differences between the conventional and naturalistic paradigms have been explicated elsewhere (Guba and Lincoln, 1981; Lincoln and Guba, 1985a; Lincoln and Guba, 1986; Guba and Lincoln, in press). Only a brief reminder about the axioms that undergird naturalistic and responsive evaluations is given here.

The axiom concerned with the nature of reality asserts that there is no single reality on which inquiry may converge, but rather there are multiple realities that are socially constructed, and that, when known more fully, tend to produce diverging inquiry. These multiple and constructed realities cannot be studied in pieces (as variables, for example), but only holistically, since the pieces are interrelated in such a way as to influence all other pieces. Moreover, the pieces are themselves sharply influenced by the nature of the immediate context.

The axiom concerned with the nature of "truth" statements demands that inquirers abandon the assumption that enduring, context-free truth statements—generalizations—can and should be sought. Rather, it asserts that all human behavior is time- and context-bound; this boundedness suggests that inquiry is incapable of producing nomothetic knowledge but instead only idiographic "working hypotheses" that relate to a given and specific context. Applications may be possible in other contexts, but they require a detailed comparison of the receiving contexts with the "thick description" it is the naturalistic inquirer's obligation to provide for the sending context.

The axiom concerned with the explanation of action asserts, contrary to the conventional assumption of causality, that action is explainable only in terms of multiple interacting factors, events, and processes that give shape to it and are part of it. The best an inquirer can do, naturalists assert, is to establish plausible inferences about the patterns and webs of such shaping in any given evaluation. Naturalists utilize the field study in part because it is the only way in which phenomena can be studied holistically and *in situ* in those natural contexts that shape them and are shaped by them.

The axiom concerned with the nature of the inquirer-respondent relationship rejects the notion that an inquirer can maintain an objective distance from the phenomena (including human behavior) being studied,

ability as an analog to reliability, and confirmability as an analog to objectivity. We shall refer to these criteria as criteria of trustworthiness (itself a parallel to the term *rigor*).

Techniques appropriate either to increase the probability that these criteria can be met or to actually test the extent to which they have been met have been reasonably well explicated, most recently in Lincoln and Guba (1985a). They include:

For credibility:

- Prolonged engagement—lengthy and intensive contact with the phenomena (or respondents) in the field to assess possible sources of distortion and especially to identify saliencies in the situation
- Persistent observation—in-depth pursuit of those elements found to be especially salient through prolonged engagement
- Triangulation (cross-checking) of data—by use of different sources, methods, and at times, different investigators
- Peer debriefing—exposing oneself to a disinterested professional peer to "keep the inquirer honest," assist in developing working hypotheses, develop and test the emerging design, and obtain emotional catharsis
- Negative case analysis—the active search for negative instances relating to developing insights and adjusting the latter continuously until no further negative instances are found; assumes an assiduous search
- Member checks—the process of continuous, informal testing of information by soliciting reactions of respondents to the investigator's reconstruction of what he or she has been told or otherwise found out and to the constructions offered by other respondents or sources, and a terminal, formal testing of the final case report with a representative sample of stakeholders.

For transferability:

- Thick descriptive data—narrative developed about the context so that judgments about the degree of fit or similarity may be made by others who may wish to apply all or part of the findings elsewhere (although it is by no means clear how "thick" a thick description needs to be, as Hamilton, personal communication, 1984, has pointed out).

For dependability and confirmability:

- An external audit requiring both the establishment of an audit trail and the carrying out of an audit by a competent external, disinterested auditor (the process is described in detail in Lincoln and Guba, 1985a). That part of the audit that examines the process results in a dependability judgment, while that part concerned with the product (data and reconstructions) results in a confirmability judgment.

While much remains to be learned about the feasibility and utility of these parallel criteria, there can be little doubt that they represent a substantial advance in thinking about the rigor issue. Nevertheless, there are some major difficulties with them that call out for their augmentation with new criteria rooted in naturalism rather than simply paralleling those rooted in positivism.

First, the parallel criteria cannot be thought of as a complete set because they deal only with issues that loom important from a positivist construction. The positivist paradigm ignores or fails to take into account precisely those problems that have most plagued evaluation practice since the mid 1960s: multiple value structures, social pluralism, conflict rather than consensus, accountability demands, and the like. Indeed, the conventional criteria refer only to methodology and ignore the influence of context. They are able to do so because by definition conventional inquiry is objective and value-free.

Second, intuitively one suspects that if the positivist paradigm did not exist, other criteria might nevertheless be generated directly from naturalist assumptions. The philosophical and technical problem might be phrased thus: Given a relativist ontology and an interactive, value-bounded epistemology, what might be the nature of the criteria that ought to characterize a naturalistic inquiry? If we reserve the term *rigor* to refer to positivism's criteria and the term *reliability* to refer to naturalism's parallel criteria, we propose the term *authenticity* to refer to these new, embedded, intrinsic naturalistic criteria.

Unique Criteria of Authenticity. We must at once disclaim having solved this problem. What follows are simply some strong suggestions that appear to be worth following up at this time. One of us (Guba, 1981) referred to the earlier attempt to devise reliability criteria as "primitive"; the present attempt is perhaps even more aboriginal. Neither have we as yet been able to generate distinct techniques to test a given study for adherence to these criteria. The reader should therefore regard our discussion as speculative and, we hope, heuristic. We have been able to develop our ideas of the first criterion, fairness, in more detail than the other four; its longer discussion ought not to be understood as meaning, however, that fairness is very much more important than the others.

Fairness. If inquiry is value-bound, and if evaluators confront a situation of value-pluralism, it must be the case that different constructions will emerge from persons and groups with differing value systems. The task of the evaluation team is to expose and explicate these several, possibly conflicting, constructions and value structures (and of course, the evaluators themselves operate from some value framework).

Given all these differing constructions, and the conflicts that will almost certainly be generated from them by virtue of their being rooted in value differences, what can an evaluator do to ensure that they are pre-

sented, clarified, and honored in a balanced, even-handed way, a way that the several parties would agree is balanced and even-handed? How do evaluators go about their tasks in such a way that can, while not guaranteeing balance (since nothing can), at least enhance the probability that balance will be well approximated?

If every evaluation or inquiry serves some social agenda (and it invariably does), how can one conduct an evaluation to avoid, at least probabilistically, the possibility that certain values will be diminished (and their holders exploited) while others will be enhanced (and their holders advantaged)? The problem is that of trying to avoid empowering at the expense of impoverishing; all stakeholders should be empowered in some fashion at the conclusion of an evaluation, and all ideologies should have an equal chance of expression in the process of negotiating recommendations.

Fairness may be defined as a balanced view that presents all constructions and the values that undergird them. Achieving fairness may be accomplished by means of a two-part process. The first step in the provision of fairness or justice is the ascertaining and presentation of different value and belief systems represented by conflict over issues. Determination of the actual belief system that undergirds a position on any given issue is not always an easy task, but exploration of values when clear conflict is evident should be part of the data-gathering and data-analysis processes (especially during, for instance, the content analysis of individual interviews).

The second step in achieving the fairness criterion is the negotiation of recommendations and subsequent action, carried out with stakeholding groups or their representatives at the conclusion of the data-gathering, analysis, and interpretation stage of evaluation effort. These three stages are in any event simultaneous and interactive within the naturalistic paradigm. Negotiation has as its basis constant collaboration in the evaluative effort by all stakeholders; this involvement is continuous, fully informed (in the consensual sense), and operates between true peers. The agenda for this negotiation (the logical and inescapable conclusion of a true collaborative evaluation process), having been determined and bounded by all stakeholding groups, must be deliberated and resolved according to rules of fairness. Among the rules that can be specified, the following seem to be absolute minimum.

1. Negotiations must have the following characteristics:
   a. It must be open, that is, carried out in full view of the parties or their representatives with no closed sessions, secret codicils, or the like permitted.
   b. It must be carried out by equally skilled bargainers. In the real world it will almost always be the case that one or another group of bargainers will be the more skillful, but at

least each side must have access to bargainers of equal skill, whether they choose to use them or not. In some instances, the evaluator may have to act not only as mediator but as educator of those less skilled bargaining parties, offering additional advice and counsel that enhances their understanding of broader issues in the process of negotiation. We are aware that this comes close to an advocacy role, but we have already presumed that one task of the evaluator is to empower previously impoverished bargainers; this role should probably not cease at the negotiation stage of the evaluation.

c. It must be carried out from equal positions of power. The power must be equal not only in principle but also in practice; the power to sue a large corporation in principle is very different from the power to sue it in practice, given the great disparity of resources, risk, and other factors, including, of course, more skillful and resource-heavy bargainers.
d. It must be carried out under circumstances that allow all sides to possess equally complete information. There is no such animal, of course, as "complete information," but each side should have the same information, together with assistance as needed to be able to come to an equal understanding of it. Low levels of understanding are tantamount to lack of information.
e. It must focus on all matters known to be relevant.
f. It must be carried out in accordance with rules that were themselves the product of a pre-negotiation.

2. Fairness requires the availability of appellate mechanisms should one or another party believe that the rules are not being observed by some. These mechanisms are another of the products of the pre-negotiation process.
3. Fairness requires fully informed consent with respect to any evaluation procedures (see Lincoln and Guba, 1985a, and Lincoln and Guba, 1985b). This consent is obtained not only prior to an evaluation effort but is continually renegotiated and reaffirmed (formally with consent forms and informally through the establishment and maintenance of trust and integrity between parties to the evaluation) as the design unfolds, new data are found, new constructions are made, and new contingencies are faced by all parties.
4. Finally, fairness requires the constant use of the member-check process, defined earlier, which includes calls for comments on fairness, and which is utilized both during and after the inquiry process itself (in the data collection-analysis-construction stage and later when case studies are being developed). Vigilant and

assiduous use of member-checking should build confidence in individuals and groups and should lead to a pervasive judgment about the extent to which fairness exists.

Fairness as a criterion of adequacy for naturalistic evaluation is less ambiguous than the following four, and more is known about how to achieve it. It is not that this criterion is more easily achieved, merely that it has received more attention from a number of scholars (House, 1976; Lehne, 1978; Strike, 1982, see also Guba and Lincoln, 1985).

Ontological Authentication. If each person's reality is constructed and reconstructed as that person gains experience, interacts with others, and deals with the consequences of various personal actions and beliefs, an appropriate criterion to apply is that of improvement in the individual's (and group's) conscious experiencing of the world. What have sometimes been termed *false consciousness* (a neo-Marxian term) and *divided consciousnes* are part and parcel of this concept. The aim of some forms of disciplined inquiry, including evaluation (Lincoln and Guba, 1985b) ought to be to raise consciousness, or to unite divided consciousness, likely via some dialectical process, so that a person or persons (not to exclude the evaluator) can achieve a more sophisticated and enriched construction. In some instances, this aim will entail the realization (the "making real") of contextual shaping that has had the effect of political, cultural, or social impoverishment; in others, it will simply mean the increased appreciation of some set of complexities previously not appreciated at all, or appreciated only poorly.

Educative Authentication. 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 rooted in the different value systems of those others. In this process, it is not inconceivable that accommodations, whether political, strategic, value-based, or even just pragmatic, can be forged. But whether or not that happens is not at issue here; what the criterion of educative validity implies is increased understanding of (including possibly a sharing, or sympathy with) the whats and whys of various expressed constructions. Each stakeholder in the situation should have the opportunity to become educated about others of different persuasions (values and constructions), and hence to appreciate how different opinions, judgments, and actions are evoked. And among those stakeholders will be the evaluator, not only in the sense that he or she will emerge with "findings," recommendations, and an agenda for negotiation that are professionally interesting and fair but also that he or she will develop a more sophisticated and complex construction (an emic-etic blending) of both personal and professional (disciplinary-substantive) kinds.

How one knows whether or not educative authenticity has been reached by stakeholders is unclear. Indeed, in large-scale, multisite evaluations, it may not be possible for all—or even for more than a few—stakeholders to achieve more sophisticated constructions. But the techniques for ensuring that stakeholders do so even in small-scale evaluations are as yet undeveloped. At a minimum, however, the evaluator's responsibility ought to extend to ensuring that those persons who have been identified during the course of the evaluation as gatekeepers to various constituencies and stakeholding audiences ought to have the opportunity to be "educated" in the variety of perspectives and value systems that exist in a given context.

By virtue of the gatekeeping roles that they already occupy, gatekeepers have influence and access to members of stakeholding audiences. As such, they can act to increase the sophistication of their respective constituencies. The evaluator ought at least to make certain that those from whom he or she originally sought entrance are offered the chance to enhance their own understandings of the groups they represent. Various avenues for reporting (slide shows, filmstrips, oral narratives, and the like) should be explored for their profitability in increasing the consciousness of stakeholders, but at a minimum the stakeholders' representatives and gatekeepers should be involved in the educative process.

Catalytic Authentication. Reaching new constructions, achieving understandings that are enriching, and achieving fairness are still not enough. Inquiry, and evaluations in particular, must also facilitate and stimulate action. This form of authentication is sometimes known as feedback-action validity. It is a criterion that might be applied to conventional inquiries and evaluations as well; although if it were virtually all positivist social action, inquiries and evaluations would fail on it. 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; the notable inaction subsequent to evaluations that is virtually a national scandal—all indicate that catalytic authentication has been singularly lacking. The naturalistic posture that involves all stakeholders from the start, that honors their inputs, that provides them with decision-making power in guiding the evaluation, that attempts to empower the powerless and give voice to the speechless, and that results in a collaborative effort holds more promise for eliminating such hoary distinctions as basic versus applied and theory versus practice.

Tactical Authenticity. Stimulation to action via catalytic authentication is in itself no assurance that the action taken will be effective, that is, will result in a desired change (or any change at all). The evaluation of inquiry requires other attributes to serve this latter goal. Chief among these is the matter of whether the evaluation is empowering or impoverishing, and to whom. The first step toward empowerment is taken by providing

all persons at risk or with something at stake in the evaluation with the opportunity to control it as well (to move toward creating collaborative negotiation). It provides practice in the use of that power through the negotiation of construction, which is joint emic-etic elaboration. It goes without saying that if respondents are seen simply as "subjects" who must be "manipulated," channeled through "treatments," or even deceived in the interest of some higher "good" or "objective" truth, an evaluation or inquiry cannot possibly have tactical authenticity. Such a posture could only be justified from the bedrock of a realist ontology and an "objective," value-free epistemology.

Summary

All five of these authenticity criteria clearly require more detailed explication. Strategies or techniques for meeting and ensuring them largely remain to be devised. Nevertheless, they represent an attempt to meet a number of criticisms and problems associated with evaluation in general and naturalistic evaluation in particular. First, they address issues that have pervaded evaluation for two decades. As attempts to meet these enduring problems, they appear to be as useful as anything that has heretofore been suggested (in any formal or public sense).

Second, they are responsive to the demand that naturalistic inquiry or evaluation not rely simply on parallel technical criteria for ensuring reliability. While the set of additional authenticity criteria might not be the complete set, it does represent what might grow from naturalistic inquiry were one to ignore (or pretend not to know about) criteria based on the conventional paradigm. In that sense, authenticity criteria are part of an inductive, grounded, and creative process that springs from immersion with naturalistic ontology, epistemology, and methodology (and the concomitant attempts to put those axioms and procedures into practice).

Third, and finally, the criteria are suggestive of the ways in which new criteria might be developed; that is, they are addressed largely to ethical and ideological problems, problems that increasingly concern those involved in social action and in the schooling process. In that sense, they are confluent with an increasing awareness of the ideology-boundedness of public life and the enculturation processes that serve to empower some social groups and classes and to impoverish others. Thus, while at first appearing to be radical, they are nevertheless becoming mainstream. An invitation to join the fray is most cheerfully extended to all comers.

References

Guba, E. G. "Criteria for Assessing the Trustworthiness of Naturalistic Inquiries." *Educational Communication and Technology Journal,* 1981, *29,* 75–91.

Guba, E. G., and Lincoln, Y. S. "Do Inquiry Paradigms Imply Inquiry Methodologies?" In D. L. Fetterman (Ed.), *The Silent Scientific Revolution.* Beverly Hills, Calif.: Sage, in press.

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

Guba, E. G., and Lincoln, Y. S. "The Countenances of Fourth Generation Evaluation: Description, Judgment, and Negotiation." Paper presented at Evaluation Network annual meeting, Toronto, Canada, 1985.

House, E. R. "Justice in Evaluation." In G. V. Glass (Ed.), *Evaluation Studies Review Annual, no. 1.* Beverly Hills, Calif.: Sage, 1976.

Lehne, R. *The Quest for Justice: The Politics of School Finance Reform.* New York: Longman, 1978.

Lincoln, Y. S., and Guba, E. G. *Naturalistic Inquiry.* Beverly Hills, Calif.: Sage, 1985a.

Lincoln, Y. S., and Guba, E. G. "Ethics and Naturalistic Inquiry." Unpublished manuscript, University of Kansas, 1985b.

Morgan, G. *Beyond Method: Strategies for Social Research.* Beverly Hills, Calif.: Sage, 1983.

Strike, K. *Educational Policy and the Just Society.* Champaign: University of Illinois Press, 1982.

Yvonna S. Lincoln is associate professor of higher education in the Educational Policy and Administration Department, School of Education, the University of Kansas. Egon G. Guba is professor of educational inquiry methodology in the Department of Counseling and Educational Psychology, School of Education, Indiana University. They have jointly authored two books, Effective Evaluation *and* Naturalistic Inquiry, *which sketch the assumptional basis for naturalistic inquiry and its application to the evaluation arena. They have also collaborated with others on a third book,* Organizational Theory and Inquiry, *Sage, 1985.*