DEVELOPING an EVALUATION CULTURE for INTERNATIONAL AGRICULTURAL RESEARCH

William M. K. Trochim
Cornell University

I am pleased to be here today to explore with you some of the major issues in the assessment of agricultural research impacts from the point of view of a professional evaluator. I thought I would concentrate on three general topics for this presentation.

First, I will give a brief introduction to the evaluation field as a whole and an overview of its principles and methodologies, so that we can share an understanding of the major terminology and issues in evaluation.

Second, several specific evaluation methodologies that might be useful in international agricultural research are introduced, along with discussion of potential applications. The first of these -- concept mapping -- is probably most useful for what is commonly referred to as "ex-ante" evaluation. This method was used with a small group of Cornell agricultural researchers in order to prepare this presentation and to provide an example more directly relevant to the agricultural context and the topic of this conference. The second method -- the regression-discontinuity design -- is a quasi-experimental approach useful in ex-post causal hypothesis testing when random assignments to condition is not feasible.

Third, and most important, I would like to ask each of you to join me in the ongoing development of an idealistic, utopian frame-of-mind that I call the "evaluation culture." This evaluation culture is a way of looking at the world — a broad-based, forward-looking point of view that sees the crucial evaluation feedback function as an integral part of our everyday life. The development of such a

culture is, I think, essential to the successful use of evaluation in the international agricultural research system as it is elsewhere. And successful evaluation and impact assessment are, I believe, the work-a-day methodological stuff that sustainable development will be made of. I will describe how this utopian evaluation culture might feel, what values it might foster, and how it will need to be adapted for the international agricultural research community.

Some Principles and Methodologies of Evaluation

I'd like to begin with a whirlwind tour of evaluation. The highlights of the tour will include:

1) a few definitions of evaluation;
2) the goals of evaluation work;
3) the planning-evaluation cycle, to show how evaluation is intimately connected with other fields and activities;
4) the most common evaluation strategies;
5) the major types of evaluation;
6) how evaluators approach the idea of evaluation standards and quality; and
7) the major evaluation questions and methods used to address them.

Definitions of evaluation

Let us begin by exploring what evaluation means. Probably the most frequently given definition of evaluation is that "*evaluation is the systematic assessment of the worth or merit of some object*" (Joint Committee, 1981).

I actually do not care much for this definition. There are many types of evaluations that do not necessarily result in an assessment of worth or merit — descriptive studies, implementation analyses and formative evaluations, to name a few. I prefer a definition that emphasizes the information-processing and feedback functions of evaluation. For instance, I prefer to say that "*evaluation is the systematic acquisition and assessment of information to provide useful feedback about some object.*"

Both definitions agree that evaluation is a systematic endeavor and both use the deliberately ambiguous term "object" which could be a program, policy, technology, person, need, activity, and so on. My definition emphasizes "acquiring and assessing *information*" rather than "assessing worth or merit" because all evaluation work involves collecting and sifting through data, and making judgments about the validity of the information and of inferences we derive from it.

The goals of evaluation

The generic goal of most evaluations is to provide "useful feedback" to a variety of audiences including sponsors, donors, client-groups, administrators, staff, and other relevant constituencies. Most often, feedback is perceived as "useful" if it aids in decision-making. But we have learned that the relationship between an evaluation and its impact is not a simple one — studies that seem critical sometimes fail to influence short-term decisions, and studies that initially seem to have no influence can have a delayed impact when more congenial conditions arise. Despite this, there is broad consensus that the major goal of evaluation should be to influence decision-making or policy formulation through the provision of empirically driven feedback.

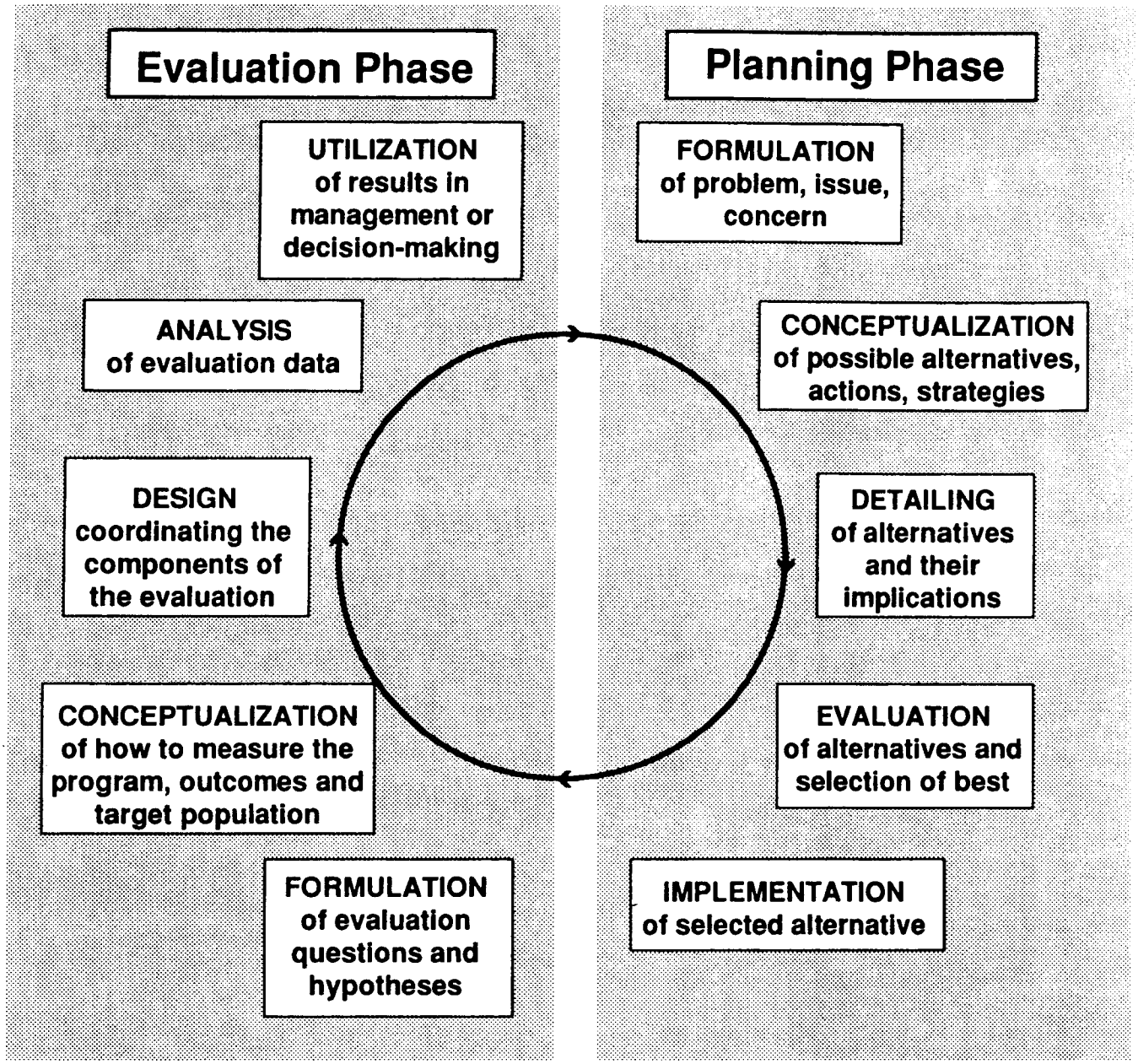
The planning-evaluation cycle

Often, evaluation is construed as part of a larger managerial or administrative process. Sometimes we refer to this as the *planning-evaluation cycle* (Figure 1). The distinctions between planning and evaluation are not always clear; this cycle is described in many different ways with various phases claimed by both planners and evaluators. Usually, the first stage of such a cycle — the *planning* phase — is designed to elaborate a set of potential actions, programs, or technologies, and select the best for implementation. Depending on the organization and the problem being addressed, a planning process could involve any or all of these stages (Nutt 1984):

- *formulation* of the problem, issue, or concern;
- the broad *conceptualization* of the major alternatives that might be considered;
- *detailing* of these alternatives and their potential implications;
- *evaluation* of the alternatives and selection of the best one; and
- *implementation* of the selected alternative.

Although these stages are traditionally considered **planning**, there is a lot of **evaluation** work involved. Evaluators are trained in needs assessment, they use methodologies — like the concept mapping one I will show later — that help in conceptualization and detailing, and they have the skills to help assess alternatives and choose the best one.

Figure 1. The planning-evaluation cycle.



The evaluation phase also involves a sequence of stages that typically includes:

- the *formulation* of the major objectives, goals, and hypotheses of the program or technology;
- the *conceptualization* and operationalization of the major components of the evaluation program, participants, setting, and measures;
- the *design* of the evaluation, detailing how these components will be coordinated;
- the *analysis* of the information, both qualitative and quantitative; and
- the *utilization* of evaluation results.

Evaluation strategies

"Evaluation strategies" mean, to me, broad, overarching perspectives on evaluation. They encompass the most general groups or "camps" of evaluators, although at its best, evaluation work borrows from the perspectives of all these camps. *Scientific-experimental models* are probably the most historically dominant evaluation strategies. Taking their values and methods from the sciences — especially the social sciences — they emphasize the desirability of impartiality, accuracy, objectivity and the validity of the information generated. Included under scientific-experimental models would be: the tradition of experimental and

quasi-experimental designs (Campbell, 1969; Cook and Campbell, 1979; Suchman, 1967); objectives-based research that comes to us from education (Tyler, 1989); econometrically-oriented perspectives including cost-effectiveness and cost-benefit analysis; and the recent articulation of theory-driven evaluation (Chen, 1990).

The second class of strategies are *management-oriented systems models*. You have undoubtedly heard of two of the most common of these: PERT, the Program Evaluation and Review Technique, and CPM, the Critical Path Method. Both have been widely used in business and government in this country. I would also include the Logical Framework or "Logframe" model developed at USAID (Binnendijk 1989) and general systems theory and operations research approaches. Two models in this category were originated by evaluators: the UTOS model (Cronbach 1982) where U stands for Units, T for Treatments, O for Observing Observations and S for Settings; and the CIPP model (Stufflebeam 1985) where the C stands for Context, the I for Input, the first P for Process and the second one for Product. These management-oriented systems models emphasize comprehensiveness in evaluation, placing evaluation within a larger framework of organizational activities.

The third class of strategies are the *qualitative/anthropological models*. They emphasize the importance of observation, the need to retain the phenomenological quality of the evaluation context, and the value of subjective human interpretation in the evaluation process. Included in this category are the approaches known in evaluation as "naturalistic" (Lincoln and Guba 1985) or "Fourth Generation" evaluation (Guba and Lincoln 1989); the various qualitative schools (Miles and Huberman 1984; Patton 1990, Fetterman 1988); critical theory and art criticism approaches; and, the "grounded theory" approach (Strauss 1967), among others.

Finally, a fourth class of strategies is termed *participant-oriented models*. As the term suggests, they emphasize the central importance of the evaluation participants, especially clients and users of the program or technology. These are the models Norman Uphoff has recommended (Uphoff 1992). Client-centered and stakeholder strategies (Stake 1976) are examples of participant-oriented models, as are consumer-oriented evaluation approaches (Scriven, 1981).

With all of these strategies to choose from, how to decide? Debates that rage within the evaluation profession (and they **do** rage, take my word for it) are generally battles between these different strategies, with each claiming the superiority of its position. In reality, most good evaluators are familiar with all four categories and borrow eclectically from each as the need arises. There is no inherent incompatibility between these broad strategies; each of them brings something valuable to the evaluation table. In fact, in recent years attention has increasingly turned to how we can integrate results from evaluations that use different strategies, that are carried out from different perspectives, and that use different methods (Binnendijk 1989). Clearly, there are no simple answers here. A word of caution is good for us; in international agricultural research, Anderson and Herdt (1990) and Horton (1990), among others, point out that the problems are complex and the methodologies needed should therefore be varied.

Types of evaluation

There are many different types of evaluations depending on the object being evaluated and the purpose of the evaluation. Perhaps the most important basic distinction in evaluation types is that between *formative* and *summative* evaluation (Scriven 1967).

Formative evaluations strengthen or improve the object being evaluated — they help *form* it. How? By examining the delivery of the program or technology, the quality of its implementation, the assessment of the organizational context, personnel, procedures, inputs, and so on. Summative evaluations, in contrast, examine the effects or outcomes of some object — they *summarize* it, so to speak. How? By describing what happens subsequent to delivery of the program or technology; assessing whether the object can be said to have **caused** the outcome; determining the overall impact of the causal factor beyond the immediate target outcomes; and estimating the relative costs associated with the object. The formative-summative distinction can be further refined.

Formative evaluation includes several evaluation types:

- *needs assessment* determines who needs the program, how great the need is, and what might work to meet the need;
- *evaluability assessment* determines whether an evaluation is feasible and how stakeholders can help shape its usefulness;
- *structured conceptualization* helps stakeholders define the program or technology, the target population, and the possible outcomes;

- *implementation evaluation* monitors fidelity of the program or technology delivery; and
- *process evaluation* investigates the process of delivering the program or technology, including alternative delivery procedures.

Summative evaluation can also be subdivided:

- *outcome evaluations* investigate whether the program or technology caused demonstrable effects on specifically defined target outcomes;
- *impact evaluation* is broader and assesses the overall or net effects — intended or unintended — of the program or technology as a whole;
- *cost-effectiveness* and *cost-benefit analysis* address questions of efficiency by standardizing outcomes in terms of their dollar costs and values;
- *secondary analysis* reexamines existing data to address new questions or use methods not previously employed;
- *meta-analysis* integrates the outcome estimates from multiple studies to arrive at an overall or summary judgment on an evaluation question.

Evaluation standards and quality

The evaluation profession is attempting to address the issues of standards and quality of evaluations in several ways. I will describe three of them here.

First, there have been several attempts to delineate a formal set of professional standards (Joint Committee 1981; Rossi 1982) much like those in the legal, auditing or accounting professions. These standards specify the roles, responsibilities and obligations of evaluators and the parties with whom they interact. The American Evaluation Association has not yet adopted any formal set of standards, but many evaluators use these two sets in an informal or advisory manner. It is unlikely that we will see formal evaluation standards adopted in the near future.

Another way to encourage standards and quality in evaluation consists of organized efforts, usually by the federal government, to review large numbers of evaluations done in a particular field. One of the most noteworthy of these studies was the Congressionally-funded review of educational evaluations that Bob Boruch led in the late 1970s. Given his key role in that effort, I will defer to Bob on this topic and hope that he will give us a thumbnail sketch of that work and its relevance to this workshop. He is also in a better position to describe the efforts of other national reviews of evaluation such as work done under the auspices of the National Science Foundation or the National Academy of Sciences among others.

The third approach to evaluation standards and quality, and undoubtedly the largest, consists of the ongoing debates in the evaluation literature about how best to do our work. I will not review this area here, but I would like to present briefly one framework for judging quality that is widely recognized and utilized in evaluation research, the *theory of validity* articulated by Campbell and Stanley (1963) and later revised by Cook and Campbell (1970).

The term *validity* here refers to the best available approximation to the truth or falsity of a given inference, proposition, or conclusion. We subdivide validity into four types: *conclusion validity* addresses the validity of statements about the presence or absence of observed relationships; *internal validity* refers to inferences about observed **causal** relationships between program and outcome; *construct validity* addresses whether the observed components in an evaluation — such as the program or outcomes — were implemented or operationalized as **intended**; and *external validity* refers to the degree to which an inference or conclusion can be generalized to other persons, places, technologies, times, and so on. For any inference or conclusion, there are always various "threats to validity" — reasons we might be wrong in that conclusion.

Imagine that we wish to examine whether there is a **relationship** between the amount of training in a technology and subsequent rates of adoption. We are interested in a **relationship**, an issue of **conclusion** validity. We complete our study and find no significant correlation between the amount of training and adoption rates. On this basis we conclude that there is no relationship between the two. How could we be wrong in this conclusion — what are the "threats to validity"? Well, it is possible that we do not have sufficient statistical power to detect a relationship even if it exists. Perhaps the sample size is too small or the measure of amount of training is unreliable. Or maybe we have violated the assumptions of the correlational test with the variables we used. Perhaps there were random irrelevancies in the study setting or random heterogeneity in the respondents that increased the variability in the data and made it harder to see the relationship of interest. Our inference that there is no relationship will be stronger — have greater validity — if we can show that these alternative explanations are not credible. We might examine the distribu-

tions to see if they conform with assumptions of the statistical test, or conduct an analysis to determine whether we have sufficient statistical power.

In a similar manner, we might analyze other potential inferences that we might wish to make from an evaluation study, determining which type of validity is most relevant and what major threats to the validity of the inferences are judged most plausible. The theory of validity and the consideration of specific threats provide a useful scheme for assessing the quality of our evaluation conclusions. This theory is general in scope and applicability, well-articulated in its philosophical suppositions, and virtually impossible to explain adequately in a few minutes. As a framework for judging the quality of evaluations, it is indispensable and well worth your investigation.

Evaluation questions and methods

Let us finally look at the kinds of questions evaluators face and the methods they can use in addressing these questions. To do so, we will return to the simple distinction between formative and summative evaluation.

In formative research, the major questions and methodologies are:

What is the definition and scope of the problem or issue, or what is the question? Here we might use formulating and conceptualizing methods such as brainstorming, focus groups, nominal group techniques, delphi methods, brainwriting, stakeholder analysis, synectics, lateral thinking, input-output analysis, and concept mapping.

Where is the problem and how big or serious is it? The most common method used here is "needs assessment" which can include: analysis of existing data sources and the use of sample surveys, interviews of constituent populations, qualitative research, expert testimony, and focus groups.

How should the program or technology be delivered to address the problem? Some of the methods already listed apply here, as do detailing methodologies like simulation techniques, or multivariate methods like multiattribute utility theory or exploratory causal modeling, decision-making methods, and project planning and implementation methods like flow charting, PERT/CPM, and project scheduling.

How well is the program or technology delivered? Qualitative and quantitative monitoring techniques, the use of management information systems, and implementation assessment would be appropriate methodologies here.

The questions and methods addressed under summative evaluation include:

What type of evaluation is feasible? We could use evaluability assessment here, as well as standard approaches for selecting an appropriate evaluation design.

What was the effectiveness of the program or technology? We would choose from observational and correlational methods for demonstrating whether desired effects occurred, and quasi-experimental and experimental designs for determining whether observed effects can reasonably be attributed to the intervention and not to other sources.

What is the net impact of the program? Econometric methods for assessing cost-effectiveness and cost/benefits would apply here, along with qualitative methods that enable us to summarize the full range of intended and unintended impacts.

This gives you some idea of what methods might be matched to which questions. All of these methods, and the many not mentioned, are supported by an extensive methodological research literature. Clearly, we already have a formidable set of tools at our disposal. But the need to improve, update and adapt these methods to changing circumstances means that methodological research and development needs to have a major place in evaluation work.

This concludes the whirlwind tour of the evaluation field.

Linking Evaluation and International Agricultural Research

Now to the topic of interest in this symposium, the relationship of evaluation and impact assessment to international agricultural research, especially to the international agricultural research system, specifically the CGIAR system. Clearly, it would be impossible in this presentation to attempt a comprehensive summary of evaluation methods. Rather, two specific methods -- one primarily used for planning and formative work (ex ante research)

and one for more summative or outcome evaluation (ex-post research) -- will be presented in some detail to illustrate some of the possibilities that exist. These methods were chosen because they are both relatively new and likely to be unknown outside of evaluation circles, have great potential for international agricultural research, and have been associated with my career (largely because I have played a key role in developing them). A more thorough review of existent methods will have to be deferred to another occasion.

Concept mapping for ex-ante evaluation

What I will do here is to practice a bit of what I preach and, instead of just talking about how evaluation might be used in the CGIAR system, show you an example of a method that might be useful to you in your ex ante evaluation work and apply that method to the issue at hand. With several colleagues I used this method to explore the role of evaluation in the assessment of the impact of international agricultural research. I present this example to you partly to illustrate a method that I think you will find useful in research planning and evaluation, and partly because the results of our exercise might shed some light on the central topic of this symposium.

The method I want to show you I call "concept mapping" or, more formally, "structured conceptualization" (Trochim 1989). It is designed to help a group lay out a pictorial conceptual framework that can be used in strategic planning, evaluation measurement construction, and many other tasks. At the center of this method — the engine, if you will — are several powerful multivariate statistical techniques — multidimensional scaling and hierarchical cluster analysis. But to the participant, the process appears similar to many other group conceptualizing, brainstorming and decision-making approaches.

I conducted a small concept mapping exercise over the past few weeks with a group of willing volunteers -- some faculty and some graduate students at Cornell. As I explain how the process works, I will illustrate with that example.

Developing the concept map

Concept mapping usually proceeds through a series of six steps. The first involves preparation for the process, including determining participants and agreeing on the focus for the mapping. Next, the participants generate a large set of ideas relative to the focus, usually through a simple brainstorming process. They structure these ideas in two ways: through sorting the ideas into groups of similar ones, and rating each idea on one or more dimensions of interest. In strategic planning, they might rate how important each idea is, while in evaluation, they might rate how much they think each idea will be affected by the program or technology. The sorting and rating information is aggregated, and multidimensional scaling and cluster analysis are performed. The end-product is a representation — a map — of the participants' ideas. Actually, you will see that several different but related maps are computed. The participants learn where their ideas have been placed on the maps, reach consensus on names for the clusters of ideas and begin to internalize what the maps mean. The final step involves developing a plan for utilizing the maps to help address whatever issues or goals the group originally had in mind.

For this pilot example, I invited a few faculty members and graduate students to participate. For the focus of the concept mapping, participants were asked to brainstorm statements that describe "specific issues in the assessment of international agricultural research impacts for sustainable development." The brainstorming session lasted about a half hour and the participants generated 98 statements.

Here are several of the brainstormed items, selected arbitrarily:

- resources to conduct research
- role of extension services
- relation to strategic planning
- managing multidisciplinary evaluation teams
- impact on non-agricultural activities
- dealing with unanticipated impacts
- training of evaluation personnel
- quality of data
- nutritional issues
- germ plasm conservation

Next, each person was asked to do two tasks: sort the statements into piles of similar ones, and rate each statement on a 1-to-5, Likert-type, response scale (where 1 meant the statement had "relatively little importance" and 5 meant that the statement was "extremely important" compared with the others). This concluded the first session, which took about two hours.

Between the two group meetings, the sorted data were analyzed using multidimensional scaling and cluster analysis, and we obtained average ratings

Map 1. Point map arrangement of 98 brainstormed statements.

for each statement across all participants.
At the second meeting the group interpreted the maps. First, they examined the 13-cluster solution, deciding on names for each cluster of statements. The first map the group examined, and the most basic one, is the point map that shows the arrangement of the ninety-eight brainstormed statements (Map 1). In general, statements more frequently sorted together are closer on the map, those not sorted together as frequently are more distant on the map. Typically, we spend some time familiarizing participants with the layout of the map. We often "take a trip" across the map, arbitrarily looking at what statements wound up in what locations.

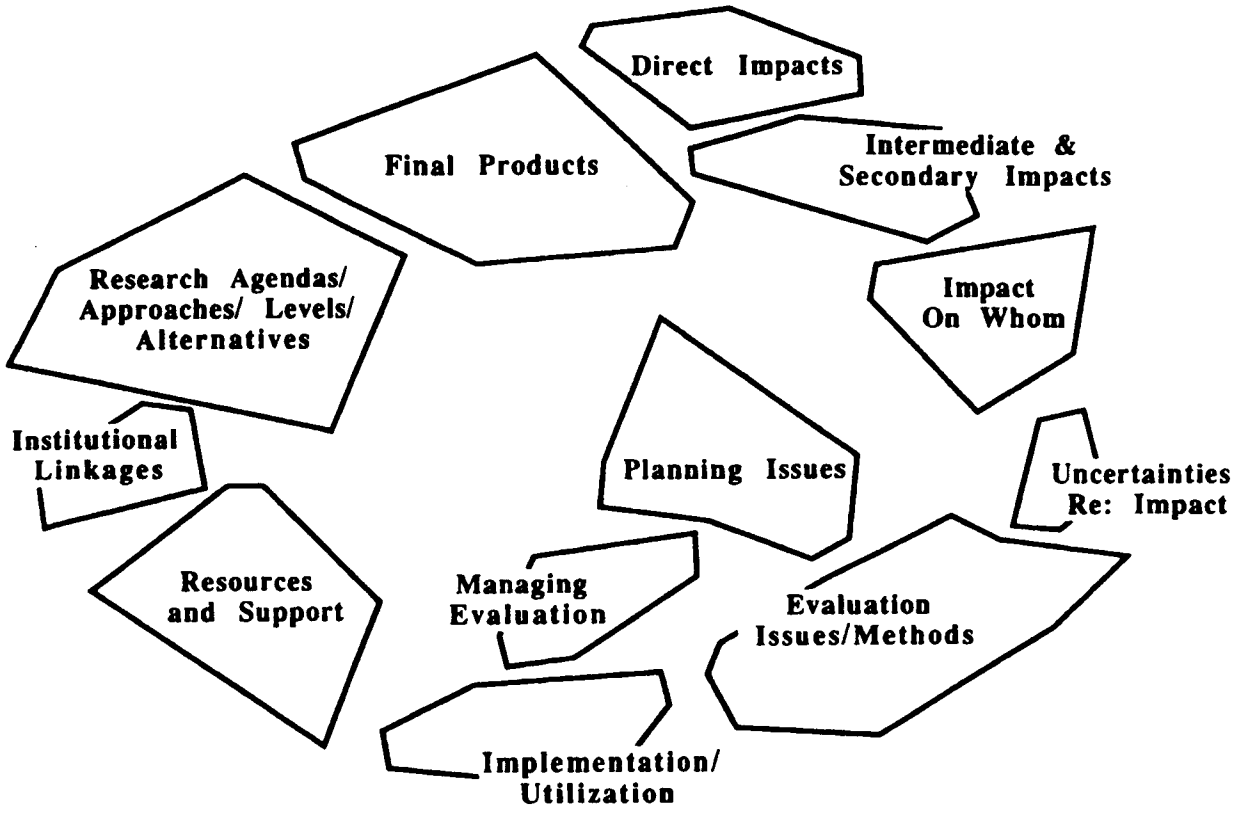
For example, on the left side of the map, statement 39, "institutional issues throughout the international agricultural system" is close to statement 71, "dependence on NARS." At the top of the map, statement 52, "germ plasm conservation," is near statement 65, "diversity of products." On the right side, statement 26, "impact of IARCs vs. impact of NARS" is next to number 42, "impact on IARC collaborations." Or, on the lower right, statement 27, "research designs for impact assessment," is next to number 10, "assessment mechanisms that can be utilized within the scientist's work."

The second map shows how the cluster analysis relates to the placement of the points (Map 2). The cluster names the participants chose are written on this map to aid in the interpretation. We can begin to see some clear patterns. The upper right side is dominated by four clusters having to do with impacts. The lower center and lower right emphasize evaluation and planning issues. The left side includes institutional and contextual factors in agricultural research.

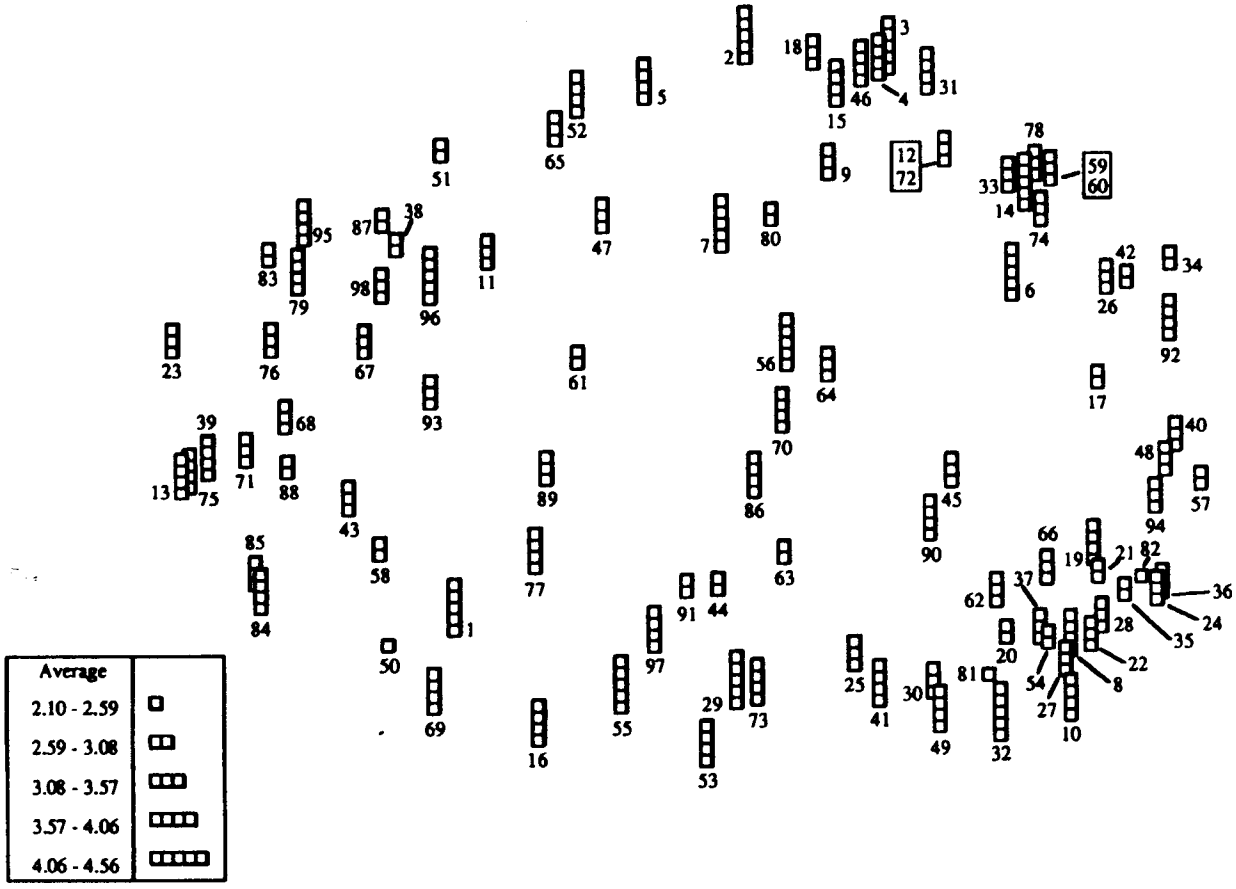
The maps presented so far are calculated based only on the way the participants sorted the statements. The next map shows the 98 points as before, but here, columns of different heights are used to show the average importance rating for that statement (Map 3).

Usually it is hard to see the forest for the trees with all the individual points displayed. It is much

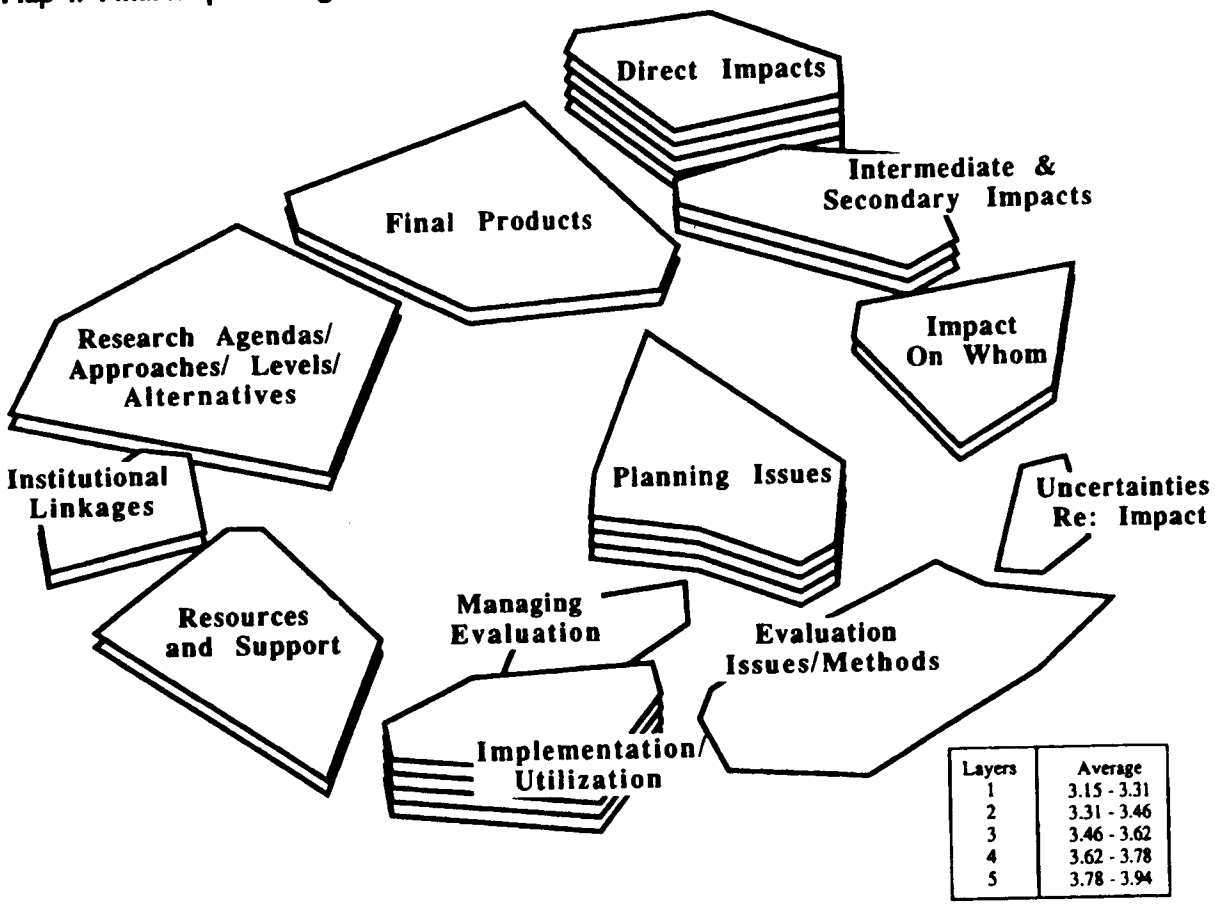
Map 2. Map of 98 brainstormed statements grouped into 12 clusters



Map 3. Map showing average importance rating for each of the 98 brainstormed statements



Map 4. Final map showing labeled clusters and average relative importance



easier to interpret a cluster map showing the average importance across all the items in the cluster. Here, the more layers in a cluster, the higher the average importance of the items in that cluster.

The final map (Map 4) shows the labeled clusters and indicates their average relative importance as perceived by this group of participants.

Interpreting the concept map example

Let us use this concept map to explore some of the evaluation issues in international agricultural research. We begin in the northeastern section of the map with the clusters related to impacts. Two of these clusters describe types of impacts: **direct** impacts in one cluster; **intermediate** and **secondary** impacts in the other. But what do these mean? The direct impacts include the statements "increasing agricultural productivity" (2), "increased adoption of technology" (3), and "improved crop varieties released" (4). The intermediate and secondary impacts include "influence on government policy" (9), "sociocultural context" (14), and "interaction of policy impact and research impact" (74). Direct effects in this example seem to be at the farm level; indirect and secondary ones are at the government or system level.

This suggests a problem in how we define impacts. How we label the impact — direct versus indirect or secondary — depends very much on where we stand in the system. If you are doing farm-level research, the most "direct" impacts consist of what happens on the farm. Farm-level productivity is an obvious choice, but there are many others including rates of technology adoption, immediate economic and environmental consequences, levels of resources expended in producing crops, and nutritional implications, to name a few. From the farm-level point of view, effects on more remote systems — government policies, resource ownership distribution patterns, sociocultural contexts, or the NAR system — are likely considered more intermediate or secondary outcomes.

Consider how this ordering of impacts changes when viewed from the perspective of the CGIAR system. The most direct and immediate impacts are effects of the CGIAR system on government policies or the NAR system. Farm-level, environ-

mental, or economic impacts are more intermediate or secondary **because they can only be achieved through the NARS, the governments of recipient countries, the agricultural extension network, and other channels**.

We need to be careful in interpreting this concept map, and throughout this symposium, to make a clear distinction between these two orderings of impacts. When we talk about "immediate" or "direct" impacts, we should ask ourselves "from whose point of view?" The plant breeder may see direct effects as tonnage yields per hectare; the IARC research manager may look for changes in government policy or extension training practices.

We cannot examine impacts unless they can be measured. If the CGIAR system is like many other evaluation areas, measurement efforts are spotty at best, with some areas of excellence and many topics that are not even routinely assessed. You may find it harder than most fields to try to construct high quality measurement systems because you often deal across cultures, language groups, and national borders in areas that may be remote and with few resources available to underwrite such efforts. The CGIAR system has an essential role to play in working with NARS representatives in the development of standards for the measurement of the most vital and common impact data.

Many of the impacts of interest to you are also important in other fields such as international development, international nutrition, and environmental studies, among others. There is much to be gained by crossing disciplinary boundaries and coordinating measurement efforts. The donors in the international agricultural system can play a major role in bringing these disciplines together. If you are like other fields, there is also probably a fair amount of re-inventing of instruments, interview protocols, checklists and rating scales that could be adapted from other work. If it is not already, the CGIAR system can provide a clearinghouse capability to help reduce such duplication.

When we move to the eastern portion of the map, we reach the cluster: Impact on Whom. This is a key issue for this symposium. The statement in this cluster that received the highest importance rating was "impact on strengthening NARS." If your goal is to evaluate the International Agricultural Research Centers, looking at the effects on the NARS ought to be a major priority — it is through the NARS that the most immediate and direct outcomes are likely to be seen.

Now to the southeast region: here is a cluster concerned with uncertainties in **measuring** impacts. The issues here are more methodological than conceptual — *how* can we best assess costs and benefits; *how* can we deal with unanticipated outcomes; *how* can we deal with changing impacts over time? This leads to the evaluation issues and methods clustered in the southeastern portion of the map. This cluster had the largest number of statements — 22 of the 98 brainstormed items were placed here. If I had done the mapping, this cluster would be divided into many more sub-clusters reflecting the range of evaluation issues we "toured" earlier. Evaluators can provide you with a clearer sense of what is involved in these matters.

The cluster on planning issues refers primarily to planning for **evaluation** rather than to planning as a task of its own. The role of training (Mook 1985) is central — you will need a critical mass of professionals who are trained in evaluation. How can this be achieved? One way is to fund a few large evaluation projects in the CGIAR system. This would probably result in increased funding for graduate students with evaluation skills, increase the incentive for some to choose evaluation as a specialization, provide new evaluators with a venue for honing their skills in this context, and increase the legitimacy of evaluation in the eyes of others in international agriculture. But in times of fiscal austerity and diminishing resources, we need to find other models for creating the critical mass of trained and committed evaluation personnel.

An efficient and relatively inexpensive way to move in this direction is to use the resources and capabilities of universities like Cornell to provide leadership and education in evaluation. You have a special contribution to make here, along with your colleagues in other international fields, because you can extend the evaluation field already well established in this country into the international domain, especially through the funding of graduate evaluation education and the training of students and professionals who come from throughout the CGIAR system. Universities can also encourage faculty like myself -- who have evaluation skills but little experience in international agriculture -- to address issues of concern to you.

Let us go south on the map — these clusters refer to the management, implementation and utilization of evaluation. The issues in these areas link the evaluation methodologies in the southeast with the institutional/contextual issues in the west. The

central issues here involve: being politically sensitive; managing interdisciplinary evaluation teams; conducting research in different cultural settings; and dealing with specific implementation and utilization constraints. While evaluators have much to offer on these topics, I cannot help but think you will be able to teach us much more. As evaluation becomes more internationalized — as it inevitably must — you will be at the leading edge of these important issues. In the near term, your efforts to conceptualize implementation and utilization issues will be essential for your work, and can represent an important contribution to the evaluation field (see the discussion by Baird 1985, for some useful beginnings in agriculture). We evaluators can help you with methods to do planning and assessment (from our knowledge of the cluster in the southeast), but you can provide the institutional and contextual experience (from your experience of the clusters in the west) to suggest how best to merge evaluation into your arenas.

The west is the portion of the map that I am least familiar with — issues related to the structure and processes of the CGIAR system. Under Resources and Support, the statement rated most important was "donor support depends on impact." But what kinds of impacts? Are donors going to be willing to support research programs that demonstrate clear impacts on NARS-level outcomes if the **causal** connection to farm-level productivity is less clear?

The irony is that as you succeed in convincing donors of the importance of looking at system-level impacts, there is likely to be an increase in competition for already scarce resources. Traditional agricultural researchers will inevitably see their share of the pie become smaller and the competition will not only be fierce, but also potentially destructive for all. It is legitimate for traditional agricultural researchers to ask "Why are resources previously devoted to improving productivity increasingly diverted into systems research?"

Evaluation advocates within the CGIAR system must take the lead in making the case for the value of evaluation. One simple, but important, way to begin making the case for evaluation in international agriculture is happening right now — this symposium we are holding. To the extent that all of the important constituencies within the international agricultural community can come together for open discussions about the appropriate role of evaluation, I am optimistic that an acceptable balance of resources can be achieved. But the process may be difficult at times, and we cannot pretend otherwise.

However it is done, I am optimistic that when it is recorded how we in the twenty-first century ultimately came to grips with sustainable development issues, it will be clear that the evaluation field we are in the process of shaping today made a critical contribution.

You face, more than most organizations in this country I am familiar with, complex institutional and intergovernmental arrangements that put unique constraints on your research. The far west portion of the map describes some of the institutional issues you know about and need to consider — the interaction of research and extension, the dependence of the system on the NARSs, institutional instability and the need to maintain continuity in research programs, the implications of private versus public sector activities, linkages with advanced laboratories, and the special situations in certain geographical regions such as Africa, to name but a few.

I suspect that we will be discussing many of these issues over the course of this symposium. Evaluators like me will not have much to add on the substance of such debates, but will be able to help you construct processes for facilitating discussion and determining the areas of consensus and disagreement. You will add greatly to the role of evaluation in international contexts by sharing your experiences in struggling with these issues.

Finally, we come full circle on the map to the cluster labeled Final Products, and I think it fitting that we conclude our tour of the map on this topic. It is ironic that one of the statements in this cluster is "the relationship to sustainable development" — a central topic for this symposium that is both unavoidable and difficult to grasp. Clearly we are only in the early stages of determining how actions in agriculture and many other fields affect the broader ecosystem that we inhabit. At this stage, one of the most important tasks is to conceptualize and define what we mean by "sustainable development." Concept mapping would be one candidate for this, but there are many others.

For instance, a group called the Global Tomorrow Coalition invented a process they call the Globescope Assembly process that essentially consists of large, locally cosponsored assemblies focused on long-term global issues. Each assembly takes from two to five days, and blends plenary sessions and workshops with focus groups designed to maximize participant response. Each assembly produces a product — an action plan, models for replication, or consensus recommendations. However it is done, I am optimistic that when it is recorded how we in the twenty-first century ultimately came to grips with sustainable development issues, it will be clear that the evaluation field we are in the process of shaping today made a critical contribution.

Using the concept map in CGIAR system evaluation

Before leaving the topic of concept mapping I want to describe other ways it might be useful in the CGIAR system. On the planning side, centers might use concept mapping at the formulation stage to involve a variety of stakeholders — including representatives from NARS, the technology distribution network, governmental positions, and farm-level organizations — in helping to identify the major issues that need to be addressed by the centers.

Operationally, concept mapping can help in planning specific center programs and activities. For instance, imagine a concept map developed to plan a training program for NARS personnel. Each of the items on the map could be a specific topic to be included in the training. The clusters would show the **general** topics to be covered and might be useful in ordering the presentation of topics. The map as a whole could give training participants an overview of the course, and keep them oriented to where they are as the course unfolds.

For evaluating this training, the map could be used like a visual checklist of the training topics, showing at a glance what material was covered and what was not — an inexpensive and straightforward form of implementation evaluation. Participants or observers could rate the quality or adequacy of the presentations of each course topic and the map could be used to show which areas were presented better. The items on the map describing specific training program activities could be used to construct a test for assessing how much participants learned. Results from the tests could then be displayed on the map, as well as in traditional tables and graphs.

Clearly, I could go on and on. For me, concept mapping is a bit like a hammer — with it in hand, everything starts looking like a nail. But the method is broadly useful and, especially when contrasted with how we currently deal with formulation and conceptualization tasks, it may be a particularly compelling alternative. One point that I hope you take away from this example is that many processes like this are available for improving planning and evaluation and could be valuable in international agricultural contexts. Some are technical and require considerable training; some are straightforward and easily implemented. Some are quantitative, some qualitative, and some are both. We need to be aware of these methodologies, test them out, and learn where they are advantageous and where they are inappropriate.

Regression-Discontinuity for Ex-Post Evaluation

I would like to turn my attention to the idea of ex-post or outcome evaluation. When evaluators think about testing outcomes or assessing impact, we usually have in mind the testing of a causal hypothesis -- whether a treatment or program can be demonstrated to have caused some outcome or result. In order to say that some program or technology "caused" some outcome or effect we have to meet several conditions. Perhaps the most difficult is the establishment of internal validity -- the determination that the observed outcome was brought about by our presumed cause and not by other potential factors. We evaluators know of no better method for assessing a causal hypothesis than the randomized experimental design. Random assignment to either the program condition or to a no-program comparison group assures a probabilistic equivalence between the groups prior to the study and increases the likelihood that observed differences afterward are attributable to the program rather than to group differences like sociodemographic or psychological variables. Random assignment is the best way we know of constructing a fair counterfactual situation in human research. The comparison or control group in such a design represents our best estimate of what most probably would have prevailed in the absence of the program.

There is a certain irony in my telling **you** about experimental design and causal hypothesis testing. We evaluators owe a great debt to agricultural research, and especially to the work in experimental design done by R. A. Fisher. As a psychology and evaluation graduate student in the

late 1970s, I studied the application of experimental designs developed in agriculture to human research contexts -- including such arcane variations as split plot and Latin square designs. Randomized experiments have been utilized widely in evaluation research -- in studies of the effect of income maintenance programs on subsequent economic and social performance of the poor; of time-of-day electricity pricing schemes on conservation of electricity; on the effectiveness of peremptory arrest of persons involved in domestic violence on subsequent violation rates; of effects of long-term residential care for the elderly compared with alternatives; and on and on. There is almost no field of evaluation that has not been studied with randomized experimental designs at some point. Boruch and Wothke (1985) describe a huge collection of samples of such studies that at last count easily numbered in the hundreds.

Nevertheless, as we begin to apply these experimental designs inherited from agriculture to human research questions, we experienced some difficulties that stem from the human nature of our research. Ethical issues are probably the most important of these -- if it is morally acceptable to use random assignment to determine whether or not persons receive a program or treatment. In many social contexts we have decided that the social value -- of learning with a fair degree of internal validity whether or not a program works -- was sufficient justification for random assignment of humans during trials. In this country, for instance, one cannot get a drug approved for sale without having conducted randomized experiments to demonstrate its efficacy. But this has always been an area of contention, and it will undoubtedly continue to be. One need only look at the current controversy regarding the testing of AZT for AIDS to get an idea of the volatility of the issue. The simple fact of the matter is that we do not and should not have the same degree of control over humans in research that we have over plots of ground in agriculture.

Although we try to conduct randomized experiments when they are feasible and justified, there will be settings where they are neither. Rather than give up on the goal of getting at the effects of programs, evaluators have developed extensive experience in research designs that might be used to help assess outcomes when randomized experiments cannot be accomplished. These methods -- often labeled quasi-experiments -- share with their randomized cousins some of the major features common to good designs, most notably, before and after measurement and use of no-treatment comparison groups. They differ from their experimental counterparts in one major and important way -- they do not use random assignment to determine what conditions people receive.

The literature on quasi-experimental design is considerable (see, for instance, the classic work of Cook and Campbell 1979); there are many variations, with different strengths and weaknesses. Along with other evaluative tools, they will be important methods for us to consider in evaluating the impacts of agricultural research. Although they are not perceived to be as scientifically credible as a randomized experiment, in some situations they may be the best methods available when we want to investigate causal hypotheses in real-world contexts.

To illustrate the use of a quasi-experimental approach to studying technology outcomes in international agricultural research, I will give an example of one specific design that I have had a major hand in developing (Trochim 1984, 1990). For our example we will construct a hypothetical scenario: we wish to introduce a new agricultural technology in a developing country, and we want to assess the impact of doing so with a fair degree of scientific credibility.

In an ideal world, we would want to control the introduction of the technology in order to be able to identify its effects as distinguished from economic trends, local variations, and other factors that affect productivity and related social, economic, health and environmental consequences. We might accomplish this by dividing the potential trial sites into some suitable units -- counties, provinces, townships, villages, or even individual farms. For our arbitrary example we will say that we have a population of villages and are able to treat villages as units. We might wish to control who gets the technology by using random assignment of willing villages to the new technology -- a type of lottery. After testing the effects of this new technology and determining its efficacy we would then introduce it more widely into the control villages.

There are many reasons why such a scheme might fail -- and I am sure that you would do a far better job of detailing them than I could ever do. But, let us imagine that, for ethical or political reasons, random assignment was deemed unacceptable. (Poorer villages might legitimately complain that random assignment would lead to at least some of them being denied a potentially better technology, one that they need more than

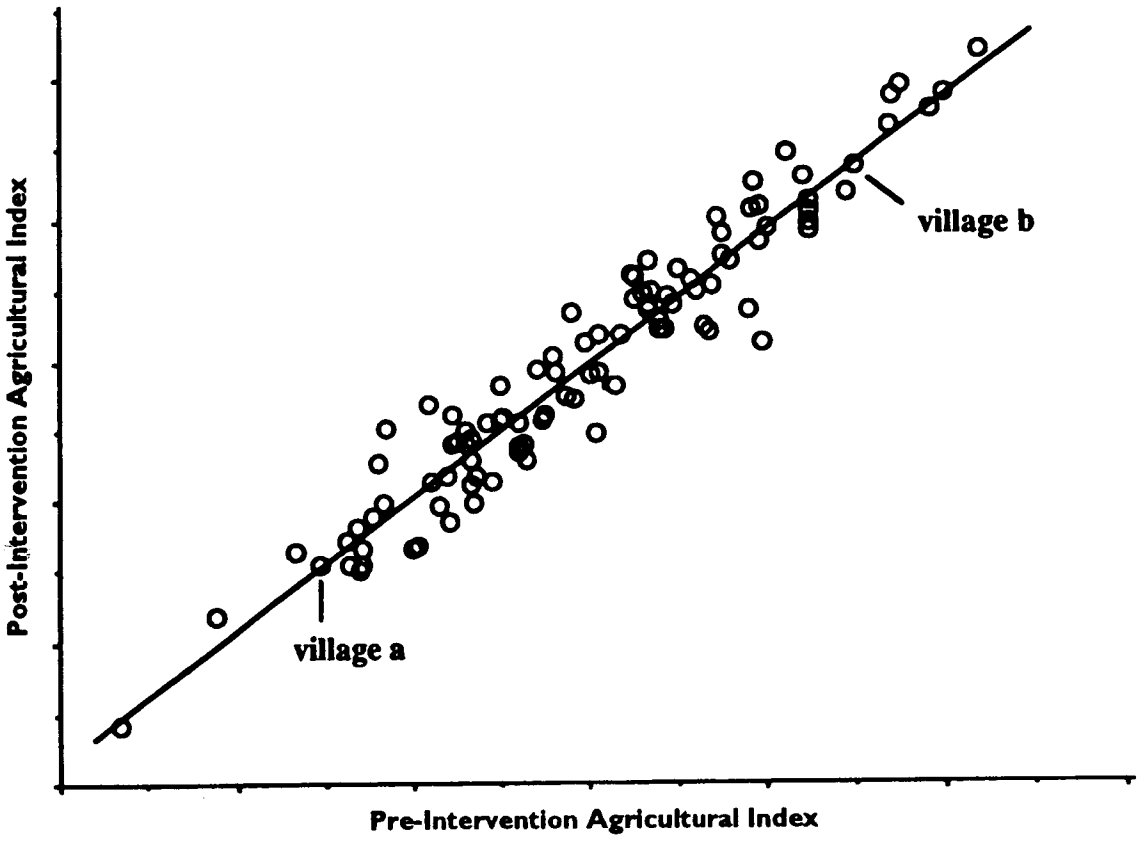
other villages do.) While random assignment may not be palatable here, it may be reasonable (again for ethical or political reasons) for us to assign the new technology to those villages that need it most. Could we do so and still study the effects of the new technology with some degree of scientific validity? Is there a way for us to meet the administrative/political desire to get the new technology to those most in need, while at the same time doing a fair job of assessing impact?

Under this type of scenario, one possible fall-back strategy to the randomized experiment is the quasi-experimental approach that we call the regression-discontinuity design. In the regression-discontinuity design, units -- in this case villages -- are assigned the new technology based on their standing relative to a cutoff value on a pretest index or variable. For our hypothetical example, we might have a measure of the agricultural performance of each village during the previous year. If we want to assign half of all the new villages to the new technology, we could set the eligibility cutoff value at the median of the agricultural performance indicator. All those villages falling below that value would receive the new technology; those above the cutoff would be considered the no-treatment controls.

I forgive you if your initial reaction is that this is a preposterous design idea -- you would be in excellent company. On first hearing of this regression-discontinuity design, most researchers think it absurd -- how can we call our two groups "comparable" when the assignment rule deliberately makes them non-comparable on a pretest assignment index? What kind of counterfactual situation would the high-scoring non-recipients represent?

To see a little more how this design works, let us look at a simple graph (Figure 2). First, imagine a "null" case -- a situation where we measure the agricultural production of our villages two years in a row with no intervention introduced. We might get a bivariate distribution like the one shown in the graph. The graph shows 100 hypothetical villages measured on some index of agricultural productivity over two successive years. Each dot on the map represents a single village. The villages clearly vary in agricultural productivity. There is a clear correlation over time -- productive villages in one year tend to be productive in the next -- but the relationship is far from perfect -- there is considerable relative movement or variability from year to year. The solid line that goes through the bivariate distribution is the simple

Figure 2. Pre-post distribution of 100 hypothetical villages on an index of agricultural productivity.



linear regression of the second year values onto the first year values.

Consider the points representing the villages labeled "a" and "b" on the graph. Village "a" is relatively low on the agricultural index in both years, while village "b" is relatively high. The distribution describes the pre-post relationship for the villages in the absence of any concerted intervention over that period. This relationship, however, includes in it all of the factors that tend to make villages more or less productive from one year to the next -- local weather, politics, demographics, productivity levels, economic factors, and so on.

Now let us imagine that we had introduced our technology to the lower scoring villages, but not to the higher scoring ones. What might we observe if the technology was effective in increasing agricultural productivity on the post measurement? For the simple case of a constant additive treatment effect, the results might look like Figure 3.

Notice that all of the cases scoring below the cutoff value on the pretest indicator received the technology. Their scores are elevated on the post index by the intervention. But those scoring above the cutoff would show the same pre-post relationship as in the previous graph. Village "a" scores higher than it would have in the absence of the intervention; village "b" scores exactly where it would have. If we had used a regression-discontinuity design to study the effects of this new technology, we would only have observed this second graph.

How would we determine that the technology was effective? Recall that in the null case, we would expect that the pre-post relationship would be describable by a continuous function. In the intervention case, if the technology is effective, we would expect that a "discontinuity" would be introduced into the "regression" line coincidental with the cutoff point that divides the two groups -- hence the name "regression-discontinuity" for this design. The high scoring controls clearly are not comparable to the low-scoring intervention cases on the pretest indicator variable.

The key assumption in the regression-discontinuity design is that in the absence of the intervention both groups are similar in their pre-post relationship -- that is, when no treatment is given we could use the same regression lines to describe both groups. But perhaps even this hypothetical scenario presumes a level of control that is seldom achievable in international agricultural settings. Let us briefly look at an even more equivocal, but perhaps more feasible, quasi-experimental design. Imagine that at first we wish to introduce our new technology in very limited trials before disseminating more widely. For demonstration purposes, we decide to try it out first on only a single village to see what happens. How could we examine impact in this case? Clearly, one thing we would want to do is to study that village in some detail during and following the course of the intervention. Good quantitative measurement within the village will probably be important here. This might include the construction of an information database to keep track of agricultural, economic, health, social, and other relevant outcome variables.

Qualitative, observational, field study methods are also indispensable here, especially if we take care to build in design features that enhance the credibility of such data. For instance, using multiple independent observers or interviewers who analyze their data in isolation before comparing results will in general be stronger than reliance on only single observers. But if we find changes from before to after the new technology is introduced, we would still probably be unsure whether or not they can be attributed to the technology itself or to other factors. A comparable control village with similarly intensive measurement would help, if one can be located and the expense can be justified. But we should also be alert for opportunities to capitalize on existing administrative data that can be utilized. If village-level agricultural or economic data are available on a regular basis, we might at little cost be able to incorporate it in a quasi-experimental analysis that is similar in principle to the regression-discontinuity design described earlier.

For instance, imagine that we decided to introduce the new technology in village "a" in the figure previously shown. If we had our agricultural index measure for that village and the other 99 non-recipients, we could examine whether the post-index level for village "a" is statistically higher that we would predict based on the pre-post regression for the 99 non-recipients considered together (see Figure 4).

In this quasi-experiment, we are testing the significance of the "displacement" of the single "point" from the control unit's regression line -- a design we now are calling "regression point displacement" design. When the administrative data is available, such an analysis is simple to accomplish, and very inexpensive. It alone will not

Figure 3. Hypothetical pre-post distribution for a regression-discontinuity quasi-experimental design

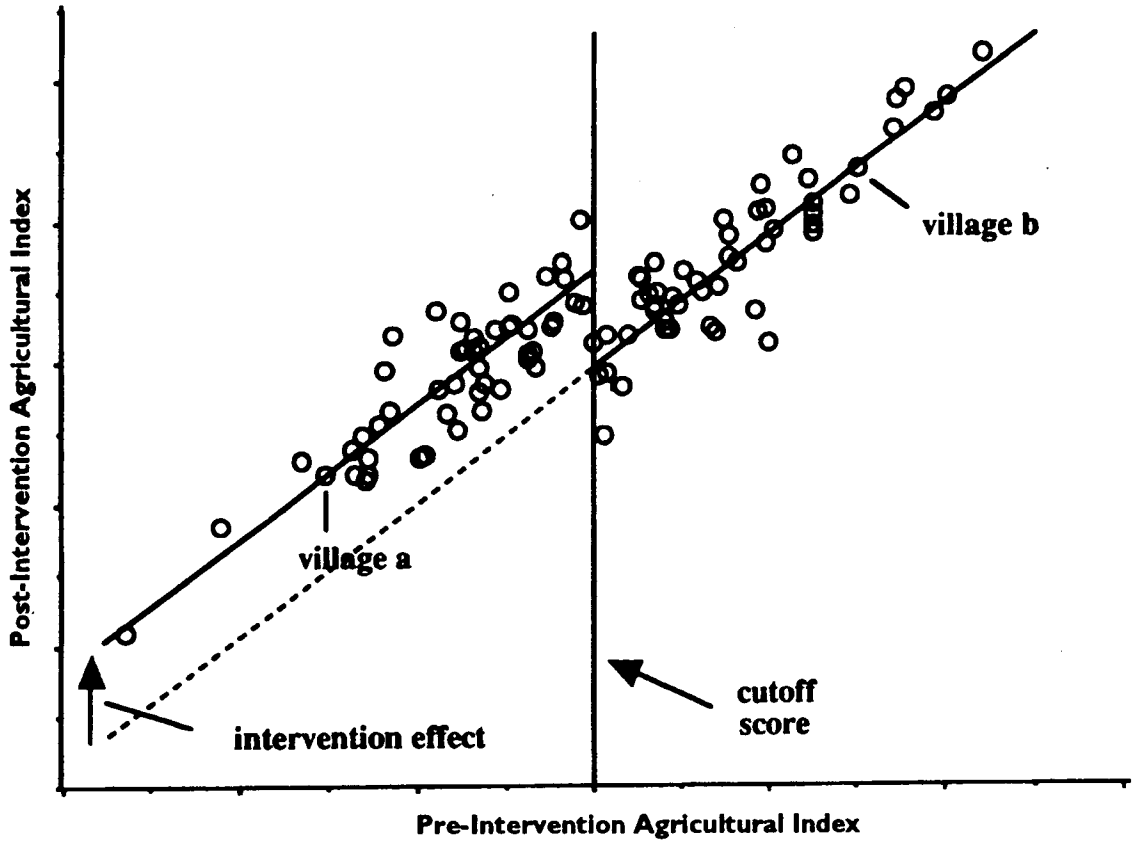
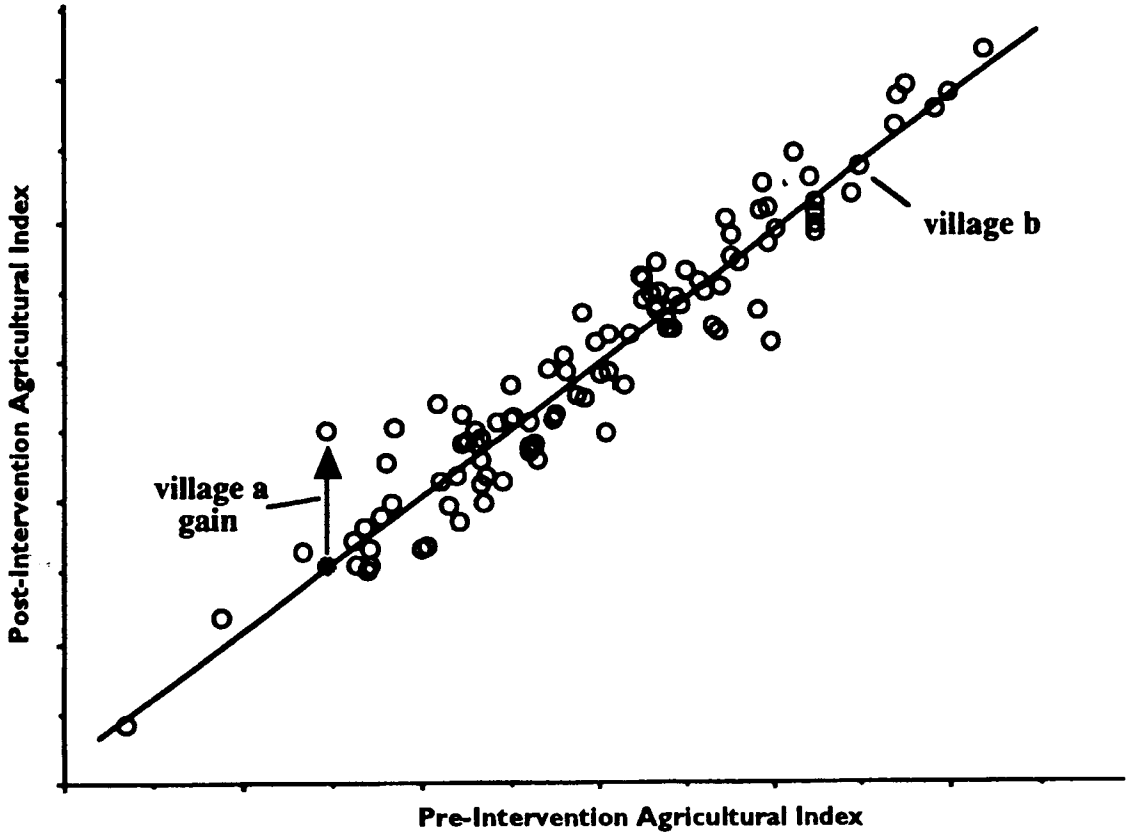


Figure 4. Hypothetical pre-post distribution for a regression point displacement quasi-experimental design



prove that the new technology brought about the post- change, but taken together with all the other information it can be a valuable addition to the evaluation.

My purpose in introducing these quasi-experimental designs is to draw your attention to the fact that there is rich domain of potential approaches to be considered for their relative advantage and degree of appropriateness for outcome assessment in agriculture. Importantly, many of these are designed to be used in natural settings, they minimize disruption of normal administrative practice, and yet enable scientifically credible, defensible assessment.

Before leaving the topic of ex-post assessment, I want to insert a cautionary note about our ability to assess specific impacts of broad, global treatment packages such as international agricultural research programs. There is no simple way to evaluate something so complex. My belief is that we need to take a multi-faceted approach to examining the broad complex hypotheses that are so hard to assess and yet impossible to ignore. In many cases we will need to break down the question into smaller ones, developing a theory of how technology gets transferred and ultimately produces effects -- postulating intervening variables such as organizational productivity of the IARCs and NARSs, researcher and administrator motivation and satisfaction, quality of record-keeping and information, and so on -- that we believe lead to better quality research. At some levels we will need to rely on our impressions, on anecdotal, qualitative, case-study approaches. But that does not mean that we need to abandon the goal of achieving scientifically credible, defensible results with such methods.

Especially for these methods we need to incorporate recent advances and suggestions such as the use of independent field note coding teams and checks on reliability of coding, qualitative data audits, internal consistency checks, and so on. Some of the smaller questions we identify as part of a larger theory of technology dissemination will be amenable to small experimental or quasi-experimental demonstration studies and, when resources permit and the need for validity is great, we should be prepared to mount such studies. We will also need to seek methods for integrating information from different data sources. Complex, multifaceted interventions invariably require complex, multifaceted methodologies. We will need to deal with such messiness, being willing to try new methods, while clinging resolutely to the desire to get at the truth of what is going on in the contexts we study.

Developing an Evaluation Culture in International Agricultural Research

My closing suggestion is to encourage you to join in the ongoing effort to develop an *evaluation culture**. Evaluation will be an essential ingredient in your continuing success in international agriculture, and in your ability to address issues of sustainable development. The key to this evaluation effort is the development of a frame-of-mind that values the role of evaluation, incorporates evaluation into all levels and branches of the international agricultural research system, and views evaluation as a natural, ongoing human activity that is an integral part of what we mean by doing a good job.

What would an evaluation culture look like? What should its values be? You should know at the outset that I fully hope that some version of this fantasy will become an integral part not only of the thinking of the international agricultural research community, but of twenty-first century thought more generally. I will point out along the way how this culture might apply to the international agricultural research endeavor. But I want you to realize that this idealistic view encourages you to become involved in something larger than just your own field — participants in an international interdisciplinary frame of mind regarding how we will use information in the twenty-first century to make the world a better place. There is no

*While its historical genesis involves many different strands, there is no doubt that my thoughts here are most influenced by the work of Donald T. Campbell (1988) who twenty years ago articulated a vision of an "experimenting society" that we would all do well to study intensively today. Much of my thought is descended from that seminal work, but in the spirit of the contentious, disputatious community of truth seekers that Campbell advocated, I have taken the liberty of extending the original vision and, I hope, improving upon it, and I encourage you to do the same. For starters, I have abandoned the term "experimental" in the "experimenting society" in favor of the term "evaluation" in "evaluation culture." I do not mean this to deprecate the value of experimentation. As mentioned earlier, I believe that in an evaluation culture, experimentation will have an important place. But I want to convey the notion that our goal is broader than experimentation narrowly defined. I have also replaced the term "society" in Campbell's "experimenting society" with the term "culture.' The term "society" implies to me something too structured, organized and operationalized for what I intend. I prefer "culture" because it refers more directly to a state of mind that characterizes a group or a society. It is this state of mind that I wish to see permeate the thinking of researchers, truth seekers, participants, donors, clients, and others, that justifies the use fo the term "culture."

particular order of importance to the way these ideas are presented — I will leave that ordering to subsequent efforts.

First, our evaluation culture will embrace an *action-oriented* perspective that *actively* seeks solutions to problems, trying out tentative ones, weighing the results and consequences of actions, all within an endless cycle of supposition-action-evidence-revision that characterizes good science and good management. In this activist evaluation culture, we will encourage innovative approaches at all levels, from the CGIAR system to the individual farm. But well-intentioned activism by itself is not enough, and may at times be risky, dangerous, and lead to detrimental consequences. In an evaluation culture, we will not act for action's sake — we will always attempt to assess the effects of our actions.

This evaluation culture will be an *accessible, teaching-oriented* one that emphasizes the unity of formal evaluation and everyday thought. Most of our evaluations will be simple, informal, efficient, practical, low-cost and easily carried out and understood by non-technicians (see Binnendijk, 1989). Evaluations will not just be delegated to one person or department — we will encourage **everyone** in our organizations to become involved in evaluating what they and their organizations do.

Where technical expertise is needed, we will encourage the experts to also educate us about the technical side of what they do, demanding that they try to find ways to explain their techniques and methods adequately for non-technicians. We will devote considerable resources to teaching others about evaluation principles. In international agriculture this will require providing resources and organizational support to train persons who have responsibilities at all different levels of the CGIAR system and who come from all over the world.

Our evaluation culture will be *diverse, inclusive, participatory, responsive and fundamentally non-hierarchical.* World problems cannot be solved by simple "silver bullet" solutions. There is growing recognition in many arenas that our most fundamental problems are systemic, interconnected, and inextricably linked to social and economic issues and factors. Solutions will involve husbanding the resources, talents and insights of a wide range of people. The formulation of problems and potential solutions needs to involve a broad range of constituencies. More than just "research" skills will be needed; especially important will be skills in negotiation and consensus-building processes.

Evaluators are familiar with arguments for greater diversity and inclusiveness — we have been talking about stakeholder, participatory, multiple-constituency research for nearly two decades. No one that I know is seriously debating anymore **whether or not** we should move to more inclusive participatory approaches. The real question seems to be **how** such work might best be accomplished, and despite all the rhetoric about the importance of participatory methods, we have a long way to go in learning how to do them effectively.

Our evaluation culture will be a *humble, self-critical* one. We will openly acknowledge our limitations and recognize that what we learn from a single evaluation study, however well designed, will almost always be equivocal and tentative. In this regard, I believe we too often undervalue cowardice in research. I find it wholly appropriate that evaluators resist being drawn into **making** decisions for others, although certainly the results of our work should help **inform** the decision makers. A cowardly approach discourages the evaluator from being drawn into the political context, helping assure the impartiality needed for objective assessment, and it protects the evaluator from taking responsibility for making decisions that should be left to those who have been duly-authorized — and who have to live with the consequences.

Most program decisions, especially decisions about whether to continue a program or close it down, must include more input than an evaluation alone can ever provide. For instance, in international agricultural research, is it inevitably desirable that a "better" plant variety from a crop productivity point of view should always be investigated by a research center or implemented in a society? Are there times when we might decide *not* to investigate or implement such a technology because it is **likely** to lead to unacceptable social dislocation, economic inequity or unfair political advantage? The weighing of the relative trade-offs of such values as crop productivity versus social and economic dislocation is inherently a political and societal concern.

While evaluators can help to elucidate what has happened in the past or might happen under certain circumstances, it is the responsibility of the organization and society as a whole to determine what *ought* to happen. The debate about the appropriate role of an evaluator in the decision-making process is an extremely intense one right now in evaluation circles, and my position advo-

cating a reluctance of the evaluator to undertake a decision-making role may very well be in the minority. We will need to debate this issue vigorously, especially for politically complex international evaluation contexts like international agriculture.

Our evaluation culture will need to be an *interdisciplinary* one, doing more than just grafting one discipline onto another through constructing multi-discipline research teams. We will need such teams, of course, but I mean to imply something deeper, more personally internalized. We need to move toward being **non**-disciplinary, consciously putting aside the blinders of our respective specialties in an attempt to foster a more holistic view of the phenomena we study.

As we consider new agricultural technology, we each should be able to speculate about a broad range of implementation factors or potential consequences. We should be able to anticipate some of the organizational and systems-related features of technology dissemination, the economic factors that might enhance or reduce implementation, the social and psychological dimensions of the technology transmission mechanisms, and especially whether the ultimate adopters can understand, know how to utilize and are willing to utilize the technology. We should also be able to anticipate a broad spectrum of potential consequences — system-related, production-related, economic, nutritional, social, and environmental.

In preparing this paper, I heard many anecdotes about well-designed crop varieties that never got adopted because we did not take into account the lifestyles, traditions and views of the farmer, or the politics and economics of a given region or country. I heard stories of crucial outcomes that were not routinely anticipated — economic inequities and dislocations, negative nutritional outcomes, potential widespread crop destruction due to over-planting single-variety monocultures, dams that failed because engineers did not trust local wisdom about flooding ferocity, and so on.

Omitting these potentially critical factors from consideration the first few times around may be forgivable, but we cannot claim dispensation any longer. A cross-disciplinary expectation that encourages us to anticipate such factors at all levels of the international agricultural system is doable now. Building in more formal interdisciplinary collaboration on research projects will be an essential part of our eventual evaluation culture (Binnendijk 1989).

This evaluation culture will also be an *honest, truth-seeking one that stresses accountability and scientific credibility.* In many quarters in contemporary society, it appears that many people have given up on the ideas of truth and validity. Our evaluation culture needs to hold to the goal of getting at the truth while at the same time honestly acknowledging that all scientific knowledge is revisable. We need to be critical of those who have given up on the goal of "getting it right" about reality, especially those among the humanities and social sciences who argue that truth is entirely relative to the knower, objectivity an impossibility, and reality nothing more than a construction or illusion that cannot be examined publicly. For them, the goal of seeking the truth is inappropriate and unacceptable, and science a tool of oppression rather than a road to greater enlightenment. Philosophers have, of course, debated such issues for thousands of years and will undoubtedly do so for thousands more. We in the evaluation culture need to consider their thinking from time to time, but until they settle these debates, we need to hold steadfastly to the goal of getting at the truth — the goal of getting it right about reality.

Our evaluation culture will be *prospective and forward-looking*, anticipating where evaluation feedback will be needed rather than just reacting to situations as they arise. We need to construct simple, low-cost evaluation and monitoring information systems when we **first** initiate a new program or technology. We cannot wait until a program is complete or a technology is in the field before we turn our attention to its evaluation.

Finally, the evaluation culture I envision is one that will emphasize *fair, open, ethical and democratic processes.* We should move away from private ownership of and exclusive access to data. The data from all of our evaluations need to be accessible to all interested groups allowing more extensive independent secondary analyses and opportunities for replication or refutation of original results. We should encourage open commentary and debate regarding the results of specific evaluations. Especially when there are multiple parties who have a stake in such results, it is important for our reporting procedures to build in formal opportunities for competitive review and response.

Our evaluation culture must continually strive for greater understanding of the ethical dilemmas posed by our research. Our desire for valid scientific inference will at times put us in conflict with ethical principles. The situation is likely to be especially complex in international agriculture

where we will often be dealing with multiple cultures and countries that are at different stages of economic development and have different value systems and mores. We need to be ready to deal with potential ethical and political issues posed by our methodologies in an open, direct, and democratic manner.

What other characteristics might this evaluation culture have? There are many other values and characteristics that ought to be considered. For now, the ones mentioned, and others in the literature, provide a starting point at which we can all join the discussion. Over the course of this symposium, you may add to the list, and I encourage each of you over the next few days and thereafter to criticize these tentative statements I have offered about the extraordinary potential of the evaluation culture that we are all in the process of creating today.

Discussion

The pertinence of evaluation in the CGIAR system
Various participants expressed their support for more attention being given to evaluation within the CGIAR centers. Although there was concern that this had the potential for being overemphasized and requiring too many resources, a general consensus was reached that current circumstances created the need for an enhanced focus on evaluation. Chandler pointed out that, in terms of productivity, the centers are dealing with diminishing returns. The development of another IR8 is not the likely solution; rather, research will increasingly focus at the margins. To do this, the centers must better define and monitor what is being done.

The IARCs' strengthened awareness and added responsibility for dealing with sustainability issues was cited as a good reason for a strengthened evaluation focus. One center social scientist thought that greater attention to detail in evaluation and an emphasis on ex-ante analysis are consequently called for in response to sustainability concerns. No one could deny the importance of donor support in bringing evaluation to the forefront in the CGIAR system. It was agreed that the donors' increased need to know what is going on in the centers and to see the effects of their research investments require that different measures be taken to "spread the CGIAR story."

The state of evaluation in the CGIAR system
Conflicting comments from participants spurred Trochim to ask whether the centers currently are or are not heavily involved in conducting evaluation. One center scientist listed several evaluation studies that had been conducted in his center over the past few years, and further suggested that "evaluation fatigue" seemed evident among the staff. On the other hand, documentation and the overriding opinion of participating center staff point to a general lack of evaluation in the centers.

Trochim, in an explanation of how these conflicting opinions sometimes arise, suggested that it often has to do with a breakdown in the feedback loop. That is to say, the information generated by evaluation is often lost in the system, frequently not reaching the public domain. Collinson, referring to his earlier presentation, pointed out that the amount of evaluation and hence "fatigue" very much depend on the constituency being addressed. There was general agreement that the center managers are conducting evaluations, but that at other levels -- for example, with respect to the centers' ultimate beneficiaries or the centers' impacts on NARS -- less attention is devoted.

Several participants voiced concern about the risk of excessive expenditures and overemphasis on evaluation in international agricultural research. It was agreed that refining what is already occurring in the area of evaluation and integrating evaluation into the daily lives of scientists, as opposed to externally imposing evaluation, were two ways of managing the resource problem that could arise.

Future possibilities for evaluation in the CGIAR
Participants expressed some curiosity and skepticism towards the proposed "evaluation culture," a prime concern being the danger that evaluation within research could be overemphasized. Trochim explained that although this is always a risk, his idea was to make it a more integrated, internalized process. In practice it would take on a less obvious, but more important role in research decision making.

Participants agreed that methods need to be simplified, be made more "quick and clean," and more frequently utilized. A mixture of methods, both formal and informal, could be applied. It was pointed out that the potential of evaluation lies in the "feedback loop," through which information is constantly being used in decision making. It was suggested that the breakdown of this information flow may be what has caused scientists' "evaluation fatigue" in some centers.

The extent to which Trochim's view of evaluation was systematic in nature came under scrutiny. He explained that by a systematic process, he was referring to an evaluation methodology that would be public, observable and capable of being retraced. It was pointed out that this was subjective but represented an effort toward objectivity, which is the best that can be attempted.

The evaluation culture was summarized simply as a way to use information so that feedback is placed in the forefront and useful evaluation information is used in the decision making process in a constructive manner.

-- Mary Ellen Mulholland,
Rapporteur

Assessing the Impact of International Agricultural Research for Sustainable Development

Proceedings from a Symposium at
Cornell University, Ithaca, NY, U.S.A.
June 16-19, 1991

Edited by David R. Lee, Steven Kearl and Norman Uphoff

Sponsored by CIIFAD,
the Cornell International Institute for Food, Agriculture and Development

Cornell International Institute for Food, Agriculture and Development
Box 14, Kennedy Hall
Cornell University
Ithaca, New York 14853-4203

The Cornell International Institute for Food, Agriculture and Development (CIIFAD) seeks to promote sustainable agricultural and rural development around the world. It helps Cornell University faculty and students work with colleagues overseas in interdisciplinary, problem-focused, collaborative ways, to generate knowledge, develop human resources and strengthen institutions so that our understanding and use of resources -- natural, human, physical and intellectual -- will provide benefits and opportunities for generations to come.