PN-AAR-750 38089

SOCIOLOGIA RURALIS
Vol. XXII No. 3/4 - 1982

23

A PRACTICAL METHODOLOGY FOR EVALUATING INTEGRATED RURAL DEVELOPMENT*

by

FRANK W. YOUNG

Department of Rural Sociology, Cornell University, Ithaca, NY, USA

In 1979 the Tunisian government set up a federally sponsored and controlled "Central Tunisian Development Authority" that spanned five governorates, involved about 400 employees, and had expenditures of five or more million dollars per year. USAID agreed to contribute to this effort on a matching basis over three years, and the proposal included a range of interventions typical of semi-arid regions: irrigation, drought-resistant seed development, potable water, rural infrastructure, small business development, upgrading health facilities, etc.

The CTDA program was evaluated in the traditional way in the summer of 1981. At that time, a joint Tunisian-American team spent several weeks reviewing progress and problems. The five-man team read a mountain of documents, worked out checklists for the many components of CTDA, developed scenarios, and then provided feedback to CTDA and ministry staff before the final report was written. Such "process evaluation" will no doubt continue to be the standard approach to evaluation in these and similar circumstances. It is rapid, relatively inexpensive, contractually predictable; and, in principle, it can penetrate the complexities, subtleties, and processes that have been going on over time and in diverse corners of the organization. Such appraisals are less efficient in evaluating impacts on the local residents and in sorting out the interventions that made a difference in the level of poverty, wheat production, environmental deterioration, etc., that are equally, if not more, important to a development project

* I am indebted to J.-P. Habicht, Michael Miller, L. Richard Meyers, Charles Perry, Norman Uphoff, Donald E. Voth, and Gerald Wheelock for comments that substantially improved this paper. The research on which the paper is based was supported by the Office of Rural Development Administration of the Agency for International Development via its sponsorship of Cornell's Rural Development Committee. I am also indebted to the Central Tunisian Development Authority for its support. None of these agencies is responsible for the opinions and conclusions expressed in the article, of course.

than its administrative efficiency and service delivery record. They also falter in dealing with the regional character of most rural development projects. What goes on in the main office may or may not correspond to events in the communities.

The two conventional alternatives to process evaluation were neither feasible nor appropriate in this situation. A household/farm survey of the 80 subdistricts in the pilot region would have been prohibitively expensive, considering the dispersion of the households and the terrain. Even assuming an adequate baseline survey, it is difficult to see how a questionnaire could ascertain the impact of such multiple public goods as improved water supplies or a more participative local government. Household questionnaires are not adapted to measuring performance or impact of whole institutions.

The other alternative is even less attractive. In a country like Tunisia, it is possible to develop indices of welfare for each of the governorates and for the region as a whole on the basis of the decennial census. The amount of change for Central Tunisia over the period 1975 to 1982, the year of the next planned census, could be assessed and the rate of change for Central Tunisia compared to that for the four other regions in the country. The problem with this approach is that no other region is comparable, so the inference that an especially marked improvement of average welfare in Central Tunisia resulted from the agency's efforts would be open to challenge. If data are available for the smaller district units, as is true of Tunisia, the inference of causation from project efforts could be made stronger, but then a new problem would arise: Does such district-specific evaluation demonstrate the *integrated* character of this development strategy? Probably not, and even an interrupted time-series design, if it were possible, would not demonstrate that an integrated impact is better than the more usual unintegrated strategy. It may be, in fact, that testing for the effect of integration is beyond the capacity of any of the evaluation designs reviewed here, but it is virtually certain that a design based on census data would be too cumbersome and limited for even a minimum evaluation of this type of program.

These two alternatives require different types of data, but their design is essentially the same. It involves before and after measurement, statistical matching according to whatever control variables are available, and an inference (usually weak) about program impact. Given the data sources, neither of these two designs will be strong with respect to measuring the programs that operated in the districts or subdistricts. Neither do the methods facilitate the measurement of the background factors that must be introduced as controls. Yet there is every reason to believe that these background factors are powerful, perhaps even stronger than the agency interventions.

So the question is whether an approach to evaluation exists that is appropriate to the rough and tumble reality of rural/regional development projects. At the outset, we can rule out randomized assignment of treatments. Such a demanding design would be unfeasible in this environment, quite aside from the political problems. Thus, we are perforce dealing with a design that is less than ideal. However, the compromised design need not give up the requirement of before-and-after measurement. Admittedly, conducting a post-test is extremely problematical because the evaluators (and perhaps even the agency) may not be around after four, five, or ten years; and it is difficult to achieve comparability even across a five-year span. There may be places where a development agency starts work at a "zero level", obviating the need for a baseline, but such cases are rare. It is assumed here that a baseline is necessary along with an appropriately timed post-test. The basic logic is as follows:

$$Y_2 - Y_1 = a + bI + e \qquad (1)$$

where Y_2 is a criterion such as wheat yields or income, measured after the intervention; Y_1 is the before measure; and I is some measure of intervention. $Y_2 - Y_1$ is the "gain score," and of course it can be standardized by dividing by Y_1. The constant "a" and the error term "e" are necessary components of any estimating equation, but they are not germane to the present problem, so they will be dropped in the subsequent discussion. The "b" coefficient is understood.

This formula implies that; apart from error and unknown causes, the intervention was fully responsible for the observed changes. That is quite unrealistic, of course, because all sorts of forces are at work in any region. In Central Tunisia, for example, the subdistricts with urban centers clearly had an advantage, and other subdistricts had higher proportions of young men who went to the coast or to Europe to work. So a comprehensive equation should identify and include such "background factors" (X_1, X_2, X_3, etc.), which, however, are often so important that they really should be called "foreground factors."

Equation (1) has another deficiency, in that the familiar gain score introduces biases that are difficult to deal with. Consequently, it is preferable to compute the gain score by regressing Y_2 on Y_1. In other words, the baseline level of the criterion variable (Y_1) is considered a variable that must be controlled in predicting Y_2. Taking these improvements into account gives us the following formula:

$$Y_2 = Y_1 + X_1 + X_2 + X_3 + I \qquad (2)$$

That is, the value of the criterion variable at time 2 is predicted by its

3

value at time 1, several background factors, and the intervention(s) under study.

Equation (2) does not specify a unit of analysis; yet a decision on the appropriate level is crucial. For example, if the intervention takes the form of agricultural roads, it is probably the district, of which there are 25 in Central Tunisia, that is the appropriate unit. If the interventions consist of a package of rural services like dispensaries, water points, primary schools, etc., then the subdistrict would be the appropriate unit. The study of farm-specific interventions requires a sample of farms.

For most rural development activities, it is likely that the subdistricts, which are typically dispersed rural communities, would be the appropriate units. There are about 200 in Central Tunisia. At the subdistrict level, a measure of intervention might be a count of new services. Alternatively, one might calculate the dollar value of project investment or the number of man-hours of project activity. Then the subdistricts that received no services or few services would serve as the control group for comparison with those that received more. Similarly, the other interventions offer contrasts of intervention/no intervention.

It will be objected that the subdistricts that received no new services (or other investment) do not constitute a good control group, because they almost certainly differ in other respects from the treatment groups. At minimum, they are probably more isolated from the urbanized subdistricts. That is certainly true, but much of this variability is taken into account by the measures of the "background factors" that are already in the equation. If X_1 signifies level of urban services, X_2 the distance from the state capital, and so forth, then we have, in fact, achieved a certain level of comparability. The precise subdistricts that serve as controls in this statistical format will vary with the set of background factors and the intervention measure, and matching will never be perfect. But it will be feasible.

The next step is probably already evident. In order to assess the impact of a given intervention, one compares the regression equation for a given Y and given I, when a set of X's are controlled, with one that omits I. The comparison turns on any difference between the two R^2's, that is, the variance explained, which is a coefficient of the efficiency of prediction. For example, if the R^2 for the equation that includes I is .60 and that for the second is .50, there is then a .10 difference. It is possible to assess whether that is significant on a statistical basis and also in terms of practical impact. Such comparisons assume fairly accurate measurement, but that is true for any evaluation.

If a difference attributable to an intervention can be demonstrated, one must, of course, examine the sign of the coefficient and, if possible, the form of the relationship. Any interpretative leverage, whether fact

4

or concept, must be brought to bear. A valid result is impossible if conceptual understanding is absent.

Because of the multicollinearity that is almost always present, this design systematically underestimates and sometimes masks the effects of interventions. It is also inefficient when the treatment group starts from a very low level of development (Cook and Campbell 1979, p. 300). It is possible, however, to introduce ad hoc corrections. In any case, these deficiencies are generic to this whole class of designs. The purpose here is to emphasize a new data collection and interpretation strategy.

A Feasible Technology for Regional Evaluations

There is considerable agreement on the evaluation logic that has just been described, and Voth (1975) has applied it in an evaluation of rural development in Illinois. As noted, there are problems with this approach, but they are known and are no more intractable than the many other problems that research involves. That leaves the question of appropriate technique. It is proposed here that the "informant survey" is a feasible and appropriate technique for district and subdistrict units. The proposed technology can be summarized as follows:

1) Select a set of *impact criteria* that, in addition to being relevant to the interventions under study, have universal applicability. In Tunisia we devised five: level of living of the poorer farmers, a crude index of income inequality, an estimate of wheat yields, a scale of restrictions on women, and a score of environmental problems.

2) As a *source of information*, use people in a public role. In Tunisia we interviewed the *omdahs*, that is, the local leaders who were also paid officials of the government. Thus they were both informed about their rural community and available for an interview.

3) Choose a *unit of analysis* that is larger than the farm/household but smaller than the province. One needs around 100 cases. In Tunisia we focused on the subdistrict (*secteur*), for which minimum census data and some agricultural information were available. Our pilot baseline study was limited to 80 subdistricts, but it could have been expanded to around 200.

4) On the basis of theory and observation, select *control variables* (i.e., the X's) operating at the district or subdistrict level that are likely to determine levels of the impact criteria. In Tunisia we made preliminary judgments that the availability of water, the presence of urban centers, road access to them and to other places, migrant work, type of agriculture, agricultural machinery and practices, political relationship to the district headquarters, local organization, religious organization, and the characteristics of the local leader would affect the

5

criterion variables, particularly poverty and inequality. Factor analysis reduced many of these variables to five main dimensions: urban services, religious institutions, mechanized agriculture, irrigated agriculture, and sedentary herding. The factor analysis of many variables gives some assurance that the evaluator has captured the significant context variables. One can never be sure that all pertinent controls have been included, but the identification of major factors, instead of specific variables, expands the content of the variables that are explicitly controlled.

5) On the basis of exploratory field work, interviews with officials, and any available reports, design a *questionnaire* appropriate to the informant and to the dimensions. The questionnaire should have the following general characteristics: it should be nonthreatening and minimally inconvenient to the informant; it should be organized according to categories like agriculture, administration, water sources, etc., that are familiar to administrators; and, most important, it should be made up of many simple objective questions so that the informant can answer yes or no with some readily known fact. As a general rule, the questionnaire should not ask the informant for his opinions or for complex judgements. Readily observable items like the condition of roads should be observed and such data collected separately according to a checklist. Preliminary fieldwork may require several months; and, of course, the final instrument should be pretested.

6) If possible, the researchers should build in various *validation checks*, such as items that are identical to census questions or responses that are subject to independent verification.

7) The *administration*, editing, and coding of the questionnaire should build in the standard quality controls.

8) Although the *analysis* of these data follows the general lines well known for sample survey data, there is one major difference. The many small separate items must be compressed into scores, scales, or other more comprehensive measures. This process not only converts the initial qualitative data to quantitative dimensions, but it transforms the natural categories used by the informant to a form more appropriate to multivariate analysis, without, however, destroying the social validity of the original reports.

9) Early in the analysis, *subregions* should be delineated and used in cross-tabulations for presenting the basic variables like population, sources of water, crops, etc. These cross-tabulations are invaluable for describing the context in which the intervention will be introduced.

10) After the general descriptive exploration suggested in (9), the set of *impact criteria* should be presented in detailed form. If possible, the measures should take the form of continuous variables. The requirement is important for later regression analysis. Similarly, the

6

principal context variables should be continuous and, if possible, normalized. As noted already, factor analysis is often an appropriate technique for summarizing the many components that make up the set of structural dimensions that will be used as background variables in assessing the impact of interventions.

It may be objected that a technique like factor analysis identifies broad dimensions of context that may obscure the impact of particular and significant variables – for example, remittances from migrant work. That is indeed a risk, and some contextual variables should be examined separately. But it is equally risky to limit the analysis to narrow measures. Social structure does not usually operate through such particular channels. If social factors have an impact, they must be institutionalized, and institutions are group-level and usually systemic. These are substantive assumptions and must be tested, but they are more likely to lead to a realistic description of the situation than the administrator's frequent preoccupation with selected features of the social environment.

11) Finallly, each impact criterion should be *analyzed in relation* to the set of context variables in order to establish the baseline relationships (i.e., equation 2 with I omitted). These regression equations serve as a final check on validity, because it is difficult to get reasonable results if the measures are not reliable and valid all the way up to this point. For example, the Tunisian baseline analysis demonstrated that urban insitutions, religious institutions, and (negatively) sedentary herding predicted rural affluence as measured by the housing scale. Theory and observation strengthened these relationships. In contrast, it was impossible to show relationships between wheat yields and any of the five context dimensions. Thus, the reliability or validity of this criterion is in doubt. Note that the conventional baseline study does not include such an analysis. One must simply accept the baseline measures as offered.

This list of steps summarizes the essentials of a baseline study, which is about all that can be accomplished until the program has had an impact, perhaps five years later. However, there are some special problems For example, the informant survey administered in Central Tunisia asked each local leader how many dry (unirrigated) farmers' households had windows, doors, stoves, sewing machines, latrines, etc. This type of question, which required the informant to estimate household attributes, differed from the majority of the questions, which dealt with community institutions, such as whether or not there was a school and, if so, when it was built. The estimates of household attributes probably strained the knowledge of some of the informants, particularly in the larger communities. Our method of asking for absolute numbers and then computing percentages seemed to work,

7

but other techniques for obtaining such information should be tried. Step 5 above should focus on this and related problems of estimation.

Household attributes seemed absolutely necessary for estimating level of living, which is a fundamental criterion of rural progress, but the other impact criteria were easier for the informants. For example, the questions dealing with restrictions on women required only a knowledge of customary practice. Similarly, the questions dealing with the environment depended on observations that any farmer would ordinarily make.

Informant surveys do have special problems of reliability and validity, but they are no more difficult than those inherent in household surveys. One can presume that three or four decades of methodological research will refine informant surveys, just as they have household surveys. Meanwhile, at perhaps one-tenth the cost, the informant survey is accurate enough for the baseline analysis and subsequent evaluation that is recommended here. There are several reasons for this:

a) The proposed universal yardsticks of progress are more likely to involve readily observed facts that a knowledgeable and long-term resident can report accurately.

b) The observations and judgments that the informant is asked to make are, if the preliminary fieldwork is adequate, important to him and to other members of the community. Therefore, if he has not observed a difference after five or more years, it is safe to say that the agency has not had an impact that is meaningful to local people. The fact that the values for the impact criteria vary widely indicates that change has been possible and, therefore, that a given subdistrict could, in principle, change for the better.

c) This is the only method that can supply sufficiently accurate data on institutional patterns for use in the subsequent evaluation equations. Neither anthropological reconnaissance nor the standard household/farm survey generates measures of the institutional structure. These, however, are of overwhelming importance in determining levels and potential of income, ecological problems, etc.

d) There is nothing in this methodology that blocks the research team from using conventional reconnaissance or even a household/farm survey. If resources are available, information from both sources can be added to the informant-based file. A comprehensive statement on monitoring and evaluating agricultural extension projects by Cernea and Tepping (1977) illustrates some of the connections that are possible. First, they propose separate farm surveys for each project within an integrated program, which is an effective way to deepen the analysis. One of their proposed criterion variables is, of course, agricultural productivity, and to the extent that the sample survey includes farmers from many villages – a minimum of 20 communities would be

necessary – it should be possible to validate the informant estimates of productivity. Third, the recommendation that Cernea and Tepping make with respect to case studies of particular projects is certainly appropriate to informant survey results. There is always a need for information about context and possible causal connections.

Despite these many complementarities, it should be apparent that the approach proposed here for evaluating integrated rural development departs radically from the conventional project evaluation model that attempts to show the connection between project inputs and impacts on specific target populations. Here we have multiple inputs, some of which may not even be assignable to particular projects. Moreover, it makes little sense to talk about target populations. The agency is oriented toward raising the welfare of the whole region, or at least that of some of the subdistricts, and it claims to be able to do so. Even if it were possible to find the target populations for a particular input, it would be difficult to show a specific effect, because the diverse interventions are designed to complement each other and have a general impact. By contrast, this design takes rural development rhetoric at its word and evaluates the effort against general development criteria for whole communities.

Complexities of Regional evaluation

Assuming that the baseline study is accurate and valid and that the research team returns after a reasonable interval, the post-test would then be administered in the same way as the baseline, except that the second questionnaire must contain detailed information on the presence and impact of the agency. If agency records are adequate for devising some of these independent indicators, so much the better. Unfortunately, it is not possible to list actual measures. This design paper is being written on the basis of a baseline study only,[1] so we cannot describe in detail how the interventions should be conceptualized and measured. They have been suggested earlier: the presence or absence of specific inputs, the location of particular input combinations, the number of new services, the number of man-hours or dollars invested, etc.

A more general point is that "intervention" will actually take many and diverse forms. There are different ways to measure the overall impact of a development project, and one could even use conventional program inputs on the assumption that the universal yardsticks will reflect their impact. When all the specific inputs are tested, along with the general interventions and the many alternative measures, there may be as many as 100 pairs of regression equations. Such flexibility is an argument in favor of this approach, because integrated rural develop-

ment programs often provide many diverse interventions. A more fundamental point is that the programs and projects are never so well planned or sharply defined that one can say with certainty that a given program should be evaluated against a particular criterion. Programs shift with circumstances and the political winds, and goals and aims are often reformulated.

There is still another advantage of this design, but it may not be seen as such by the agency director. A fundamental assumption of this approach is that a treatment impact can be claimed only after all relevant variables of context have been controlled. Of course, no evaluation can anticipate or measure all possible control variables, but the major dimensions should be entered into the analysis alongside the inputs. When this is done, it will frequently happen that background variables will account for all or most of the explained variance. The intervention may then show no demonstrable effect. It is possible, of course, that the intervention is so highly correlated with a contextual factor (e.g., new clinics with the road system) that its effect is hidden. This design weakness has already been acknowledged. On the other hand, such a hidden effect should make an administrator wonder whether the intervention (clinics in this case) could ever function alone.

Such a result is difficult for an administrator to accept. If wheat yields have risen sharply over five years, it is only human to claim success for the agency's actions. It is difficult for an administrator to believe that the environment is strong and rooted in history and that his interventions are rather weak in comparison. Development efforts are usually comparable to the attempts by applied meteorologists to change the weather by cloud seeding. Sometimes their small and well-targeted interventions are successful, but more often the stronger forces of weather prevail.

Of course, a finding that it was a particular dimension of institutional context rather than the intervention that accounted for an observed change in a criterion variable is important. The alert administrator should ask himself how he can use such knowledge to advantage. Unfortunately, such findings usually pose the problem of nonmanipulable variables. If analysis shows that proximity to an urban center is the primary determinant of increased average incomes, the administrator is hard put to increase the level of urbanization in the space of five or even ten years. If the service centers are small and the regional development agency is powerful, it may be able to create new centers and facilitate growth in the old ones. But in the usual case, the administrator is concerned with building agricultural roads and bridges, drilling wells for irrigation or potable water, or distributing dispensaries. Indeed, analysis may show that these many efforts, which usually

require much thought, investment, and administrative guidance, are gradually incorporated as components of the structural context. For example, rural service centers typically subsume specialized establishments like clinics and credit agencies. The sum of these specialized establishments defines an urban center, and it is this level of development that is the operative variable in rural development. Thus, the administrator may realize after five years of hard work that he has simply elaborated the social structure that is already there; he has not really intervened in a new way. That realization may be more humbling than the knowledge that one's interventions had no measurable impact.

A final consideration returns us to the choice of level of analysis. Even if evaluation shows that interventions at the subdistrict level had an impact, it is always possible that the agency efforts were wholly or partially determined by activities at a higher system level, say, the district political or urban organization. If data at the district level are available, such relevant variables should be included in the subdistrict comparisons. It may also happen that events at the different levels move in opposite directions. One can imagine, for example, a series of interventions that produce significant positive change in a criterion variable but at the same time the average level of the whole region declines. A decrease of just a few cents in the government-imposed price of wheat could have such a region-wide effect and possibly invalidate the success of many specific interventions. Or the opposite could happen: the average income of the whole region could rise despite the fact that all interventions failed to enhance the incomes of the target subdistricts.

Every experienced administrator is familiar with forces that he cannot control, and it is thus no novelty to suggest that evaluation may reveal such forces. If they emanate from the national or regional levels, it is not likely that the statistical analysis will reveal their presence explicitly, but a strong indirect case may be made on the basis of general knowledge. Thus, even though the proposed design is necessarily limited when it comes to the possible cross-cutting effects of state policy, it at least calls attention to them. Sometimes adjustments are possible.

NOTES

1. The results of the baseline study are summarized in Young et al., 1981a, 1981b, and 1981c.

REFERENCES

CERNEA, M. M. & B. J. TEPPING (1977), A System for Monitoring and Evaluating Agricultural Extension Projects (Washington, D.C.: World Bank Staff Working Paper no. 272)

COOK, T. D. & D. T. CAMPBELL (1979), Quasi-Experimentarion: Design & Analysis Issues for Field Settings (Cnicago: Rand McNally College Publishing Co.)

VOTH, D. E. (1975), An Evaluation of Community Development Programs in Illinois, *Social Forces*, 53 (June), 635-647

YOUNG, F. W., F. BERTOLI & S. BERTOLI (1981a), A Baseline for Rural Development in Central Tunisia (Report prepared for the Central Tunisia Development Authority (Ithaca, NY: Department of Rural Sociology, Cornell University)

YOUNG, F. W., F. BERTOLI & S. BERTOLI (1981b),Design for a Microcomputer-Based Rural Development Information System, *Social Indicators Research*, 9 (September), 283-312

YOUNG, F. W., F. BERTOLI & S. BERTOLI (1981c),Rural Poverty and Ecological Problems: Results of a New Type of Baseline Study, *Social Indicators Research*, 9 (December) 495-516

ABSTRACT

As an alternative to conventional project evaluation, which does not meet the challenge of multiple inputs in integrated rural development, it is proposed to survey multiple project impact at the community level by means of a before-and-after informant survey. Multiple regression is used to control on important background factors as well as the initial level of the criterion variable. The community-level informant survey approach is in most respects more appropriate for integrated regional development programs than is qualitative "process analysis" or before-and-after designs that use census or survey data.

RÉSUMÉ

A la méthode traditionnelle d'évaluation de projet, qui ne rend pas compte du problème des effets multiples du développement rural intégré, cet article propose une alternative: celle-ci consiste à étudier ces impacts à travers une enquête préalable et une enquête postérieure auprès d'informateurs dans les communautés rurales. La technique de régression multiple est utilisée pour tenir compte aussi bien des facteurs de base que du niveau initial des variables d'évaluation. A beaucoup d'égards, cette approche par une enquête au niveaucommunautaire se révèle plus adéquate pour évaluer les programmes de développement intégré ou régional que l'analyse qualitative du changement ou les modèles "avant-après" qui utilisent des données de recensement ou de surveys.

KURZFASSUNG

Als Alternative zur konventionellen Projektevaluierung, die nicht ausreichend die unterschiedliche Bedeutung verschiedenartiger Inputs bei der integrierten ländlichen Entwicklung berücksichtigt, wird vorgeschlagen, die vielfältigen Auswirkungen von Projekten durch Befragung von Schlüsselpersonen auf der Gemeindeebene zu ermitteln. Multiple Regression wird zur Kontrolle wichtiger Hintergrundfaktoren und des Ausgangsniveaus der Zielvariablen eingesetzt. Der Ansatz einer Befragung von Referenzpersonen auf der Gemeindeebene ist in vielfacher Hinsicht besser geeignet für integrierte regionale Entwicklungsprogramme als die qualitative "Prozeßanalyse" oder der Vergleich von Ausgangs- und Endsituation an Hand von Daten der amtlichen Statistik oder Umfragedaten.