MEASURE *Evaluation* Bulletin

Number 3, 2001

Monitoring Population and Health Program Efforts with Composite Indices

HIGHLIGHTS FROM THIS ISSUE

- Composite Indices to monitor national program efforts are now available for family planning, maternal and neonatal health, and AIDS.
- Composite national indicators have commonly been used for international comparisons in many social and economic fields (e.g., Human Development Index, economic indices).
- A solid research foundation is necessary to enhance the credibility and long-term utility of national composite indices.
- The Family Planning Effort Index has been successfully used for three decades, and five global rounds of data collection have shown gradual improvement over time.
- A Maternal and Neonatal Program Effort Index has been developed and was used in 49 countries in 1999.
- An AIDS Program Effort Index aims to measure political commitment and program effort in HIV/AIDS and is likely to play an important role in monitoring global and national efforts to expand the response against AIDS.

Previous Issues of the MEASURE Evaluation Bulletin:

1. Monitoring the Quality of Care in Family Planning
2. Indicators for Monitoring and Evaluation of AIDS Programs

The MEASURE *Evaluation* Bulletin is published four times a year by the MEASURE *Evaluation* Project. Each issue addresses a specific theme and the contents are summary papers based on research and technical assistance supported by MEASURE Evaluation.

The MEASURE *Evaluation* Bulletin is made possible by support from USAID under the terms of Cooperative Agreement HRN-A-00-97-00018-00. The opinions expressed are those of the authors, and do not necessarily reflect the views of USAID.

Monitoring National Progress with Composite Indices

Amy Tsui

- √ **National composite indices are designed to enable comparisons across countries, regions, provinces or systems, and are used in many fields such as economics, human development and health.**
- √ **The credibility and long-term utility of national composite indices depend on a solid research foundation.**
- √ **In the development of well-known indices such as the Human Development Index, leading economic indicators and the Family Planning Program Effort Index, careful attention has been given to issues of validity and reliability and weighting of the components.**

This bulletin shares findings from recent assessments of national effort on family planning, maternal and neonatal health and HIV/AIDS. The composite indices are the Family Planning Effort Score, the Maternal and Neonatal Health Program Index, and the HIV/AIDS Program Effort Index. The first is well established, while the latter two are newly developed. First, we provide some background on the general use of composite indices for monitoring national progress.

National composite indices used for monitoring development trends share several common features. They are designed to enable comparability across aggregate units, such as countries, provinces or systems, as well as time. They are usually multidimensional in nature and involve numerical scoring to permit national ranking. Also, in addition to being able to measure current conditions of national interest, some composite indicators, such as leading economic indicators, have been developed to forecast future levels.

As national rankings, composite indices have the power to provoke debate, and, some would argue, induce policy change. Ranking national performance by an index serves to raise public awareness about the developmental need and to motivate officials to address deficits in infrastructure or resources. Benchmarking national progress by setting target values for indices can draw policymakers' attention to welfare issues by focusing their discussion on the content and measurement of the indicators. Disaggregating the values of a composite indicator and its constituent components to apply to subnational areas can engender a programmatic response from lower level administrative units. These summary measures thus serve a distinct purpose of focusing international and national discourse and organizing resource allocation around efforts to improve the social and human condition.

Examples of National Composite Indicators

Among the more well-known composite indices to chart international and national progress on various aspects of human welfare are the United Nations (UN) Human Development Index (HDI) and Human Poverty Index (HP-1) for developing countries.[1,2] These have been joined recently by two other UN indices – the Gender-related Development Index (GDI) and Gender Empowerment Measure (GEM). As composite indices, the HDI, HP-1, GDI and GEM are, by definition, each constructed from base indicators. The underlying logic is that use of more than one indicator will improve measurement reliability of the construct, i.e., human development, poverty, gender equity and gender empowerment. Thus the HDI is a weighted composite of life expectancy, educational attainment, and adjusted Gross Domestic Product (GDP) measures, while the Human Poverty

Index is a separately weighted measure of deprivation on the same three dimensions. The GDI also uses the three measures – longevity, knowledge, and standard of living – to assess disparity in achievement for women and men. In contrast, the GEM employs indicators of gender presence in administrative, professional/technical and parliamentary positions to gauge the relative empowerment of men and women in political and economic spheres of activity.

In the economic sector in the U.S. and other industrialized countries, composite indices have a long history of research and application. For example, the Leading Indicators Index, which captures time series data on business cycles to enable anticipation of future downturns or recoveries, is built on ten key economic series, such as the nation's money supply, vendor performance, unemployment insurance claims, new manufacturing orders and consumer expectations. Financial market indices, such as the Dow Jones Industrial Average or Standard & Poor's Composite Index, track the general health of a selected group of stock prices; equivalent indices have been constructed for countries in Europe and Asia with significant market economies. The Gini Coefficient is another measure of economic welfare used to track income inequality in a population and is calculated by many government agencies for local areas. The Pan American Health Organization has prepared a Gini Coefficient and Concentration Index to track health inequalities to determine where its technical assistance is needed.

A number of organizations have sponsored composite assessments of national well-being in other domains. For example, the Heritage Foundation's Index of Economic Freedom surveys a range of economic policy measures for 161 countries, and the Freedom House's Annual Survey of Freedom in the World examines political rights and civil liberties in a similar set of countries. Population-level indicators of subjective perceptions about the quality of life and lifestyle satisfaction are also quite prevalent. On the health front, the World Health Organization's studies of the global burden of disease have led to the development of the Disability-Adjusted Life Expectancy (DALE) measure. The DALE is the expected number of years to be lived in full health equivalents, that is, overall life expectancy downwardly adjusted for years of ill-health weighted for severity. Moreover, WHO has recently calculated Health System Performance measures for 96 countries[3] to compare the efficiency of their health systems in translating health expenditures into health, measured by the DALE. On the reproductive health front, Population Action International has ranked 133 countries on a Reproductive Risk Index, composed of ten indicators of reproductive health, such as adolescent fertility, HIV/AIDS prevalence among adult males, the total fertility rate, and the maternal mortality ratio.[4]

Methodological Issues

The use of composite indices has not been without criticism. Complaints can range from the indices being conceptually flawed to being inaccurately or unreliably measured to the rankings being inappropriately interpreted. As a result of such criticisms, social and economic indicators research has matured as a science, leading to continual indicator adjustments and a diversity of composite indices. Because methodological issues are not trivial to the construction of composite indices, it is worth reviewing several related to validity, reliability and weighting here.

Validity

Fundamental to the utility of any index or composite indicator is whether it accurately represents and measures the construct of interest. How well, for example, is national poverty measured in terms of the selected base indicators of adult illiteracy, lack of health service access, lack of safe water, child malnutrition, maternal mortality and premature adult mortality? Should other indicators be added into the mix, such as lack of adequate housing, inadequate adult nutrition, or unemployment? To answer this, one will need to address three validity issues:[5]

- *Content validity* - the extent to which the indicator adequately represents the concept
- *Criterion validity* – the extent to which the indicator predicts or agrees with the criterion indicator, such as comparison against a "gold standard"
- *Construct validity* – the extent to which relationships between indicators agree with relationships predicted by theories or hypotheses

All three validity issues are relevant to the foregoing question and perhaps content and construct validity even more so, given the multidimensionality of the Human Poverty Index. To uphold a composite index's content validity, it is important to guide its construction and refinements with a range of theoretical insights. Relevant theory from the behavioral, physical, biomedical or social sciences can be brought to bear in clarifying the conceptualization of human poverty. As multi-dimensional measures, composite indices are often heavily informed by the disciplinary or professional preferences of their developers. In the case of human poverty, a more clinically inclined index developer might argue to base or augment the underlying dimensions with nutritional and morbidity status measures.

If there is a gold standard criterion for human poverty, the index values should be judged against this to establish criterion validity. However, unlike their clinical counterparts, global constructs of development often lack objectively verifiable measures and standards. For example, laboratory tests of the presence of a viral or bacterial infection enjoy relatively unambiguous criteria for validating their measurement ability. A li-

BOX 1
WEIGHTS IN INDEX CONSTRUCTION

Indices such as the HDI, HP-1 or DALE are based on substantial research into the relationships among the constituent indicators and their relative importance for measuring the construct of interest. Their relative importance is reflected in numerical weights assigned to each prior to being combined into a total score. For example, suppose we are interested in constructing a composite indicator of national health and conceptualize its constituent indicators to be infant mortality, adult mortality, annual per capita income, percentage of public expenditures on health, and proportion of the population with access to good primary health care (PHC). What mathematical function best captures the relationship of the five indicators to the overall construct of national health? Should the composite index be constructed as a weighted sum of the five indicators? Or should the five be equally weighted in composing this National Health Index (NHI)? A background study of the variation in the values of these indicators across countries and over time might suggest that instead of equal weights (1.0), the following weights be assigned: -0.35 to infant mortality, -0.05 to male adult mortality, 0.15 to per capita income, 0.20 to health expenditures, and 0.35 to the PHC coverage. Using hypothetical indicator values for country X, we would have:

NHI = -.35 x (IMR=42) - .05 x (Annual death rate per 1000 men > 45 years = 12] + .15 x (logged pc GNP = $750) + .20 x (% budget on health=10) + .35 x (% population with PHC = 65)

or = -14.7 - 0.6 + .288 + 2 + 22.75 = 9.74

By itself an NHI value of 9.74 is not very meaningful. However, suppose optimum conditions define the maximum index value to be 35.53, then country X's NHI score would be 27.4 (9.74/35.53 x 100) of the maximum health standard, using a 0 to 100 point range. Assessing other countries' values on the constituent indicators would produce a range of NHI scores that could eventually be ranked.

gase chain reaction (LCR) based assay to detect *Neisseria gonorrhoeae* DNA in a biospecimen can be rated in terms of its sensitivity and specificity as to how well it matches culture-based test results for the same. The concept of poverty, defined in human or social terms, however, might not be seen as having a physiological criterion to use as a gold standard criterion. One could consider, though, testing the social measures against anthropometric, physiological and physical measures of malnutrition, stress, illness and environmental hygiene if one believes they serve as valid criteria of poverty.

Construct validity should, whenever possible, be confirmed with statistical methods, such as confirmatory factor and latent structure analysis. These methods are based on correlational analyses and examine the interrelationships among a set of variables or measures to explain them in terms of a limited number of unobserved (latent) variables. For example, the question of whether the number of dimensions to the Human Poverty Index should be augmented can be tested with appropriate data and confirmatory factor analysis. It is quite possible that because national factors are often highly correlated with each other, HP-1's present formulation has both optimal content and construct validity.

Reliability

Reputable time series data are the preferred data sources for indicator measurement and index construction. The reliability of a composite indicator's measurement will be heavily influenced by the quality of the data used to measure each constituent indicator. Often measuring these constituent indicators requires extant and objective data, that is, data that have been collected and published or made publicly accessible by professionally responsible parties, such as infant mortality rates, primary health care access, or government health budgets provided by ministries of health or national statistical offices. Three relevant reliability issues are

- Test-retest reliability – Extent to which there is a high correlation between measurements taken at different points in time

- Inter-rater reliability – Extent to which measures obtained by different raters for the same concept are highly correlated
- Internal consistency reliability – Extent to which a subject's responses to items related to a common concept are highly correlated

Continuing to the use of the human poverty construct to illustrate these issues, we might be concerned if the test-retest reliability of the Human Poverty Index degrades because two constituent indicators, such as the lack of safe water or health service access, are not measured uniformly over time. The minimum standards for access, for instance, as proportion of population covered, might be independently altered over time. Similarly, if the measures over time are drawn from different sources with different methods of standards rating, the HPI's reliability can be compromised. For instance, child immunization levels are often extracted from health information systems, special immunization surveys, and maternal reports in population surveys. Highly variable estimates of a single indicator, such as the percentage of children fully immunized by age 1, can arise because of different measurement methods.

Last, in terms of internal consistency, this reliability issue is perhaps most germane to the three composite indices described in this bulletin. All three indices rely on a select sample of knowledgeable informants to rate a large number of items gauging national effort on family planning, maternal and neonatal health, and HIV/AIDS. How internally consistent are the raters within the subset dimensions to the indices? For example, when an individual rater judged a set of items for maternal and neonatal health policy, how correlated were his or her responses, as opposed to how correlated were various raters' scorings. Some background analyses by the index developers suggest relatively high consistency in respondents' answers on item sets for some areas and less for others, as might be expected.

Poor data quality and measurement error obviously pose significant threats to the validity and reliability of an index. Indicator developers should always address the various types of measurement error and detail efforts taken to minimize them.

Weighting components

Because composite indices are constructed from individual base indicators, another type of question that will arise is how should these first be combined into a summary score and then how should the items be internally weighted? Should the base indicators be linearly or nonlinearly combined? How important for measuring reproductive health risk, for example, is it that the maternal mortality ratio and HIV/AIDS prevalence have the same weight in an overall score as the total fertility or adolescent fertility rate? Should any one of these constituent indicators be accorded greater or lesser importance in their contribution to the summary score? Deciding on the weights requires statistical analysis of the correlations among the indicator variables. Factor analysis in general, and confirmatory or exploratory factor analysis in particular, have been commonly employed methods. Box 1 illustrates how weights are applied to base indicators to produce a composite index.

Conclusion

It should be clear that the credibility and long-term utility of national composite indices depend on having a solid research foundation. The well-known ones like those in the HDI family, the Human Poverty Index, leading economic indicators, or Global Burden of Disease measures have such a foundation. The three indices discussed next have benefited from the same in their development.

Notes

[1] United Nations Development Program. 2000. Human Development Report 2000. Oxford University Press, New York.

[2] C. Kaul and V. Tomaselli-Moschovitis. 1999. Statistical Handbook on Poverty in the Developing World. Oryx Press, Phoenix, AZ.

[3] World Health Organization. 2000. The World Health Report 2000. Health Systems: Improving Performance. World Health Organization, Geneva, Switzerland.

[4] For on-line access to some these indices, see:
Index of Economic Freedom – *www.heritage.org/index*
Freedom in the World Index – *www.freedomhouse.org/ratings/index.htm*
Disability-Adjusted Life Expectancy – *www-nt.who.int/whosis/statistics/dale/dale.cfm?path=statistics,dale&language=english*
Reproductive Health Risk Index - *www.popact.org/resources/publications/worldofdifference/rr2_introduction.htm*

[5] The discussion on validity and reliability is drawn from the following: L. Aday, 1991. Designing and Conducting Health Surveys. Jossey-Bass Publishers, San Francisco and J. Bertrand and A. Tsui, 1995. Introduction to Indicators for Reproductive Health Program Evaluation. The EVALUATION Project, Carolina Population Center, University of North Carolina at Chapel Hill.

Three Decades of Tracking National Family Programs through the Family Planning Effort Index

John Ross and John Stover

- √ **The Family Planning Effort Index has been a very useful way to track national program efforts for the last three decades.**
- √ **Family planning programs have gradually become stronger throughout the world since the seventies, with the weakest programs improving most.**
- √ **Since the 1994 Cairo conference effort scores have further improved in most countries.**

The first family planning effort (FPE) scores were collected in 1972. A standard methodology has been used since 1982. This methodology uses scores on 30 features of country family planning programs that are calculated from questionnaire responses. These FPE questionnaires are completed by a select group of experts who are active in, or very familiar with, that country's policies and activities. Four types of expert respondents are used: (1) local officials implementing the FP program, (2) donor staff close to program operations, (3) knowledgeable nationals not specifically involved in program or policy management, and (4) knowledgeable foreigners.

The questions are designed to capture program inputs, gauge their strengths and weaknesses, and measure improvement over time in the relationship between effort and outcomes. The questionnaire includes about 120 items, which yields a set of 30 scores. Each one ranges between 0 and 4, with 0 being no effort and 4 strong effort. The 30 features are grouped into four areas: policy environment (8 questions), services and support (13), evaluation and records (3), and access or method availability (6). An overall index compiling all thirty scores (maximum total of 120) represents the general level of family planning effort in evidence in that national setting. Scores are generally presented as percent of the maximum, e.g., 80 points equals 67% of the maximum.

1999 Country Results

Total scores in 1999 range from a low of 29 to a high of 86. Six programs have total scores of 75 or above: China, Indonesia, Taiwan, Vietnam, Thailand, and Mexico, all of which are generally recognized for the strength of their family planning programs. These six, and others at the upper end of the range, generally score well on all four components. At the lower end of the range, seven countries have total scores of 35 or below: Sudan, Congo, Gabon, Uruguay, Costa Rica, Argentina, and Venezuela. Most of these countries scored well on at least one component but very poorly on others.

Strength Categories

In previous rounds, programs have been classified into four broad categories of effort on the basis of the percentage of the maximum possible score:

- Strong: 67% or higher
- Moderate: 46-66%
- Weak: 21-45%
- Very weak/none: 0-20%

According to this classification, programs in 13 countries are "strong," programs in 53 countries are "moderate" and 23 are

Figure 1. 1999 FPE Index Results, by Strength Category

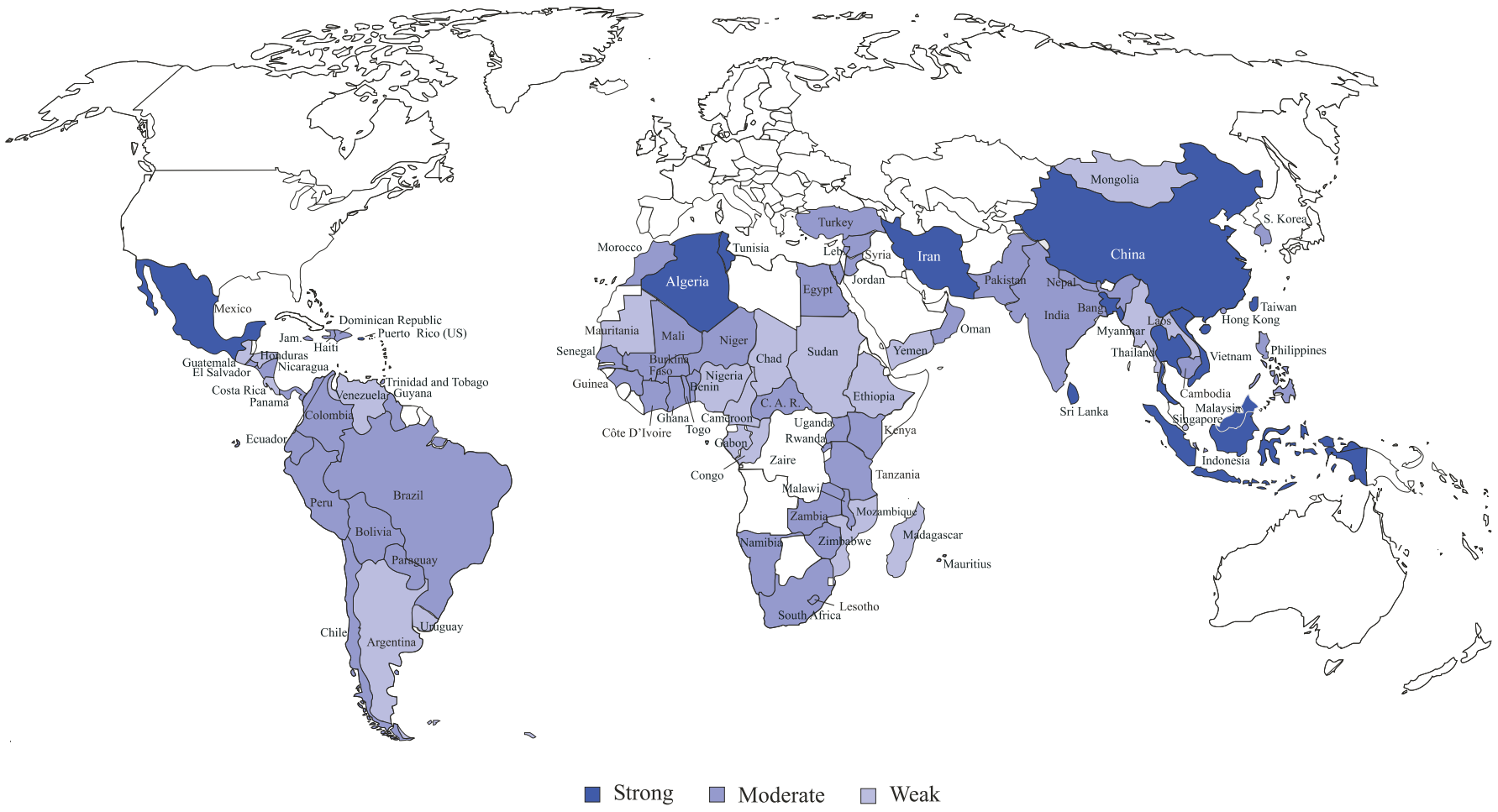
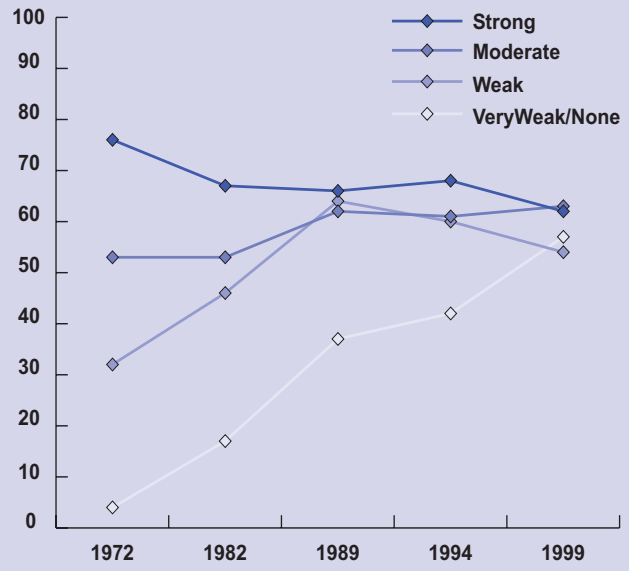


Figure 2. Increases in Effort over Time, by 1972 Effort Cohorts



"weak" (Figure 1). No countries were classified as "very weak/none" in 1999. As a group, the weaker countries need to improve in almost all features, not just a single one, in order to move into the stronger categories.

Trends by Strength Category

In 1972 over 60 countries were classified as very weak/none. Over the years more and more countries have instituted policies and programs and worked to improve them. By 1989 most countries had moved out of the lowest category and joined the weak or moderate categories. Between 1994 and 1999 the weak group lost members to the moderate group, so that by 1999 the largest number of countries is found in the moderate category. There has been very little change over the years in the number of countries classified as having strong programs, but the transition in the number of countries classified as very weak/none, weak, and moderate has been striking. By 1999, no countries were classified as very weak/none and only 19 were weak.

Another way of displaying the trends appears in Figure 2. It keeps all countries together according to their classification in 1972. The average score for countries classified as "strong" in 1972 dropped by 1982 but then only slightly to 62 (top line). Those classified as moderate in 1972 have since increased their average score somewhat from 53 to 62. The largest changes appear in those countries originally classified as weak or very weak/none. They increased their scores dramatically over these 27 years to within 10 points of the higher categories. The average score for all countries increased from 20 in 1972 to 54 in 1999. The dominant trend has been for the weak group to rise toward the strong group. By 1972 the strong group was already at a high level and has remained there over the past 27 years.

Results and Trends by Region

The regional averages are shown by the four program components for 1999 in Figure 3. The widest variation in scores clearly occurs in method availability. The range is only 15-20 points for the components of policy, services and evaluation, but over 50 points separate the lowest region (Francophone Africa) from the highest region (East Asia) in method availability. Most regions now have policies in place and have programs with important elements of service delivery and evaluation. However, the implementation of these programs, to actually deliver methods to the population, sharply differentiates the high-effort countries from the low-effort ones. A relatively full choice of methods is available to those living in most East Asia countries, while many programs in sub-Saharan Africa provide less choice and reach only certain segments of the population.

Trends in overall program effort are shown by region in Figure 4. All regions show improvement since 1972, although East Asia declined from an earlier peak to an average FPE of 64 in 1999. However, at 64 East Asia still shows the highest average score. Latin America seems to have reached a plateau of 51-50% of the possible maximum score, although as in all regions there are major variations. In 1999, for example, Mexico had a total score of 75, but Venezuela only 29. Africa and the Middle East are still climbing, albeit starting from extremely low effort scores.

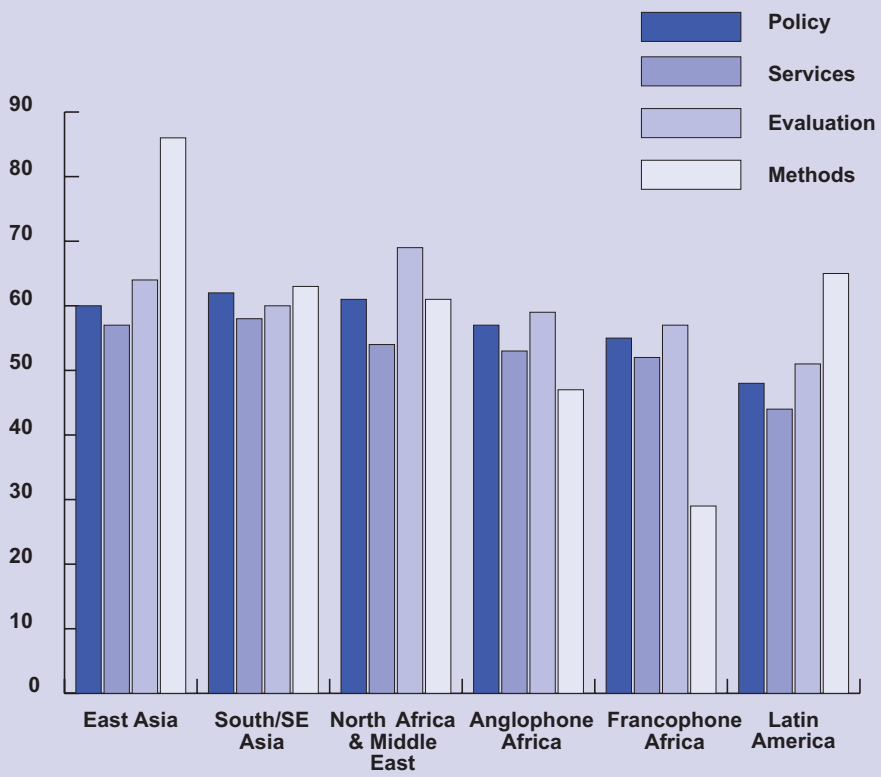
Trends by Population

The picture is different on a population basis, and more favorable. While in 1972 only 36% of the population in surveyed countries lived in countries listed as strong, by 1982 that had increased to 62% and by 1999 to 68% (Figure 5). Only 6% of the population in surveyed countries lived in countries that had programs scoring weak in 1999, and none scoring very weak/none, compared to 29% in 1982.

FPE and Social Setting

There is ample historical and contemporary evidence that fertility levels decline if a country becomes more socio-economically developed. There is also plenty of current evidence that strong family planning programs are associated with a decline in fertility. An important question for family planning programs is what difference such programs have made net of improvements in social setting. A comparison of contraceptive prevalence rates with FP program efforts (as periodically measured

Figure 3. FPE Index Scores by Component and Region, 1999



by FPE index scores) and the social setting (measured by a different index, composed of educational, economic, and health variables) suggests a strong association with both factors, as shown in Table 1. Average contraceptive prevalence is highest in countries with high FPE Index scores and high Social Setting Index scores. Absence of a conducive social setting seems to have a somewhat more negative effect on the contraceptive prevalence than lack of a strong FP program effort, as the contraceptive prevalence means in the "Low" social setting category tend to be much lower than the means in the "Low" FPE category (overall, 16% compared to 29%). Where FP Effort may have peaked, therefore, further improvements in reducing fertility rates could certainly continue to occur to the extent that education continues to reach increasing numbers of people, economic productivity and growth continue to improve lives, and amelioration of health services and outcomes continues to improve the social setting in many countries.

Conclusion

Globally, family planning effort continued to strengthen during the last five years, improving by about one-eighth over the 1994 level. However, this brings the average country score to only 54% of maximum, which leaves a great deal of room for further improvement. Nevertheless the strongest programs have never risen much above 80% of maximum, which raises the question of what can reasonably be expected. Against the standard of 80%, the 54% score in 1999 represents two-thirds of what appears to be the top range of the effort scores. Moreover, on a population basis, the picture is more favorable, since most of the developing world's population lives under programs in the stronger categories and a small percentage, compared to early years, are living in countries with very weak or no programs.

Notes

From the MEASURE *Evaluation* Working Paper, No. 20, "Family Planning Effort: Scores and Trends," by Ross and Stover, May 2000.

Figure 4. FPE Effort Score by Region of the World, 1972-1999

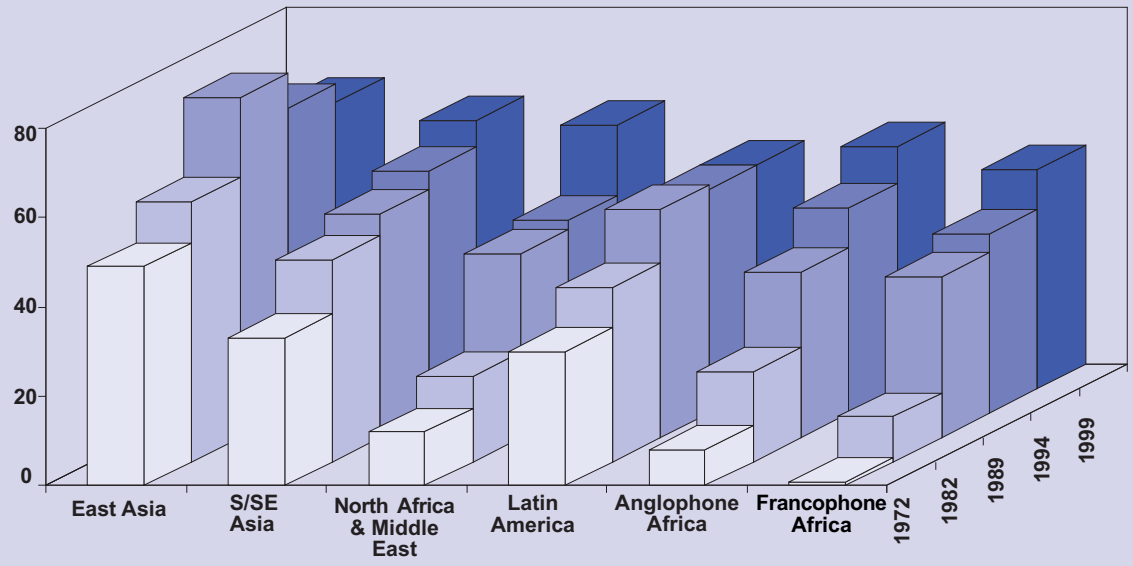


Figure 5. Percentage of the Population Living in Indexed Countries, by FPE Index Strength Categories

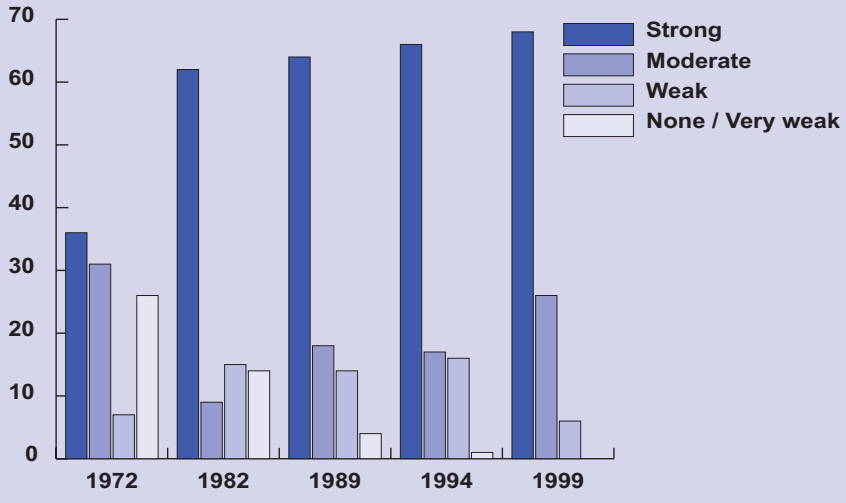


Table 1. FPE Index Categories, Social Setting Categories, and Average Contraceptive Prevalence Rates (CPR)

FPE Index:	High	Upper middle	Lower middle	Low	
Social Setting					**Mean CPR by Social category**
High	Mean CPR: 73 Korea, Jamaica, Colombia, Cuba, Mauritius, Mexico	Mean CPR: 67 Brazil, Panama, Singapore, Trinidad		Mean CPR: 49 UAE, Kuwait, Costa Rica	65
Upper middle	Mean CPR: 62 Syria, Iran, Sri Lanka, Thailand, Tunisia	Mean CPR: 57 Nicaragua, El Salvador, South Africa, Algeria, Turkey, Egypt, Philippines, Dominican Republic, Peru	Mean CPR: 54 Jordan, Paraguay, Honduras, Oman, Namibia, Ecuador	Mean CPR:38 Iraq, Mongolia	53
Lower middle	Mean CPR: 57 Morocco, Zimbabwe, Botswana, India, Vietnam, Indonesia, China	Mean CPR: 23 Senegal, Lesotho, Pakistan, Ghana, Kenya	Mean CPR: 25 Cote d'Ivoire, Nigeria, Zambia, Cameroon, Guatemala, Bolivia	Mean CPR: 36 Papua New Guinea, Congo, Myanmar, Gabon, Mauritania	36
Low	Mean CPR: 31 Rwanda, Togo, Bangladesh	Mean CPR: 14 Tanzania, Mali, Nepal, Guinea	Mean CPR: 13 Ethiopia, Benin, Haiti, CAR, Niger, Malawi, Uganda, Burkina Faso	Mean CPR: 13 Sudan, Laos, Yemen, Chad, Cambodia, Bhutan, Madagascar, Mozambique	16
Mean CPR by FPE index category	60	45	28	29	**Overall Mean CPR: 41**

Rating Maternal and Neonatal Health Programs in Developing Countries

Rodolfo Bulatao and John Ross

- √ **Building upon the success of the Family Planning Effort Index, the Maternal and Neonatal Program Effort Index (MNPI) was developed.**
- √ **The MNPI questionnaires were completed by more than 1,000 expert raters in 49 developing countries in 1999.**
- √ **Overall, the MNPI appears to yield useful measures for various aspects of program effort, which are consistent with external data.**
- √ **Comparison of the MNPI ratings in 1999 with three years before suggests improvements over the last three years in almost all categories.**

Little has been done to measure country program efforts in maternal health. Building upon the success of the Family Planning Effort Index, maternal and neonatal health services in 49 developing countries were rated in 1999 by experts in each country in order to provide a comparative assessment [1]. This assessment covers preventive and curative measures, as well as the infrastructure required for service provision. The expert ratings are summarized as the Maternal and Neonatal Program Index (MNPI).

How is the MNPI constructed?

Each expert was asked to rate the national maternal and neonatal health program in one country on an 81-item questionnaire. The items refer to many related areas: essential obstetric services, antenatal care, newborn care, family planning, control of sexually transmitted infections, etc. Items on maternal and neonatal health policy and related health promotion, training and research were also included.

Experts rated services on a scale from 0 to 5, where 5 was meant to indicate that a statement was "completely true." The opposite end of the scale, a rating of 0, was meant to indicate that it was "completely false." In reporting results, these ratings are multiplied by 20, so that they run from 0 to 100, where 100 represents maximum effort on an item. In rating access to services, raters were asked directly to indicate the percentage of pregnant women with adequate access to each service. In addition to rating current programs, respondents produced similar, retrospective ratings of programs as of three years previously (effectively in 1996).

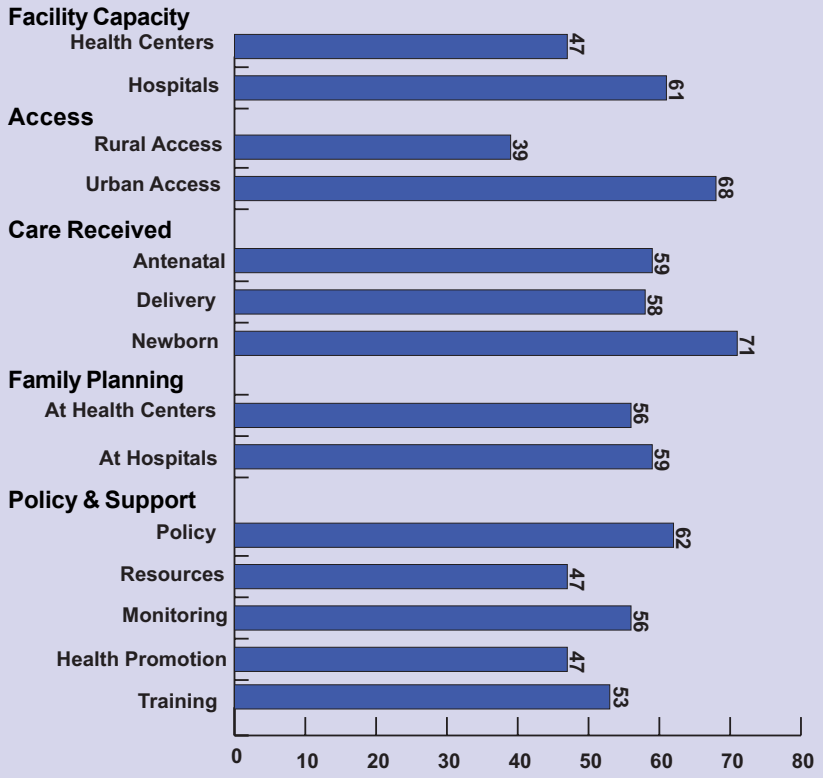
The MNPI assessment in 1999 covered 49 countries, including 21 in sub-Saharan Africa (divided into Francophone and Non-Francophone), 13 in Latin America and the Caribbean, 10 in Asia (excluding the Middle East), and 5 in the Middle East and North Africa [2].

How well are programs doing?

Facility capacity

District hospitals score somewhat better than health centers, though they are far from perfect. Hospitals are best at doing the things that health centers are supposed to do, but they are expected to go beyond that, for example hospitals should have the capacity to provide blood transfusions. On average across countries, they are just close to even odds for having such additional functions.

Figure 1. Summary Rating for 49 Developing Countries



Access to services

Access to maternal health services was rated separately for rural and urban areas, and urban access is much better than rural access. The urban rating (68%) and the rural rating (39%) constituted by far the largest difference in Figure 1. Particular items reflected this disparity: for example, raters estimated that 81% of urban women have access to a 24-hour district hospital, compared with 58% of rural women.

Care received

The ratings for care received are somewhat higher than the ratings for facility capacity or access, indicating that pregnant women and newborns have somewhat better than even odds of receiving several types of care. This is partly because this set of items places proportionally less emphasis on obstetric emergencies and more emphasis on routine types of care. The best chances for receiving care involve immunization.

Family planning provision

Ratings for family planning provision combine elements of facility capacity, access, and care received. These ratings range from 36 to 71.

District hospitals do better than health centers, and they do best at being able to insert IUDs (rated 71). They also tend to have contraceptive pills in stock, an area in which health centers are not too far behind. Health centers do worst at having progestin-only pills for breastfeeding women (rated 49). Hospitals do even worse, however, at providing male sterilization (rated 36).

Policy and support services

Ancillary services are divided into five categories: policy, resources, monitoring and research, health promotion, and staff training.

Broad policy is generally the strongest area. Having a basic policy (rated 72) and having a service director with a high rank in the bureaucracy (rated 67) rate comparatively well. Similarly, such other policy items as allowing appropriate personnel to provide services, developing policies through consultation with interested groups, and providing frequent public statements of support, get better mean ratings than most other support-service items. The weakest areas (rated just above 50) where policy is concerned, are policies favoring treatment of abortion complications and active implementation of policies through high-level reviews and action plans.

The weakness of implementation is also reflected in poor scores in the area of resources. Only half of the replies suggested that the budget is adequate (rated 48). One resource item, the existence of an active private sector, is rated only slightly more likely than not (rated 58).

The other three items under the policy and support services are similarly mixed. Overall, they rate from 47 to 56, in the middle range, with some internal diversity depending upon the particular items they contain. (See the full report [2] for details.)

Table 1: Composite Scores

Level	Latin America and the Caribbean		East and Southeast Asia		South Asia	
Moderate (70-89)	Jamaica	83.1	China	75.4		
	Dom. Rep.	72.9	Vietnam	73.9		
	Peru	72.1				
Weak (50-69)	Mexico	66.1	Philippines	69.2	India	56.2
	Brazil	64.1	Myanmar	57.1		
	Paraguay	58.1	Indonesia	52.4		
	Ecuador	53.4				
	Nicaragua	50.6				
Very weak (30-49)	Honduras	49.7	Cambodia	33.0	Bangladesh	31.5
	El Salvador	47.9				
	Guatemala	40.4				
	Bolivia	39.1				
	Haiti	31.6				
Extremely weak (10-29)			Pakistan	24.6		
			Nepal	16.9		

	Middle East and North Africa		Francophone Sub-Saharan Africa		Non-Francophone Sub-Saharan Africa	
Moderate (70-89)	Iran	80.9			South Africa	73.3
	Egypt	74.5				
	West Bank	72.9				
Weak (50-69)	Algeria	66.4	Congo, Rep.	51.9	Zimbabwe	65.5
			Ghana	56.6		
			Malawi	53.9		
			Sudan	52.4		
Very weak (30-49)	Benin	48.9	Tanzania	47.2		
	Madagascar	48.1	Kenya	42.5		
	Rwanda	44.3	Mozambique	42.2		
	Mali	42.4	Nigeria	40.4		
	Guinea	40.0	Uganda	40.3		
	Senegal	39.7	Zambia	37.3		
	Congo, DR	39.4	Angola	35.4		
Extremely weak (10-29)	Yemen	29.4	Ethiopia	27.5		

Country Ratings

Countries vary a good deal in the ratings, and one way to show this is with the access ratings. These were done separately for urban and rural sectors, unlike other items, on which rater made national judgments. In Table 1 national ratings are shown by applying weights proportional to the population in each sector.

On the access indicator countries vary greatly, for example Iran and Pakistan, neighbors, are almost at the opposite extremes. For convenience, countries in the table are grouped by level of national access, from "moderate" ratings of 70-89 down to "extremely weak" ratings of 10-29. Each region generally has a mix of countries at different levels, although the central tendencies differ. The Francophone sub-Saharan countries rate low and are the most tightly clustered, most of them receiving "very weak" national access ratings. Non-Francophone countries also rate low but show a broader range than the Francophone ones, from South Africa at the top to Ethiopia at the bottom.

The MNPI is improving over time

According to the raters, the weak scores for maternal health programs actually represent some improvement over past scores. Ratings as of three years previously indicate improvement over time, in all regions, on virtually all items.

Figure 2 shows the differences between the average scores by category in 1996 and 1999 across 49 countries. A 10-point gain over three years is the norm. The level of gains may be somewhat overstated for methodological reasons. Any rater wishing to indicate some gain would have to shift at least 1 point on the scale from 0 to 5, or 20 points. Nevertheless relative gains for different countries or on different items should still be meaningful.

Gains have been greater overall in the two Asian regions than in other regions. The items on which East and Southeast Asia shows the most gains, absolutely and also in comparison to other regions, are those having to do with care received. The region also shows relatively good gains in policy and support services, especially health promotion. For South Asia, on the other hand, relative gains are better than average in policy and support services and also in hospital capacity.

Over the last three years, raters estimate that adequacy has improved 10 points on the typical item. If their judgments are accurate, that would be fairly good performance, and, if sustained, could lead to substantial improvement over time. However, not all regions have been progressing at a similar rate. It remains to be seen why improvements have been smaller in some cases, and what the effect has been of barriers such as poor policy, limited donor support and economic problems.

Checking 'Conventional Wisdoms'

In the full working paper [2] we examined the findings from the study by comparing the results against what appears to be the conventional wisdom (CW). Here is a selection of 6 of the 24 comparisons.

CW By region, programs are weakest in sub-Saharan Africa and in South Asia.

This is only partly true. South Asia indeed receives the lowest overall ratings. But overall ratings for sub-Saharan Africa differ only slightly, on average, from those for Latin America or the Middle East. On the other hand, if one looks at regional access to services weighted by country populations, sub-Saharan Africa is clearly behind the other regions. This seems to imply that the maternal services that do exist in sub-Saharan Africa are not necessarily worse than elsewhere, but given proportionally large rural populations, their coverage is more limited.

CW Routine services for pregnant women are more likely to be provided than emergency services, especially in rural areas.

This seems generally the case, though there are exceptions. Items relating to care received, which mostly have to do with routine care, have higher ratings than items relating to facility capacity and access to services, which on balance focus more on emergency services. Those items on care received that deal with emergency obstetric services do receive low ratings, below other types of care. Nevertheless, there are some types of care that could be considered routine – such as STI and HIV counseling and scheduling a check-up – that receive even lower ratings than emergency care.

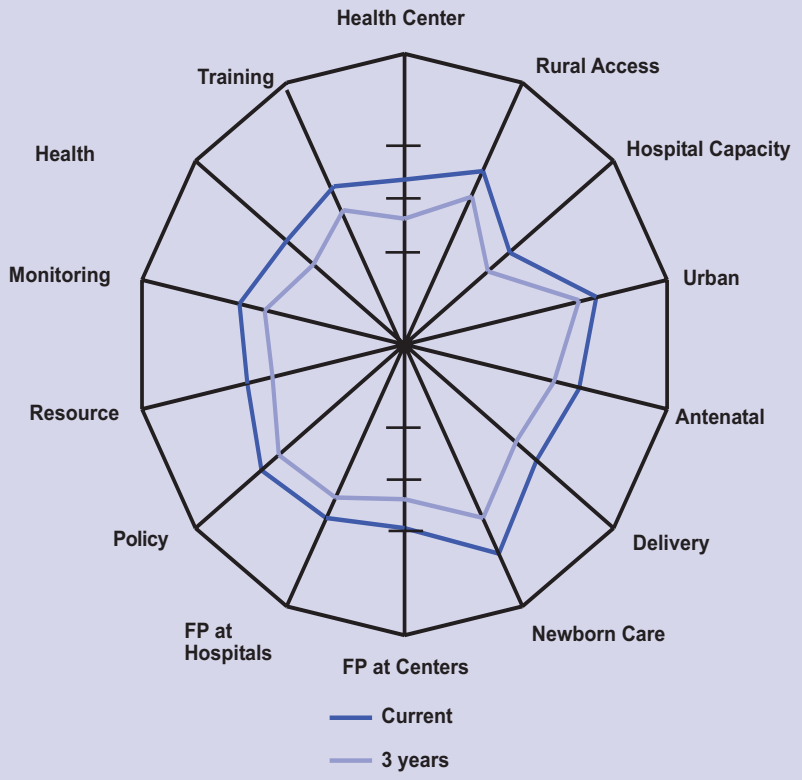
CW Maternal health services are often provided on demand, as an apparent need arises, rather than individual pregnancies and individual women being systematically checked and followed up by health providers.

On the contrary, antenatal care in general receives relatively good ratings. Nevertheless, the one item that might reflect systematic follow-up, the scheduling of a check-up 48 hours after delivery, receives a mean score across countries of only 42. This may be symptomatic, though it does not prove the conventional wisdom.

CW Due to substantial donor assistance for family planning, its services are usually more satisfactory than other maternal health services.

Actually family planning items receive ratings that are only around average, and some ratings are well below average. Ru-

Figure 2. Mean Rating across 49 Countries, by Category



ral access to postpartum family planning, for instance, is only 36 percent. The items on family planning were not selected to set standards more stringent than in other areas and probably do not do so. Perhaps family planning faces special obstacles, and the considerable assistance that has been available has on average only served to neutralize such obstacles. Or, family planning may rely proportionally more on services provided outside health centers and hospitals, so that ratings based on services at these facilities overlook other areas of strength. Still another possibility is that the assistance provided for family planning has rubbed off, to be of equal benefit in other areas of maternal health services, so that family planning services do not stand out.

CW Where public sector services are inadequate, private facilities provide more maternal health care.

There is no evidence of this across regions. If anything, the reverse appears to be the case. Where ratings are particularly low, in South Asia, an active private sector is also least evident.

CW- Maternal health care services have made little progress since the Cairo conference.

The data, to the extent that they are accurate, disagree. Over the last three years, raters estimate that adequacy has improved 10 points on the typical item. If their judgments are accurate, that would be fairly good performance, and, if sustained, could lead to substantial improvement over time.

Conclusions

How good are these ratings? A definitive answer is not yet available, but some indications exist that the ratings are at least reasonable. Raters who were program administrators and raters who were services providers hardly differed in their ratings of different program areas. Raters outside a program, particularly if medical doctors, appear to have given slightly lower ratings, though the differences were usually not significant. These personnel comparisons suggest no large biases in the ratings.

The proportion of births with a trained attendant present, from DHS data, agrees well with the current ratings (as of 1999) for births attended. The R2-correlation across countries is 0.70. The correlation is even stronger, at 0.83, with ratings for three years previously – effectively for 1996. For the proportion receiving at least two tetanus injections, the correlations were also strong, at 0.62 for current ratings and 0.74 for ratings three years previously. The mean percentage of births with a trained attendant across the 23 DHS countries is 55, virtually identical to the mean current rating of 56 for these countries and higher than the rating of 43 for three years previously.

Apart from more intensive analyses of the data, various questions require further investigation. Why do developing-country regions appear similar in some ways but so different in others? Is national access to services indeed the fairest way to make comparisons? Why are some countries rated much better than others, and do income, education, program leadership, or other factors account for the differences? Much remains to be learned from these expert ratings of maternal and neonatal health programs in developing countries.

Notes

[1] The assessment was conducted by the Futures Group International (TFGI), funded through the MEASURE *Evaluation* project. TFGI developed a questionnaire, identified a consultant for each country to be studied, and coordinated the recruitment of experts to produce ratings collected in 1999 and early 2000.

[2] For a more complete discussion of the results and the methods used, refer to the MEASURE *Evaluation* Working Paper, No. 26, "Rating Maternal and Neonatal Health Programs in Developing Countries," by Rodolfo A. Bulatao and John A. Ross, August 2000.

Monitoring Political Commitment and Program Effort in HIV Prevention and AIDS Care: The AIDS Program Effort Index

John Stover

- √ **The AIDS Program Effort Index (API) is a composite index designed to monitor political commitment and program effort in the areas of HIV prevention and AIDS care.**
- √ **The API is composed of 100 individual items grouped into 11 categories, which are rated by 15-25 knowledgeable people in a country.**
- √ **The API was used in 38 countries in 2000 and is likely to be used increasingly in the near future to monitor global and national efforts to expand the response against AIDS.**

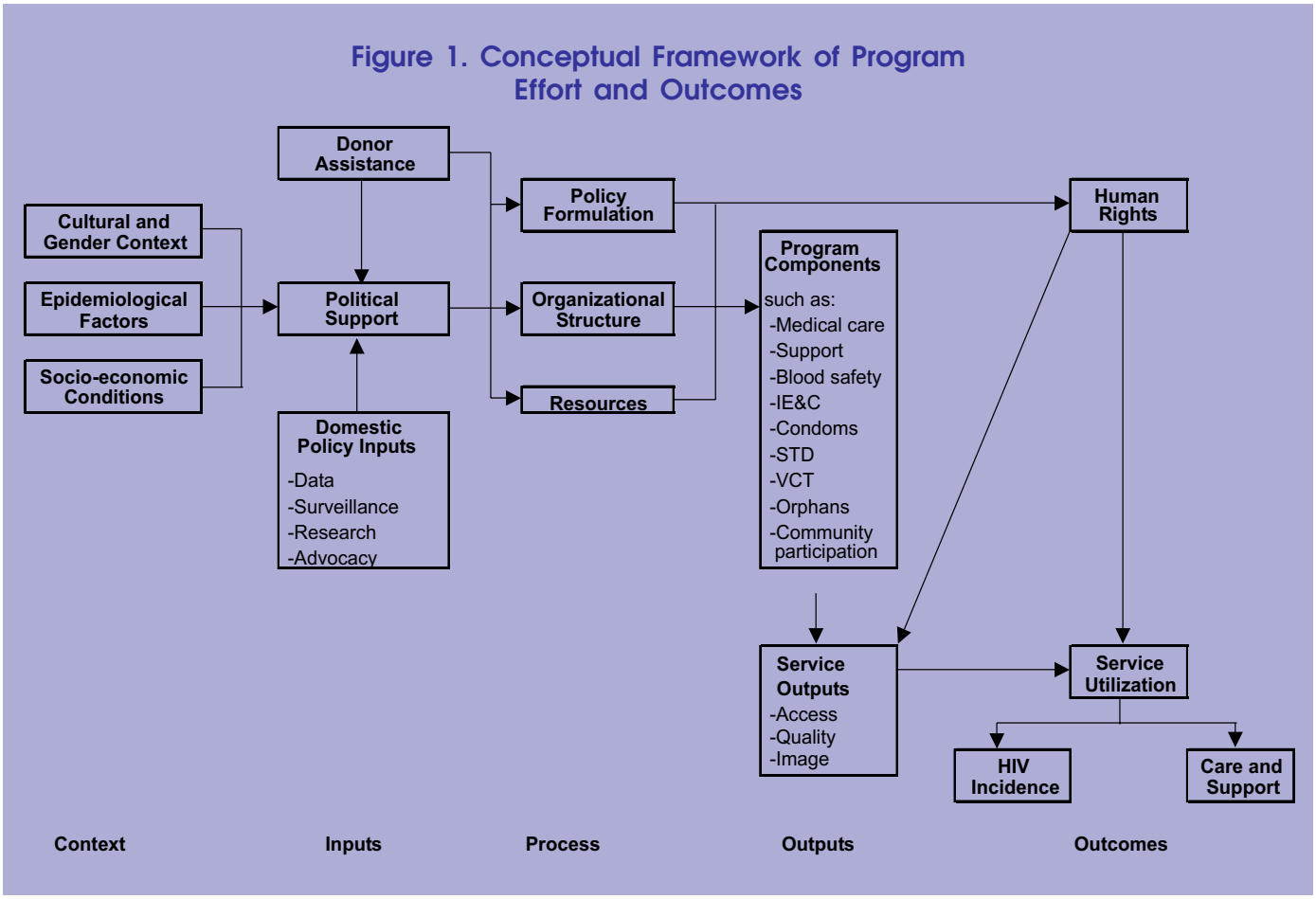
Many factors may affect the success of national HIV/AIDS programs, including political commitment, program effort, sociocultural and economic context, and resource availability. Building upon the experience with the Family Planning Effort Index, the POLICY project, USAID and UNAIDS developed a program effort score for HIV/AIDS and other STD programs: the AIDS Program Effort Index (API). The API is a composite index designed to measure political commitment and program effort in the areas of HIV prevention and care. Instead of tracking low-level inputs, such as training workshops conducted and condoms distributed, the API is intended to measure program effort independent of program outputs. For example, program effort includes items such as the degree of political support, the amount of participation in the program and the resources devoted to the program, but it does not include output measures such as the proportion of acts protected by condom use.

There are many uses for scores that measure program effort independent of output. At the global level, an effort score can be used to analyze the independent contribution of program effort to program success in a variety of social and cultural settings. The API can also be used in an analysis of the relative contribution of international organizations. Global level use of the API may contribute to monitoring of the goals set at the United Nations General Assembly Special Session (UNGASS) on HIV/AIDS in July 2001 and to assess progress made in the context of the USAID-supported Expanded Response. At the country level an effort score can be used to compare the national effort against that of other countries with similar settings or problems. The scores can also be used as a diagnostic tool, to indicate which program areas are weakest and which are strongest and to suggest corrective action. In this context the term "national program" encompasses not only the formal government program but also includes efforts by individuals, nongovernmental associations, communities, and others.

Conceptual framework

Figure 1 shows the conceptual framework for the relationship between HIV/AIDS program effort and desired outcomes. This framework is adapted from a similar framework developed for family planning services by Tsui and others (Bertrand, Magnani and Knowles, 1994).

Figure 1. Conceptual Framework of Program Effort and Outcomes



The instrument

The API is a composite indicator composed of 100 individual items grouped into 11 key categories. Each item is scored on a scale of 0-5 by knowledgeable individuals. The item scores are averaged for each category to produce a category score that does not depend on the number of items in the category. The category scores form a profile describing the program effort of each country. The components are as follows:

- Political support
- Policy formulation
- Organizational structure
- Program resources
- Evaluation, monitoring and research
- Legal and regulatory environment
- Human rights
- Prevention programs
- Care programs
- Service availability
- United Nations' role

Box 1 shows an example of the eight items that are rated to obtain a policy formulation score.

The Respondents

Judgments are provided by 15-25 people in each country. Respondents are not meant to be a representative sample but are carefully selected for their knowledge and viewpoint. The goal is to find the 15-25 most knowledgeable people from a variety of backgrounds. Usually, the respondents are selected from a variety of backgrounds, such as the National AIDS Control Program, Ministry of Health, other government organizations, NGOs, researchers, academics, major religious groups, community-based organizations and donors.

A primary purpose of the API is to measure change. Ideally, data from multiple rounds are available with intervals of at least two years. In the initial round the participants are asked to rate each item twice, once for the current situation and once for the situation two years ago.

Box 1. Policy Formulation
(Respondents are asked to rate the validity of the following statements on a scale from 0 to 5.)

1. *A favorable national policy exists.*
2. *Formal program goals exist.*
3. *Specific and realistic strategies to meet program goals exist.*
4. *A national coordinating body exists and functions effectively.*
5. *Ministries other than Health are involved in policy formulation.*
6. *Policy dialogue and formulation involves NGOs, community leaders, and representatives of the private sector, women's groups and special interest groups.*
7. *International organizations have facilitated policy formulation through the provision of technical assistance and guidelines.*
8. *International organizations have facilitated planning through the provision of technical assistance and guidelines.*

Results by Region

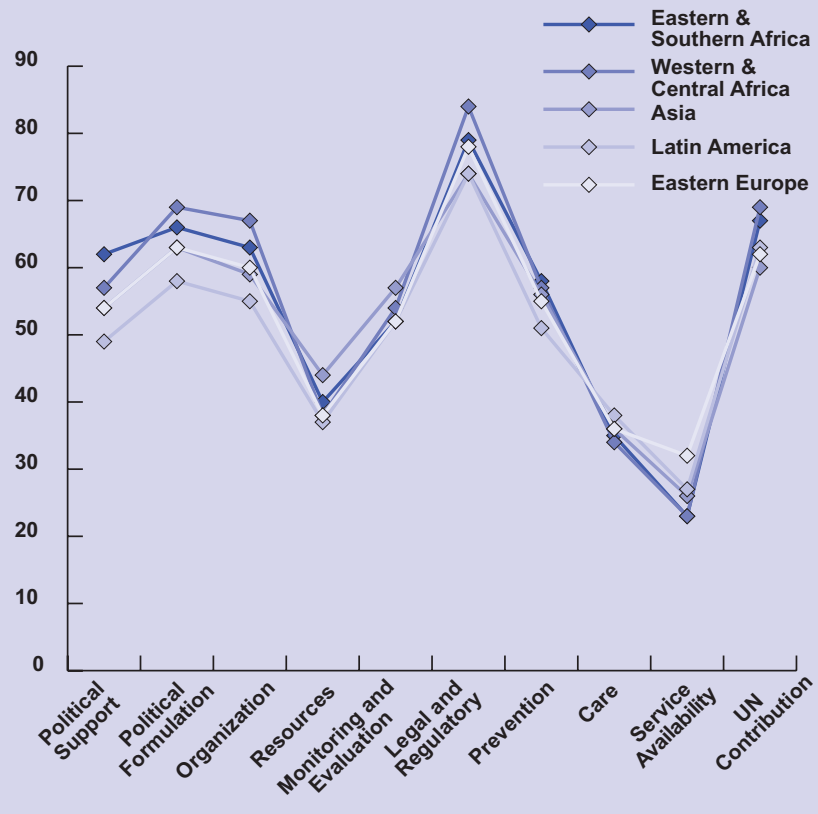
For 2000, there are results from 38 countries. Figure 2 presents the average score for ten components by region. Country rankings will be reported in the future. Although it would be possible to rank countries, it is likely that respondents in each country used different standards in rating effort. Furthermore, analysis revealed that respondents did not adequately understand the scoring of the human rights component. That component is being revised and will be scored at country meetings in early 2001.

- Programs are judged to be doing a particularly good job on legal and regulatory issues, with scores above 70%. This indicates that the laws, regulations and practices generally support effective interventions. For example, in most countries condom advertising is allowed and there are few restrictions on who may receive STI services.
- Policy formulation is judged to be good. Respondents in most countries reported that formal policies and laws were in place that established program goals and strategies, organized a multi-sectoral effort and involved a variety of stakeholders in policy dialogue.
- The organization and structure of the national program was also judged to be relatively good. Most countries have a national government program in place and attempt to include non-governmental organizations and representatives.
- Only about 25% of the total population in each country surveyed have reasonable access to care and prevention services. This suggests that there is still a large amount of improvement required in providing access to basic preventive and care services.

Changes from 1998 to 2000

Figure 3 shows the change in component scores by region from 1998 to 2000. The respondents judged that there had been a large increase in political commitment and policy formulation during the past two years, especially in Eastern and Southern Africa. A number of countries in Eastern and Southern Africa have passed and implemented new national HIV/AIDS policies, including Kenya, Ethiopia, Uganda and Zimbabwe. In addition, more and more leaders are speaking about HIV/AIDS. This increase has raised the scores for political commitment and policy formulation from around 40% to about 60%. Scores for the other components also increased on average, but by much smaller amounts.

Figure 2. AIDS Program Effort Index by component and region - 2000



Conclusions and Next Steps

1. All countries studied have organized at least some reasonable effort. No country received a total score (averaged across all components) lower than 39. On the other hand, no country received a total score higher than 77. Thus, there is considerable room for improvement in all countries.

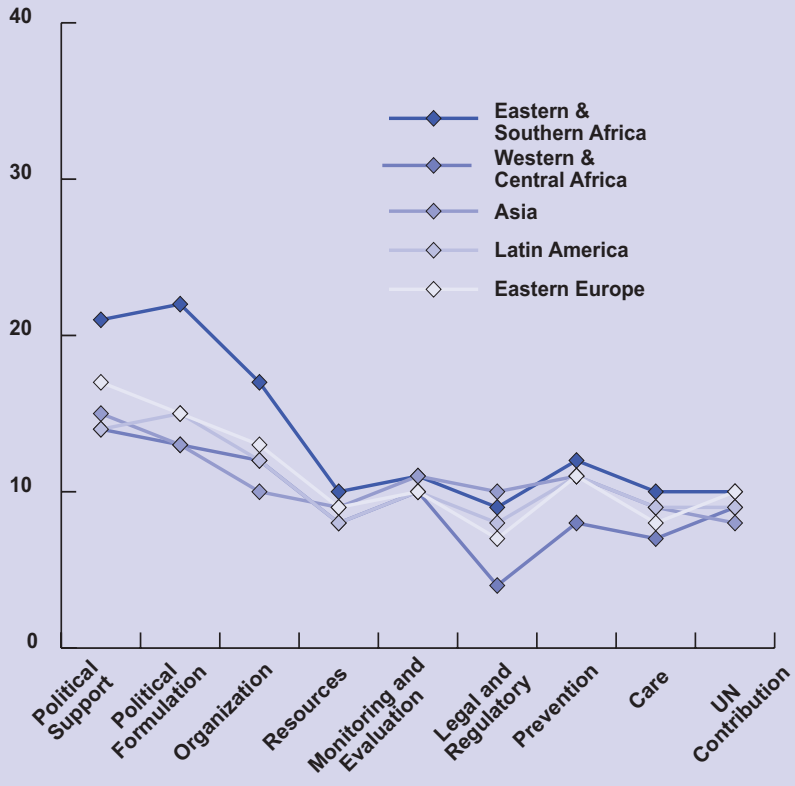
2. Respondents judged that the best efforts have occurred in the legal and policy areas. The highest scores were given to the legal and regulatory structure and policy formulation. Even here, however, considerable improvement is required, primarily to ensure that the legal structure that is in place is used to protect the human rights of people affected by HIV/AIDS.

3. The political commitment of national leaders to confront HIV/AIDS has been a major concern to many. Commitment has been weak in the past and this has affected programs in a variety of ways. In the past two years, however, political commitment has increased more than any other component. The increase has been especially marked in Eastern and Southern Africa. Although political commitment is still lacking in many areas, it is encouraging to see that it has been increasing in recent years.

4. One of the weakest areas is resources. Respondents felt that the resources devoted to HIV/AIDS programs are inadequate to support an effective response. Although respondents felt that resources had increased over the last two years, the increase was quite small compared to the other components. The increased political commitment has not yet led to a similar increase in resources.

5. The API shows quite clearly that the effort being made to care for people living with HIV/AIDS is the weakest component of most programs. Care is the lowest

Figure 3. Change in API from 1998 to 2000 by Component and Region



rated component in all regions and the service availability items relating to care were also the lowest rated.

6. Service availability is a major problem for most countries. Even in the capital cities the majority of the population does not have access to most services. The best scores were given to safe blood, condoms and STI services.

7. United Nations agencies and other international donors are making a significant contribution to program effort. Respondents judged international assistance to be a positive factor in most country programs. The contribution is greatest for policy, planning and prevention and weakest for care.

8. A separate effort will be undertaken in early 2001 to get international experts to compare program effort across a range of countries. These scores will be used to rank countries on a consistent scale. Results should be available by the middle of 2001.

Notes

[1] The API work was funded by USAID and UNAIDS and carried out by the POLICY project at the Futures Group International, 1050 17th Street, NW, Suite 1000, Washington, DC 20036. For the full report, "Measuring the level of effort in the national and international response to HIV/AIDS: The AIDS Program Effort Index (API)," from February 2001, see The Futures Group International website. www.tfgi.com.

[2] Bertrand, Jane T., Robert J. Magnani and James C. Knowles. 1994. *Handbook of Indicators for Family Planning Program Evaluation.* Chapel Hill, NC: The Evaluation Project.