

UC Berkeley

IURD Reprint Series

Title

Validating Policy Indicators

Permalink

<https://escholarship.org/uc/item/6ht45333>

Author

Innes De Neufville, Judith

Publication Date

1978-09-01

Peer reviewed

File Copy 1



JUDITH INNES DE NEUFVILLE

Validating Policy Indicators

Institute of Urban & Regional Development
University of California, Berkeley

Reprint No. 167

Reprinted from:

Policy Sciences 10 (1978-79),

© Elsevier Scientific Publishing Company, Amsterdam—Printed in Scotland

The Institute of Urban and Regional Development serves the faculty and students of the Berkeley Campus conducting research into a wide array of policy-relevant questions. Its foci of inquiry are the processes of urban and regional growth and decline, the consequences that alternative styles of governance have for patterns of development, and the scientific bases for development policy and for policy planning.

Toward these ends, current research is directed to the economics of urbanization in the West and in the Third World; techniques of environmental management; effects of using new energy sources; emergent patterns of suburbanization and central area reconstruction; evaluation of alternative modes of transportation; design of effective means for supplying social services; contemporary issues in housing policy; regulation of urban growth and uses of land; determinants of urban livability and the factors shaping urban life; and potential improvements in methods of analysis, evaluation, and planning.

The Institute's members are drawn from departments throughout the Berkeley Campus, and their research findings are published in their respective professional journals and in commercially distributed books. The Institute prints and distributes many of their research papers; a catalog of publications is available on request.

Validating Policy Indicators

JUDITH INNES DE NEUFVILLE*

Department of City and Regional Planning, University of California, Berkeley, California

ABSTRACT

Although validity of an indicator, that is, whether it means what it purports to, is critical to its legitimate use in policy, there are few systematic approaches to validation. The reason lies largely in the traditions of positivist social science which require the separation of fact and theory and allow no place for meaning and valuation. The paper proposes approaches to test the validity of both indicators and the conceptual frameworks on which they are based. The most powerful forms of validation require the application of a priori concepts and models. Validation is partly judgmental and never simply mechanical. The validity of a measure is dependent on the context of its use. Social and political processes are an important part of validation, particularly in determining the appropriateness of the assumptions. If an indicator is to be valid, it should be chosen or designed iteratively with the policies it will help to shape.

Introduction

Quantitative measures have demonstrably played pivotal roles in the development and implementation of public policy. They are the flags for public problems, shaping the understanding of their nature and solutions. They become the embodiment of goals and the language of policy discourse and ultimately evolve into standards for the design of programs and criteria for the allocation of public funds (de Neufville, 1975). Despite this potential of quantitative measure or indicators to shape policy, policy analysts and decisionmakers have no agreed upon nor systematic method to assure that the indicators they use are valid and appropriate for their tasks. Indeed, the notion of validity is only poorly understood by those who use policy indicators, though it is the central issue in the choice or design of a measure.

A valid indicator, in the most common definition, is one which measures what it purports to. Validity subsumes many other questions, such as reliability or freedom from bias. These are more frequently tested than the most important aspect, whether the measure *means* what it is being used to mean.

Validity is an elusive question on which the intellectual traditions of policy science provide little help. To validate a measure is to determine if it matches an idea. Three very different realms are involved, those of the empirical data, of the formal constructs, such as the theory and the indicator, and of the meaning. The positivist

* The author wishes to thank Sir Geoffrey Vickers, Professor Martin Krieger and Dr. Norman Shulman for their thoughtful comments on an earlier version of this paper.

foundation of modern social science tells us we cannot "know" meanings except through observation and theorizing about data. Fact and theory are separate, the one being in the world and the other in our heads. The notion of matching one to another is problematic as the language and logic of each is different and there are no independent ways to assess each. Moreover, "meaning" has a connotation which is unscientific and subjective. The tendency has been to neglect the issue and focus on bringing the facts and theories together.

As indicators play roles in policy, as actions hinge upon them, it turns out that meanings matter. Some indicators are accepted in political processes and some are not. Some "work" well and some do not. Some are argued over endlessly. In all this, a sort of validation does take place, though a more subtle and elaborate one than social scientists have employed. This validation is necessary because the "scientific" method falls short. An indicator exists at the intersection of fact and theory. Though it provides a way of seeing facts, it is at the same time a construct or hypothesis about what the fact looks like. It is the lens through which to observe the world. It cannot be tested against a fact because it is one, or the closest approximation that can be achieved. Poverty, for example, can only be identified with the aid of poverty line, and air pollution with the help of an index weighting pollutants. Ultimately an indicator must be tested for validity against some other sense of reality than the empirical data which it organizes. It is these other senses of reality which provide the framework for systematic validation. Thus, other views of knowledge will be called upon to provide the intellectual foundation and rationale for validation of the measures and analysis used in policymaking. One source is likely to be the view that knowledge is socially created (Berger and Luckman, 1966) and therefore meanings are known and shared by groups within a culture. This speaks to validation through social process and social agreement. It says that indicators mean what it is agreed they mean, whether tacitly or explicitly, and that measures developed within the ivory tower are without meaning. Indicators can also be seen as analogous to ordinary language, with meanings which are shared, cumulative and learned through experience (Pitkin, 1972). This would imply that it is in use that indicators acquire meaning and suggests that common sense views and nonexperts are appropriate judges of validity. Much knowledge, moreover, comes with the aid of the analytic function (Polanyi, 1969) and this can be drawn upon too for validation. Finally, ideas from phenomenology, which give central importance to the perception of the observer may help us to make sense of validation.

While the systematic linking of these epistemological traditions to the validation process is still in the future, it is possible to chart some validation strategies which draw on these as well as on policy experience with indicators which have been accepted as valid. The intent of this paper is to provide a guide to validation for the designers and users of indicators. The fundamental assertion is that validation cannot be a mechanical procedure but must be a continuous process involving both designers and users, putting to work all that they know and understand. A validity coefficient will never be permanently attached to an indicator as its correctness will depend on contexts, uses, and users, and its meanings will accumulate and evolve

over time. At best, the design and choice of indicators for policy will be of a single piece with the processes of policy formation and implementation.

Definitions

Some terms should be clarified. An indicator is a construct, a set of procedures for collecting and combining data to stand in for a concept. It is a shorthand to describe, discuss, or manipulate ideas about the world. It enables one to organize empirical observations, connect them with ideas, and give them substance. The indicator of the unemployed is the product of applying a counting rule to populations to identify some members as unemployed. The idea it represents could be the number who want work, or the number who need work. The counting rule is different according to which concept is involved. The indicator also may represent a concept growing from theory. The unemployment rate, for example, is a measure of pressure on the job market and fits into macroeconomic theory.

A concept is part of language referring to shared experience of a phenomenon. Its meaning may be highly complex as its full implications come out through use in a variety of contexts. One's concept of a house, for example, depends on whether one is an urban or farm dweller, corporate executive or nomad. One can define it by its physical characteristics, by what one does in it, or both, and one would differentiate it from related concepts, like apartment buildings, depending on what one wanted to say about it. The logic is created in response to a need for some common explanation for behaviors and for communication (Berger and Luckman, 1966).

This view of concepts implies first that an indicator inevitably means less than a concept. The indicator represents one of the concept's dimensions or represents the concepts less well in some contexts than in others. Therefore, understanding the validity of the indicator means understanding the concepts and their contexts, as well as the indicator. Indicators derive, through the concepts, from the analytic paradigms which help to make sense of things. Yet these mind sets come from outside the data so indicator designers have to take responsibility not only for the measure but also for the conceptual frame they apply. Validity of an indicator has to include validity of the framework.

Types of Indicator Validity

The degree of validity of a measure is dependent on the extent to which it is built on, or can be incorporated in theory. This, in turn, affects the type of contexts in which it may be used and the conclusions that may be drawn from its use. Highly empirical measures have only limited validity and application while the more theoretical ones are also the more possible to validate.

Operational Validity

Indicators created in the tradition of operational measurement, in which a measure is simply the result of performing a specified set of operations, are of questionable value for policy purposes. If an indicator does not purport to measure

something distinct from itself, validation has no meaning. Moreover, the effect on the measure of introducing a policy change cannot be predicted.

Operational indicators such as the IQ measure and the leading indicators of the business cycle have predicted values of other variables such as school achievement or the lagging business indicators. However, the IQ measure is simply a score on a series of questions which may reflect genetically given intellectual capacity, learning, and some random component. No one can say, as the test was developed without concepts about the differences between hereditary and acquired capabilities and without a clear view of the various dimensions of intelligence. The business cycle indicators are a subset of figures on economic activity clustered together because their rises and falls tend to precede those of other indicators. The leading indicators have no special meanings as variables influencing the economy.

At best such indicators can be said to be operationally or instrumentally valid. They can lay claim only to validity insofar as they predict something else or suggest the existence or size of some vaguely perceived phenomenon. However, as one does not know why they worked in the first place, it is impossible to be sure they will continue to work under changed circumstances. The only fundamental connection between an operational measure and its meaning is one of correlation. The blooming of the apple trees may be a useful leading indicator of the rise in employment that comes in the spring when the weather improves, but the indicator has no value if a depression sets in in May. Moreover, if the apples do not bloom because of a blight it will not be reflected in a change in employment patterns.

While most indicators appear to have more of a substantive relation to a phenomenon than in this example, this impression can be misleading. If the rationale for an indicator is only one of correlation, its movements are unpredictable and its behavior under changed circumstances unpredictable. Its valid uses are limited to description or forecasting in reasonably steady conditions but it cannot be a guide to action. Such indicators are not valid as an aid to policy prescription as one cannot, without a model, tell how the measure is related to interventions. Nor should one use them as goals or standards for policies as their meanings are unclear. If the IQ level in a group declines, one does not have a clue where to look for the reasons nor what problems it may reflect. To improve the IQ measure by focusing it more on hereditary intelligence is not possible unless an idea of the meaning of intelligence is elaborated. In addition, an operational indicator can be ascribed a meaning it does not have, and it can be used as if the concept is no more than the indicator. Intelligence thus is reduced to IQ. While this is commonly the fate of measured social phenomena, it is particularly serious when the measure has no clear foundation in ideas, because there is no way of using it with judgment and appropriate constraints.

Experiential Validity

Even an operational indicator may gain a somewhat higher degree of validity if its meaning can be tested against experience. If IQ seems to label the right people "intelligent," people whom one senses in other ways are intelligent, it has more meaning. While it is still poorly specified, it can be applied in more situations.

Experiential validity cannot be fully explicated but may be recognized intuitively. An obvious connection between measure and concept may exist, because the concept is simple in operational terms. Population seems equivalent to numbers of people, for example. Or the understanding may come through shared social norms relating, for example, poverty and income levels. Experience says that 300 days of illness is not good health, but two days a year is. Such validation can be useful and at times may be the only kind possible, particularly when the concept an indicator represents is poorly verbalized or when the context of reality changes, but explicit understandings lag behind intuitive ones.

The United States' poverty indicator is a case in point. When it was designed in 1964, it meshed reasonably well with prevailing conceptions of the income level that goes with poverty. But by the end of the decade after the efforts to use the measure in allocating funds and setting benefit levels, there came a rising chorus of complaints from users about its validity. It no longer corresponded to the opinions expressed in the Gallup Poll about what it takes to live on, though earlier, it had matched well. It turned out that although the "poverty line" had been updated by the price indices for the components each year, it had not kept pace with the cost, much less the standard, of living because the food component of the indicator, which had been set when the cost of food was relatively high, was too great in the sixties. As food prices had not increased as much as other prices, the total index was not increasing fast enough. This and other problems with the index are now well-recognized and the experts are working toward a more valid indicator (U.S. Department of Health, Education and Welfare, 1976).

Experiential validation often works best to suggest invalidity rather than to specify what is valid. It is a misleading criterion when used alone. Many valid indicators, like the reading on a car's speedometer, are intuitively meaningless. If indicators are only valid in an experimental way, it is unclear in what way they match the concept or fail to do so, as it may have merged with the indicator. For example, years of school may seem a good indicator of level of education—in fact, there are senses in which the two ideas are virtually identical. But education implies a good deal more than time served in a classroom. The indicator may not serve well if the concern is with achievement.

Theoretical Validity

Measures interpreted with the aid of theory (Tropman, 1976) can be linked to meanings separate from themselves and thus are amenable to more powerful validation. They are useful in more contexts and can be improved, if they fall short. Model and measure can be iteratively tested against one another and reality. The models involved may be conceptually complete theoretical systems like Newtonian mechanics or simple common sense models defining how one understands things to work. An example of the latter is "the volume of cars on a road, along with speed and weather conditions, determines the number of accidents." This model suggests indicators of volume, speed and weather are of interest and may determine much about how, precisely, they should be measured. Such causal explanations or models

are invariably at work in policy analyses though they are simple and often unrecognized. To explain the meaning and test the validity of an indicator, these models must be made explicit no matter how simplistic they appear to be.

There is extensive literature on how and why measurement with theory is important and how it is done (e.g., Blalock and Blalock, 1968; Torgerson, 1958; Koopmans, 1947; McIntosh et al., 1977). Several arguments are worth bringing out here. First, on any topic there is an infinite variety of observations that could be collected and ways they could be combined. To make the selection implies a prior notion about the nature of the problem or if not, imposes some structure by default. Even to measure the number of books on a shelf, one needs some model of what a book is, which is intimately connected with what a book is for or why one wants to know. For a child anything with outside binding, pages, and printing inside might count as a book; for a librarian two such volumes might be one book if they had one title and author, while others might count as periodicals as they require separate shelving. A faculty tenure committee might regard a handwritten manuscript as a book, but disregard many carefully bound volumes because content was the concern rather than format. In short, the concept takes on meaning in a context, and that context must be defined for the indicator to be evaluated.

Moreover, any indicator will fall short of representing the full concept, particularly since the concept takes on different content in different situations. No indicator of poverty or health or housing will account for these equally well in all situations or for all participants. However, if what the indicator represents and how it is related to the concept are defined, it can be used with appropriate caution or rejected. The indicator may be one dimension of the concept, as morbidity may be a dimension of health, while mortality and a sense of well being are others. To interpret its use as an indicator of health, one would have to postulate how it would be likely to move in relation to the other dimensions of health and under what circumstances it might diverge from these.

The United States Census has tried to use availability of plumbing as a stand-in for housing quality, though it is only one dimension of the issue, on the assumption that it would correlate with housing quality. Users, however, have found plumbing to be an invalid indicator of housing quality, and have extrapolated from 1960 data their own estimates.

Validity could have been tested by looking at the measurement model as there is no logical reason why the existence of plumbing should be paired with freedom from deterioration, particularly when great numbers of low quality units around the turn of the century were built in the U.S. with full plumbing.

An indicator may be related in many ways to a concept. It may be temporally related, always preceding or always following. It may reflect the phenomenon under certain assumptions about other variables. But it will always be most useful and possible to validate if it grows out of a causal model. The height of a column of mercury in a vacuum is a valid index of temperature not only because the link has been tested in many situations but also as the causal relationship between heat and expansion of mercury has been modeled. Education has some causal relation to social status, therefore, understanding the dynamic of this relation along with the

role of factors, like ethnic background, permits the use of education measures as indicators of social status (see Bunge, 1975, for further discussion of the issue).

Theoretical validity demands indicators which are themselves conceptually unambiguous, either because their meaning is simple and directly comprehensible, like the percentage of school-age children enrolled in school, or because they are designed according to some organizing principle. For example, people in many situations might be called unemployed, but the U.S. employment survey selects out those who are putting pressure on the labor market. This principle may not be ideal from a welfare standpoint, but it works for some economic analyses, and in any case, what it means and does not mean is reasonably clear.

Measurement with theory helps maintain one's awareness of the gap between concept and measure and avoid use of the indicator as if it were the concept. It also provides guidance to improving the indicator and gives the option of interpreting results as due to qualities of the indicator rather than the phenomenon.

Tests for Theoretical Validity

Validity for an indicator developed with models may be tested in various ways. One should first examine the model and the justification for selecting the indicator and decide on the basis of whatever experiential or analytical knowledge one has whether it makes sense and in what situations it might not do so. Secondly, one can compare the indicator's movements with those of indicators of approximately the same concept measured in different ways (e.g., Andrews and Crandall, 1976). Employment reported by business and employment reported by individuals are cross-validating measures. Theory is required to interpret any divergence in the movements. These may or may not invalidate the measures as each corresponds to a slightly different concept or phenomenon that could be expected to behave differently at some points. Another cross-validating set of measures is the poverty line as calculated by the technicians and that identified by the public. Cross-validation regularly occurs as different analysts latch on to different indicators and come to different conclusions, which others try to reconcile. In the Forties, self-reported employment in the monthly survey did not rise as fast as industry-reported employment. This was a clue to statisticians about the invalidity of one of the measures. It was found that people were misunderstanding the questions on employment in a systematic way and under-reporting (Bancroft and Welch, 1946).

Theoretical indicators can also be validated by looking at their movements in relation to indicators of other variables when an interrelationship is presumed. Thus, one looks at relative speed, direction, and timing of changes in several linked variables to see if they correspond to predictions. Job vacancy rates or help-wanted advertising are often used as validation measures for the unemployment rate. Of course, if they do not bear the expected relation—if say, vacancy rates increase along with unemployment—one has to decide whether it is because the indicators mean something different from what one assumed or because the model relating them and therefore the prediction that job openings would lower unemployment was wrong. At least one has a clear choice.

A quasi-experimental approach is to look at how an indicator reacts to the introduction of a change in the system which one predicts would have an effect on the measured phenomenon. The designers of policy indicators cannot do what the scientist might to test the indicator under laboratory conditions, but can wait for some well-defined change to occur, such as the introduction of a new policy like a tax cut, an income maintenance program, or a mortgage policy. The most pressing question at such a time is almost invariably about the effects of the program rather than the validity of the indicator, however, and there is pressure to assume the indicator is valid. Other validation approaches may be more productive.

Standard Practices in Indicator Design

Normal practice for indicator design falls somewhere among these approaches. Most indicators are based on a little theory, some experience, and are to some extent arbitrary. The result is often inconsistent and hard to interpret. The concept the indicator is to represent is seldom specified by designers or users, and few clear hypotheses are made about the relation of one indicator to others. One cannot realistically expect indicators to be totally consistent and conceptually clear, however, because the actual practice of measurement of a complex phenomenon requires many choices for which there are neither theories nor obvious criteria. However, validation becomes more difficult the further a measure diverges from theoretical and conceptual consistency.

Ad Hoc Indicators

Such measures appear for a variety of reasons from a variety of sources. Agencies which need suddenly to back up program proposals with data will reach for the first handy figures, usually outputs of an administrative process, which seem to relate to the problem. Thus, the number on welfare may become an indicator of social disorganization, though it is in some situations more of an indicator of the job market. Moynihan's use of the percentage of female-headed families as an indicator of disorganization (U.S. Department of Labor, 1965) was more possible to validate as it grew out of theories about the function of the family in maintaining social order. The theory itself may have been invalid, and testing this, too, is part of the problem of indicator validity. On the other hand the theory is essential to give content to the concept of social disorganization and therefore to evaluate the adequacy of the indicator to represent it. Thus, we need some notion of why people go on welfare, and what social disorganization is, and how it occurs to know if the welfare rolls provide a good representation of the problem.

Indicators Produced by Data Reduction

Another type of empirical measure is the indicator developed with the aid of data reduction techniques like factor or cluster analysis. These techniques each imply a model defining the nature of the phenomenon but are often applied as mindless hunting for patterns in data or even an imposition of patterns on the data. To interpret the meaning or validate such measures the models have to be made explicit.

Composite Indicators

A third common type of measure is a composite. Recognizing that simple measures fall short of representing the full concept, many designers try to capture the elements in a single measure. Often their motivation is to find a normative indicator that will demonstrate that things are going well or badly overall without further analysis. Typical of these efforts are the attempts to build quality of "—" indices. But whether it is air, water, housing, or life itself, such measures always combine disparate things. There can be no common metric for them without a theory and set of values allowing one to combine and weight them. Moreover, the combining effort requires a use for the index which would make it sensible to reduce the complexity to a single dimension and tell us what dimension that should be. One might, for example, reduce air pollutants to amount of illness caused if one were particularly interested in public health, or to economic costs if one wanted to focus on expenditures on antipollution measures. One must decide a priori, moreover, whether its components interact or move in nonlinear ways.

Even with an a priori concept, it may be difficult to apply a clear guiding principle when the indicator is composed of many parts and the concept is a broad, open one as many social policy concepts are. Some design choices are arbitrary and some made on different principles than others. For example, the standard family budget (U.S. Department of Labor, 1967) is intended to represent a norm for family living levels, but this is a vague concept and it fits no theory that might clarify it. The quantities and qualities of the various categories of consumption items were selected on different bases. The housing standard was derived from a combination of American Public Health Association criteria and the actual available housing, the food quantities from a nutritionist-designed diet, modified to account partially for actual preferences, and clothing items were selected solely by looking at consumption patterns. Thus, in some respects the norm is what one ought to have and in some way it is what people do have. The indicator is a valid enough approximation to the norm as a criterion for setting benefit or wage levels in non-controversial cases, where all that is required is some reasonable, impersonal criterion. It has not, however, been influential in highly public policy discussion or in union wage disputes because it got neither the understanding nor trust from all participants that would have been critical.

Because of its lack of clarity and connection to a theory, moreover, the budget has been little used or analyzed by researchers in models where it might have been tested. By contrast, the unemployment rate has been accepted widely by those on opposite sides of the arguments. A detailed comparison of the cases suggests that a principal reason for the difference was the greater conceptual clarity of the latter measure and its greater connection to theory. That economic theorists were able to use the indicator in their models meant that its validity was tested in many ways, and needed improvements were identified and made. Moreover, the fact that it had such a validation process not only clarified what kinds of uses were appropriate, but also legitimated them for the public (de Neufville, 1975).

Virtually all measures have some areas of indeterminacy where the concept or

theory gives no design guide. The unemployment rate is no exception. The idea of counting available labor does not tell one how to classify the person who is looking for work but will turn down almost anything, nor about the woman who would work if someone offered her a job, but is not looking, nor about the person who has a job and is looking for another. These questions were settled in a somewhat arbitrary way in the U.S. employment survey by deciding all with no work but actively looking would be counted as unemployed. The decision was partly to provide an unambiguous criterion, and partly because of prevailing values in the 1930s about who needed work. In any measurement process, one begins with a fairly open idea (Kaplan, 1964) which gradually is defined and closed by the measurement choices one makes. In this sense, no indicator can ideally measure a concept. However, one can try wherever possible to limit this unpredictable or inconsistent component, recognizing that it makes validation and interpretation difficult.

Validity of Conceptual Frameworks

A focus on the validity of an individual indicator is insufficient, particularly for policy and program indicators. Since validity of a measure is determined with the help of a model, the model too must be valid, along with the intellectual or ideological framework from which it grows. The problem is not simply whether the model is correctly structured or whether a different model could explain the phenomenon better. It is more important whether the right question has been asked, whether the desirable solution is likely to follow, and whether explanation is sought in the right realm, with an appropriate way of arguing and using information. The problem comes down to whether the framework reflects the values and understandings of the world that should be reflected in the policy choice. Whether one believes this normative choice is a pluralist one or one to be made by a technocratic elite, the reality remains that the choice of model and framework for analysis and indicator design frames the likely policy solutions. It must therefore be an explicit part of the decision and choice.

Paradigms as Frameworks

Most indicators are developed and validated from within a paradigm, the body of problems, analytic methods, theories, and research styles that is shared by a group of practitioners (Kuhn, 1970). Understanding the paradigm is important for policy analysis because it contains the formulation of a problem which in turn has the seeds of the solution (Rittel and Webber, 1973) as well as the definition of priorities. Moreover, within a paradigm the sources and structure of the data are linked in complex ways to the model problem and the methodology and research style work together, reinforcing one another and determining when an indicator will be accepted. Exposing paradigms may be difficult, however, as they are so much a part of one's training that they are internalized and invisible.

To illustrate, let us compare what an economist and a sociologist might do if asked to produce a poverty indicator. Perhaps the economist's objective might be to explain poverty as economic welfare (Morgan and Smith, 1969). The model might find the source of poverty in the inadequacy of demand to sustain high employment

or the inability of some to participate in the labor force. The social structure and political power arrangements would be taken as given, and it would be assumed that individuals operate first and foremost to maximize their economic welfare. The economist would be likely to use aggregate national data to design and test the indicators. Policy solutions would focus on improving the capacity to participate in the labor force, or on priming the economy.

A sociologist, on the other hand, might develop a poverty indicator from within his focus on structure of the social system. He might be as interested in the consequences as the causes of poverty and would look for both in terms of roles and patterns of social interaction. Poverty would not exist in an absolute sense, but in relation to other classes, so he might define it as relative income and social status. The economic situation would be an unexplored given. The data could come from interviews and participant observation, and the results might be evaluated according to how well they meshed with theory on social stratification. The indicator itself might be based on self or group assessments of who is poor and include noneconomic criteria, such as the ability to cope, or aspiration levels. Income might be only one component of the sociologists' measure. It is a commentary on the dominance of certain paradigms in society that the economist's conception probably meshes more intuitively with popular views of poverty.

Most sets of indicators, if closely examined, reveal some of the dimensions of the paradigm that produced them. Ramsoy (1974), for example, describes how different types of health indicators reflect different frames of reference and implicitly, therefore, different policy directions. Health indicators based on the individual as patient imply a negative conception of health as absence of disease and solutions in terms of doctor-patient relations. Indicators based on population characteristics fit an epidemiological frame of analysis and a concern for prevention and solutions that may have nothing to do with diagnosis or treatment of individuals, like the building of a sewer or an advertizing campaign against smoking. An economic frame of reference measures health in terms of days lost from work and is conducive to an analysis of the costs of inability to work. The solutions are likely to focus on the most productive people and the most disabling conditions.

Functional Frameworks

An indicator serves a function or series of them within the system where design or use goes on. The nature of this relationship is part of the information relevant to the validation of a measure. One series of functions of indicators are the manifest ones, the openly stated objectives. These may be, for example, to inform someone, to focus public attention, to objectify information for discussion, to assist society to adjust to higher levels of well-being and to accumulate information and make analysis more feasible and powerful. The appropriateness or validity of an indicator is partially a matter of its ability to serve some of these functions.

Latent functions of indicators too must be considered, particularly as controversies over indicators often are simply arguments over the validity of certain latent functions. The particular formulation of an indicator can affect the likelihood of the adoption of a policy, and the probability that a specific agency will take

responsibility for it, and the composition of the group which will benefit. An indicator noting dramatic increases in substandard housing every year makes a policy of removing structural defects, repairing plumbing or even replacing the housing very likely, depending, of course, on the definition of substandard. Policies to change the housing in other ways, to improve the neighborhoods, or encourage self-help housing will not be furthered by such an indicator, but the interests of the construction trades will be, as perhaps also of housing agencies.

One may rightly respond that an indicator alone could not be the cause of such policies—that many other forces are at work. One of the most important functions of indicator design and institutionalization, however, is to get agreement on the nature of a problem so that it will be easier to adopt a policy. The definition of poverty in the U.S. has been the subject of extended political dispute because inherent in any measure will be immediately visible policy choices, such as the level of benefits and target groups as well as many more basic questions about whether poverty should be attacked by direct assistance, jobs, or social work. The discussion over the measurement of unemployment is largely due to its policy implications. So long as the measures do not include the partially employed and those too discouraged to look for work, the policies are more likely to focus on the skilled white male than on the marginal worker, and on adjusting the market rather than searching out the needy, or restructuring work.

A related concern with the function of an indicator is that it may give power to experts who produce it or know how to interpret and manipulate it. Although the fear of technocratic power (Springer, 1970) is often expressed, in practice, indicators which are institutionalized serve as weapons for those outside formal governing systems. Interest groups and the media use them effectively to criticize government policies, and legislatures set standards for bureaucratic performance which cannot readily be evaded. In general, it has been Congress that has taken the initiative in proposing new indicators and institutionalizing them over the years, while the executive branch has been actively resistant. President Franklin Roosevelt opposed an employment survey during the Depression, not wanting public criticism when unemployment was increasing nor pressure, when it was decreasing, to cut the WPA budget. Official indicators have been threatening to those in power on many occasions (de Neufville, 1975).

The question of what interests and ideologies are served by an indicator will always have a mixed answer, but it will always be relevant to validation. The basic view of this whole argument is, in any case, that one should choose the main lines of policy before the analysis instead of vice versa. This in turn raises the question of whether there may be no function for analysis or indicators. The response to this is that policy and indicators are at best iteratively developed as each gradually informs the other. Both are political processes as well as technical problems.

Contextual Validity

The indicator's model has to fit the context in which it is to be applied. This context may be either a culture or a time and includes both the circumstances and the understanding of them. The indicators and concepts they represent take on different

meanings in different contexts. In the transfer of indicators and their accompanying models across times and places an indicator can lose its meaning and become dysfunctional. Despite the pitfalls, however, formal statistical continuity and comparability are widely valued. International organizations are working to develop common indicators on which to compare nations (Christian, 1974) and already use such indicators as literacy rates, Gross Domestic Product, and percentage of population which is urban to compare development levels of nations.

Cross-cultural invalidity is the most obvious type. Divorce rates as an indicator of family stability may be valid in a country where divorce is a normal concomitant of marital break-up but invalid where there are religious or legal sanctions against it (Davis, 1965). It is hard to imagine a common operational definition of a household which would have the same meaning in every country. It could be a family, but that has different meanings in various cultures. It could be those people living in a single housing unit, but again the common definition for a unit becomes problematic when some facilities are shared and there is no common view of what facilities are necessary. Whatever definition we choose, the phenomenon would be a different one in different contexts. A country opting for rural development may find that percentage of population which is urban is a poor measure of development. The underemployment index is an example of the insidious ways the transfer of an index from one context to another may be dysfunctional. Myrdal (1968) argues that the measure of those working less than full time had a different meaning in South Asia than in developed nations. In the latter it represented a surplus labor force that could be utilized. In South Asia, however, the individuals could not normally be viewed as surplus labor as many were unprepared to do more work. However, the use of the indicator encouraged an argument that the countries could be developed economically by simply finding work for the underemployed.

A second problem of contextual validity results from changes over time within a single culture. An indicator may have to change if it is to continue to represent the intended concept. As the circumstances and understanding change, an indicator's meaning changes. Infant mortality was a valid indicator of the health of the population when communicable disease was the dominant health problem and life expectancy was low. Today in developed countries it is almost meaningless as a health index since the majority of health problems are nonfatal or are diseases of old age. The standard family budget changed over the years as the concept of income adequacy evolved (Lamale, 1958). Fresh fruit and meat became necessities as they became possibilities. With the addition of indoor central heating, clothing needs decreased, and as views of citizen responsibility evolved along with educational levels, a radio and daily newspaper became essential.

Methods to Validate Frameworks

The idea of validating a framework is paradoxical, as it requires another framework and draws us into an infinite regress. Yet for policy indicators we do have a reference point, which is in some socially or politically shared values. While those are partially conditioned by the nature of the indicators themselves, societies

do have processes for public choice, through which policies gain legitimacy and acceptance. These may be through legislatures, interest group activity, courts, or administrative action or a combination. If indicator design can be integral to these processes, the resulting measures can have a validity in the sense that they represent a shared understanding of the nature of the world and of the standards of what is acceptable. The validity is neither more nor less than that of any policy or program design made through normal processes of planning and social choice.

The difficulty is that frameworks are so much internalized by analysts that they seldom become part of the explicit discussion. Although various public decisionmakers may vaguely recognize them, they seldom can expose them without help, and their response may be to reject or misuse a measure. A beginning is to discuss frameworks explicitly as the radical economists are in posing their alternative to the neoclassical paradigm. The more valid framework is one that provides a more complete or satisfying explanation than another. Over time framework validation may be essential as the framework loses its relevance, but is a slow process and likely to meet much opposition from those who have large stakes in the old approach (Teitz, 1974). The first clue, however, to framework invalidity is apt to be in the intuitively recognized invalidity of an indicator. Since formal models and indicators as they are used and adjusted come to mesh neatly, a structured validity test may do no more than assure the consistency of the two. At the intuitive level, however, a discrepancy with experience may well be recognized first.

One model of how an indicator can be validated through a public choice process is in the U.S. unemployment indicator (de Neufville, 1975). This measure was developed in the course of a public debate about the size and nature of unemployment and what to do about it that extended from at least 1920 to 1940. A first attempt at measuring unemployment in 1930 met with tremendous public outcry because it failed to mesh with popular conceptions of how many were unemployed or who they were. The measure that statisticians finally developed for the monthly employment survey was responsive to the prevailing Depression policy concern with the unemployed breadwinner. Eminent statisticians, newspaper columnists, expert task forces, and congressmen, and organized industry and labor all played roles in the design of the measure. Later as the measure was used by the Council of Economic Advisors and the Congressional Joint Economic Committee's reports on employment and proposed policies, its fundamental assumptions were subjected to powerful validation tests. A congressional subcommittee to oversee the statistics heard technicians explain how the indicators are constructed, what they mean, and how they work. They also heard the expert staff of business and industrial organizations talk about how the indicator worked for them and criticize it for the ways they saw it as invalid. This was an educational process for legislators, interest groups, and members of the public, and all became more facile at using and interpreting as well as criticizing the measure's assumptions. Though the technical questions might not be understood by all, at least users could feel there was no deliberate deception, which was a form of validation. If the members of opposing interests who understood the measure were prepared to accept it, then the wider public could too. This validation process involved getting social and political

agreement on the measure while also getting a deeper understanding of how it worked and what it would do. It involved testing it in use and seeing whether the policies it suggested were desirable and changing it purposively when they were not.

The net result of this process of use, discussion, and adjustment was that the unemployment indicator acquired a socially ratified validity, and a validity in the sense that it worked for many purposes. A scathing article in *The Reader's Digest* (Daniel, 1961) finally challenged the measure, questioning the concepts and methods and suggesting they were created for the prime purpose of promoting particular policies. The article caused such a stir that President Kennedy set up a committee to investigate the data and two sets of congressional hearings were held. The opposing interests and experts came unanimously to the indicator's defense and silenced further discussion. The key features of this process were its openness to the public, the participation by informed interests, professional and experts, the adversary nature of the discussion, and the institutionalized process for getting feedback from the indicator users and for helping them to understand its meaning. Since the indicator was regularly used in analysis and economic policy decisions, there was ample opportunity to see how it worked, and what it meant as well as to ensure that it meshed with the understanding and values that prevailed at the time. All of these are essential ingredients of validation of policy indicators.

Recommendations

As a practical matter, how can one deal with this elusive problem of validity? If we accept that we will never achieve perfect validity and aim simply for improvements, there are realistic things that both producers and users of indicators can do.

Indicator designers can follow several principles. They can, wherever possible, measure with theory and make their own paradigms and the models connecting indicator and concept explicit. Even relatively simplistic models will help the user of the indicator to know what it might mean and under what circumstances, and when odd results may be explained by a faulty model. Simple indicators with a clear organizing principle and comprehensible meaning can often be most easily linked with a concept. The designer should test the indicator's movements against other indicators of the same phenomenon or indicators of phenomena whose movements have a predictable relation to the one under study and provide an interpretation of discrepancies between the indicators' behavior in relation to each other and their expected behavior. Where such tests are not possible, the designer can provide the user with some idea of how the indicator may be tested over time—to what it might be compared and what expectation one might have for its behavior if it is valid. The designer can also make clear when an indicator has not been validated.

Governments setting statistical policy or setting up indicator systems should not be overly concerned with centralization of statistics, elimination of overlap and consistency definitions. The decentralized approach of the U.S. (Feldman, 1971) means that many of our indicators are valid for particular uses in ways they would not be if they came out of a central, all-purpose statistical agency. Moreover, the redundancy and inconsistency has the positive function of allowing validation by

comparison and in relation to alternative frameworks instead of a single one. The borrowing of indicators from other countries without understanding their logic and underlying models and testing these against the local realities is risky. By the same token, public policy should not depend on comparisons with measures in other nations unless more than spurious comparability can be assured.

Statisticians often place demands and constraints on the statistical system that may not be in the best interests of governments. Formal continuity of construct in an indicator series may preclude continuity in content as societies evolve and the intellectualizing about them changes. In the interests of objectivity it is often argued that data should simply be presented in tabular form and data collection methods described. The user cannot, however, validate measures without models. One could argue that most official statistics are not indicators; they are not designed to reflect particular concepts in a particular context, but are for a range of purposes. Measures like numbers of housing units or median population age are perhaps accurately described this way as they are intuitively clear and can fit many models, but most government statistics are more pointed than these and inevitably become indicators. Housing is given by number of people per room, employment rates by race, and wage levels by sex. The choices represent distinct views about the nature of problems and priorities. The new national social indicator publications are even more deliberate (e.g., U.S. Department of Commerce, 1977; Canada, 1977). And indeed policy is the object. On the whole official statisticians are not unaware of their political role despite the lip service they may pay in public to their value-neutral position. Nor are politicians unaware of the policy implications in accepting an indicator.

U.S. statistical policy has been a compromise. Statisticians remain insulated from politics by not discussing the implications of the data. It would be difficult for the public to deal with the idea that the statisticians make choices that affect policy. Experience has shown that indicators are more likely to be accepted as honest if presented without comment. Ironically, however, if public expectations about the value-free objectivity of data could be changed, validation would be more possible. To take the edge off this arbitrary power of statisticians, governments can develop a process of designing indicators that is open and legitimate so the models can be chosen publically. The difficulty is that to do this may be seen as an abrogation of governmental control. Moreover, public validation of indicators is costly and it is only worthwhile in proportion to the consequences of its use.

Those who use indicators should keep in mind that validity of an indicator depends on context, and involves the kind of question they are asking, the nature of the problem, and the understanding that they have.

The users need to be very clear on the conceptual framework and the models used in the indicator design, and then to use whatever information, analytic method, or experience they have to link it to their own model and objectives. The indicator does not have validity in itself but only insofar as it matches the user's understanding and purpose. Users, too, are part of the validation process, and when an indicator does not do what they want or expect, they should complain and comment to those who produce it. The user's experience is critical to learning the meaning of an indicator and thus to validation.

Conclusion

Validation of policy indicators will never be complete and fully systematized, but it is not an empty exercise. Rather it is an essential part of exposing the meaning of indicators and making communication possible among groups who are part of policymaking (Johnston, 1978). Validation of indicators depends on analysis, judgment, intuition, professionally acquired understandings and skills, feedback from applications, and social and political processes. It will be important in the coming years to improve and build upon these methods of determining the validity of indicators and to assure that more people recognize and practice them. Dealing with the issue is critical to legitimate, ethical policy analyses.

REFERENCES

- Andrews, Frank M., and Rick Crandall (1976). "The Validity of Measures of Self-Reported Well-being," *Social Indicators Research*, 3: 1-19.
- Bancroft, Gertrude, and Emmett H. Welch (1946). "Recent Experience with Problems of Labor Force Management," *Journal of the American Statistical Association* 41: 303-312.
- Berger, Peter, and Thomas Luckman (1966). *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Garden City, New York: Doubleday.
- Blalock, H. H., Jr., and A. B. Blalock, eds., (1968). *Methodology in Social Research*. New York: McGraw-Hill, chapters 1 and 2.
- Bunge, Mario (1975). "What is a Quality of Life Indicator," *Social Indicators Research*, 2: 65-79.
- Canada, Statistics Canada (1977). *Perspective Canada II: A Compendium of Social Statistics*, Ottawa.
- Christian, David E. (1974). "International Social Indicators: The OECD Experience," *Social Indicators Research*, 1: 69-186.
- Daniel, James (1961). "Let's Look at Those 'Alarming' Unemployment Statistics," *Reader's Digest*, September.
- Davis, Kingsley (1965). "Problems and Solutions in International Comparison for Social Purposes," *America Latina*, 8, June-March.
- de Neufville, Judith I. (1975). *Social Indicators and Public Policy: Interactive Processes of Design and Application*. Amsterdam: Elsevier Scientific Publishing Company.
- Feldman, Paul (1971). "Commissions on Statistics, Statistics on Commissions," in *U.S. President's Commission on Federal Statistics*, Report, Vol. 2. Washington, D.C.: U.S. Government Printing Office.
- Johnston, Denis F. (1978). "Postlude, Past, Present and Future," *Annals of the American Academy of Political and Social Sciences*, 435: 286-294.
- Kaplan, Abraham (1964). *The Conduct of Inquiry: Methodology for Behavior Sciences*. San Francisco: Chandler Publishing Company.
- Koopmans, Tjalling (1947). "Measurement without Theory," *Review of Economics and Statistics*, 24: 161-192.
- Kuhn, Thomas (1970). *The Structure of Scientific Revolution*. Chicago: University of Chicago Press.
- Lamale, Helen H. (1958). "Changes in Concept of Income Adequacy Over the Last Century," *American Economic Review*, 48 (2): 291-299.
- McIntosh, William Alex, Gerald Klogan, and Leslie D. Wilcox (1977). "Theoretical Issues and Social Indicators: A Societal Process Approach," *Policy Sciences*, 8: 245-268.
- Myrdal, Gunnar (1968). *Asian Drama*. New York: Pantheon.
- Pitkin, H. F. (1972). *Wittgenstein and Justice*. Berkeley: University of California Press.

- Polanyi, Michael (1969). *Knowing and Being*. Chicago: University of Chicago Press.
- Ramsøy, Natalie R. (1974). "Social Indicators in the United States and Europe: Comments on Five Country Reports," in *Social Indicators 1973: A Review Symposium*. Washington D.C.: Social Science Research Council.
- Rittel, Horst, and Melvin Webber (1973). "Dilemmas in a General Theory of Planning," *Policy Sciences*, 4: 155-169.
- Springer, Michael (1970). "Social Indicators, Reports, and Accounts: Toward the Management of Society," *Annals of the American Academy of Political and Social Science*, 388, March.
- Teitz, Michael B. (1974). "Toward a Responsive Planning Methodology," in *Planning in America, Learning from Turbulence*, D. Godschalk, ed. Washington, D.C.: American Institute of Planners.
- Torgerson, Warren (1958). *Theory and Methods of Scaling*. New York: John Wiley and Sons.
- Tropman, John (1976). "The Social Meaning of Social Indicators," *Social Indicators Research*, 3, (3/4): 371, December.
- U.S. Department of Commerce (1977). *Social Indicators 1976*, U.S. Government Printing Office.
- U.S. Department of Health, Education, and Welfare, Poverty Studies Task Force 1976. *The Measure of Poverty*. U.S. Government Printing Office.
- U.S. Department of Labor (1965). *The Negro Family: The Case for National Action*. Washington, D.C.: U.S. Government Printing Office.
- U.S. Department of Labor, Bureau of Labor Statistics (1967). *Three Standards of Living for an Urban Family of Four Persons*, Bulletin No. 1570-5. Washington, D.C.: U.S. Government Printing Office.