BETTER PRACTICES in *Evaluation*

Measuring Provider Performance

Challenges and Definitions

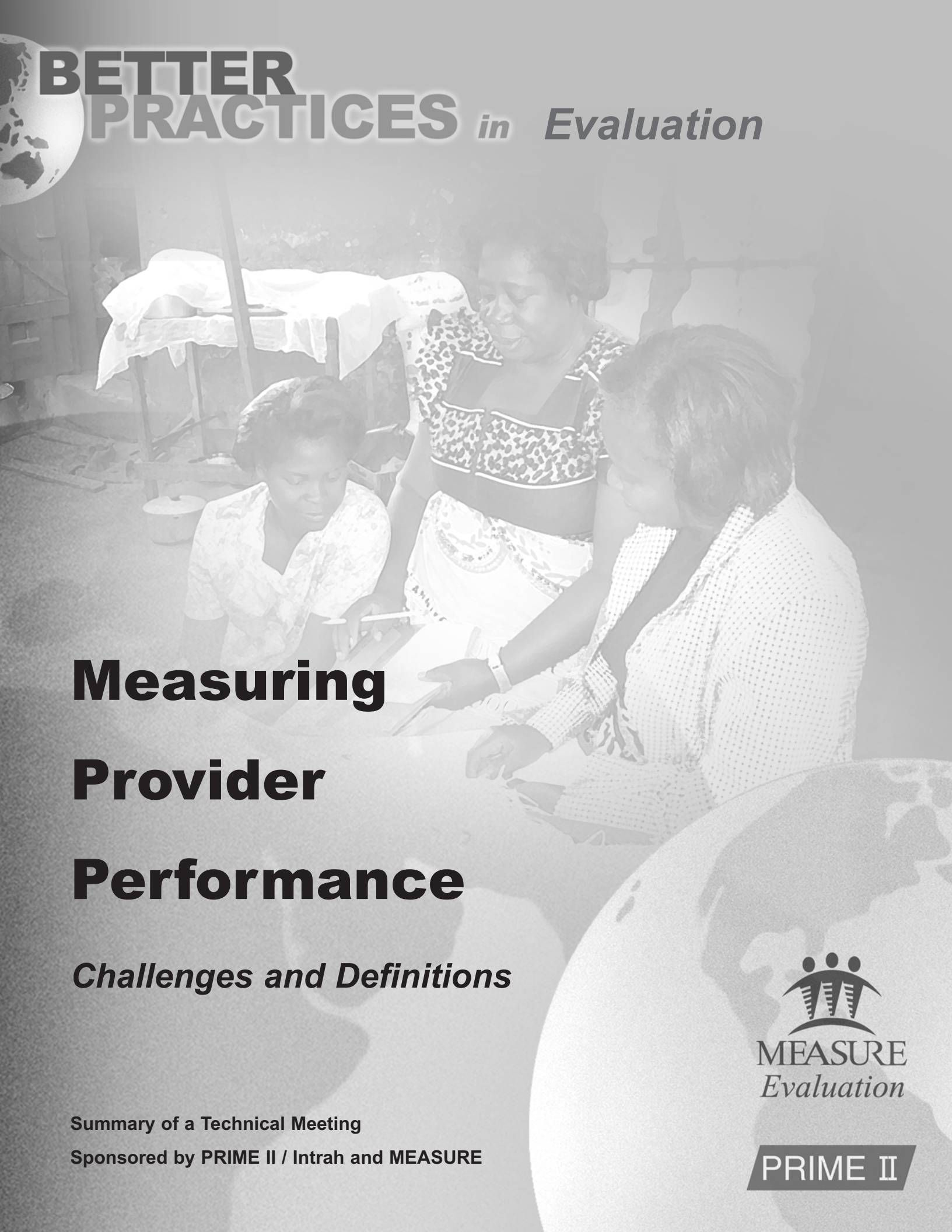
MEASURE Evaluation

PRIME II

Summary of a Technical Meeting

Sponsored by PRIME II / Intrah and MEASURE

BETTER PRACTICES *in* *Evaluation*



Measuring Provider Performance

Challenges and Definitions

MEASURE *Evaluation*

Summary of a Technical Meeting

Sponsored by PRIME II / Intrah and MEASURE

PRIME II

Contents

Performance concepts	**1**
The definition of performance in Human Performance Technology	1
Measuring provider performance: Principles and indicators	2
Adapting the HPT definition to the FP/RH field: Examples and measurement issues from the PRIME II Project	3
Measurement issues	**4**
Measuring behavior	4
Measuring accomplishments	5
Toward a single indicator combining behavior and accomplishments	6
Examples of performance-related measurements used by organizations and projects	6
Defining and measuring provider performance	**9**
On the indicators selected to measure performance	10
Study design and methods for measuring provider performance	11
Reaching consensus	12

Figures

Figure 1: Provider performance within the PRIME II Project M&E model	3
Figure 2: Differences in CSI and ORH skills between trained and untrained paramedics	4
Figure 3: Average monthly clinic attendance before and after paramedics' training	4
Figure 4: Performance in the process of monitoring labor, 2000	6
Figure 5: Percentage of health facilities meeting basic infection prevention standards	7
Figure 6: Provider compliance with clinical guidelines when administering injectable contraceptives	7

Tables

Table 1: Information collected for each service assessed by the SPA	8
Table 2: Comparison using observation vs. facility audits as unit of analysis for service provision practices	8
Table 3: Changes in the mean scores upon application of the QMT	8
Table 4: Example of the use of the service test for diagnostic purposes	9

Appendices

Appendix 1: Client rights and provider needs	13
Appendix 2: Calculation of performance: Bangladesh example	14
Appendix 3: Short list of QIQ indicators	15
Appendix 4: Main measurement and interpretation problems encountered with methods to assess the quality of care	16

PRIME II Better Practices
Number 1
November 2002

Acknowledgements

This publication was prepared by Alfredo Fort.

Thanks to Candy Newman and Catherine Elkins for valuable inputs to earlier drafts of this publication; to David Nelson and Rebecca Mann for editing and formatting; and to Barbara Wollan, David Nelson and Reynolds Richter for assistance in organizing and documenting the meeting.

Measuring Provider Performance:
A Technical Meeting
The University of North Carolina at Chapel Hill
December 6-7, 2001

List of Participants

Ruth Bessinger
MEASURE *Evaluation*

Dale Brethower
Western Michigan University

Catherine Elkins
MEASURE *Evaluation*

Alfredo Fort
Intrah/PRIME II

Nancy Fronczak
MEASURE *Evaluation*/DHS

William H. Jansen II
Intrah/PRIME II

Federico León
The Population Council

Marc Luoma
Intrah/PRIME II

Richard F. Mason, Jr.
Intrah/PRIME II

Erin Mielke
EngenderHealth

Allisyn Moran
JHPIEGO Corporation

Constance Newman
Intrah/PRIME II

David Nicholas
Quality Assurance Project/University Research Corporation

Anne Packman-Young
Management Sciences for Health

Linde Rachel
Management Sciences for Health

Heidi Reynolds
Family Health International

Kai Spratt
JHPIEGO Corporation

John Stanback
Family Health International

Lauren Voltero
Intrah/PRIME II

Performance concepts

While the need to measure "performance" in the field of family planning and reproductive health care (FP/RH) is widely recognized, there is no consensus on a standard definition of the term. Consequently, when organizations and projects describe or measure "performance," particularly in the context of health worker or program evaluations, the term may be used in ambiguous and confusing or even contradictory ways. During a review of the PRIME II Project by the Communications, Management and Training division at USAID/Washington in early 2001, this fundamental issue arose repeatedly: What is *performance?* How do you define performance?

In response to these discussions, PRIME II saw an opportunity to help shape and advance the dialogue on performance measurement among FP/RH professionals and organizations. The first step was to collaborate with MEASURE *Evaluation* and invite 18 monitoring and evaluation specialists from ten organizations working in FP/RH training and service delivery to participate in a technical meeting in Chapel Hill, North Carolina, in December 2001.

The primary objective of the meeting was to build consensus on the concept of performance in the context of individual providers of FP/RH services. PRIME and MEASURE established this focus to encourage participants to work toward definitional consensus at the health worker level rather than trying to cover all of the broader issues involved in the measurement of performance at the organizational or systemic levels. The agenda included three specific objectives:

1. Presenting theoretical constructs and empirical applications of provider performance (or performance-related) measurement and indicators
2. Discussing how institutions and projects have defined and measured such indicators and the results and challenges arising from these exercises
3. Synthesizing the experiences of the group to arrive at a consensus on the definition of provider performance in FP/RH service delivery.

In order to fulfill the first two objectives, participants from each organization presented a topic, followed by comments and discussion from the group. To meet the third objective, an extended discussion session took place on the second day of the meeting.

The definition of performance in Human Performance Technology

Dale Brethower, professor emeritus of psychology at Western Michigan University and past president of the International Society for Performance Improvement (ISPI), presented a view of performance from the field of Human Performance Technology (HPT). He began his presentation by laying out the standard definition of performance adopted by ISPI in 1961: "behavior (what the performer does) plus accomplishment (the result of the behavior)." Brethower's thesis, however, focused on what he called the "bottom line" of performance: "costs and results." He based his assertion on two assumptions: 1) every behavior or activity carried out by a performer has a cost; 2) clients and organizations will consider the "valued result" of that cost. Brethower provided examples of performance costs and results in the health care arena. For instance, provision of flu shots—a costly activity—should produce the desired result of lower incidence of flu cases; delivery of contraception—with attached costs—should result in "informed and protected clients" and "lower incidence of [unintended] pregnancies."

Brethower then used the flu shot example to illustrate how an evaluation might reveal why an activity or cost did not produce the intended result. Perhaps not enough people were vaccinated to produce the required group immunity. Or maybe a radically different strain of flu appeared in the area and was resistant to the vaccine. The evaluation would reveal that the activity of providing the flu shots was carried out but was quantitatively (i.e., insufficient numbers vaccinated) or qualitatively (e.g., aimed at the wrong virus agent) misdirected. In either case, the conclusion would be that performance was subpar, even though providers had administered the vaccines correctly.

Brethower went on to elaborate some of the broader implications of applying the HPT definition to health care. He emphasized that performance measurement needs to focus on "what is done and what is accomplished" in a culturally sensitive way because imposing recipes from one culture onto another only generates multiple problems. Using an

example from training design, he also pointed out the importance of standards for performance. According to Brethower, not having standards is "a deadly design mistake, because the mission of assuring that people know how to perform competently cannot be fulfilled." Performance standards, he argued, bring a whole new meaning to the phrase "quality of care." Correspondingly, designers of performance-related FP/RH interventions need to look within the behavior component of performance to address the challenge of improving quality while reducing unnecessary costs.

Finally, Brethower alerted participants to another implication of focusing on performance in the real world: "the two parts [of performance] are often pulled apart by organizational practices: administrators are concerned with costs, and medical personnel are concerned with results." Arising from experience, Brethower's remark clearly signaled the challenge of applying an integrated definition of performance across individuals and institutions used to viewing one or the other component in isolation.

Measuring provider performance: Principles and indicators

Catherine Elkins of MEASURE *Evaluation* presented concepts and challenges inherent in monitoring and evaluation (M&E) of the quality of provider performance and the outcomes of performance-related FP/RH interventions. She focused in particular on the process of constructing solid M&E plans and the characteristics of good FP/RH program indicators.

In order to develop an M&E plan to measure provider performance, Elkins emphasized, program implementers and evaluators need to reach consensus on basic issues such as the *units* (e.g., client-provider interaction, individual provider, service delivery point) and the *dimensions* (e.g., quality, efficiency, preventive vs. curative care, counseling) of measurement. Agreement on the appropriate degree of attempted precision in measurements (e.g., post-training testing vs. follow-up observation, minimal thresholds vs. qualitative gradations) is also necessary, ideally with consensus on the interpretation and implications of the measurements or indicator values thereby obtained. Elkins reminded meeting participants that while *monitoring* seeks to investigate the status quo of a phenomenon, *evaluation* includes its contextual features. Evaluation is a constructive "analytical exercise" that attempts to determine "the amount of the change in outcome that is due to the program or intervention." The purpose of M&E as a whole "is to measure program effectiveness."

The central part of Elkins' presentation offered a substantial review of the characteristics and applications of good indicators for measuring provider performance. Regardless of what is being measured as performance, the necessary principles to ensure appropriate and accurate measurement remain the same. Sound indicators should be:

- *valid* ("the phenomenon it measures matches the result it is designed to measure")
- *reliable* (measurement error is minimized)
- *precise* ("clear, well-specified definitions")
- *independent* (non-directional and capturing "a single dimension at a certain point in time")
- *timely* (measurement "at appropriate intervals" related to program goals and activities)
- *comparable* (units, denominators and other components of measurement allow comparison across different populations or approaches).

Without assessing their definitions of performance, Elkins presented a range of performance-related indicators that have been used for monitoring and evaluation of FP/RH training and management interventions. Categories include:

- Training Events: measuring training inputs and effort; counting trainees who complete courses
- Training Participants: gauging trainees' learning
- Trained Health Service Workers: measuring training-related knowledge and skills back on the job; assessing change in competence due to prior training events
- Providers: measuring specific performance skills on the job; gauging efficiency, quality of care in practice
- Teams: measuring performance as team members; assessing supervisory and problem-solving roles
- Facilities: measuring performance of the services provided; adherence to guidelines, protocols
- Systems: measuring management, logistics, adequacy and efficacy of support; testing referrals; impact

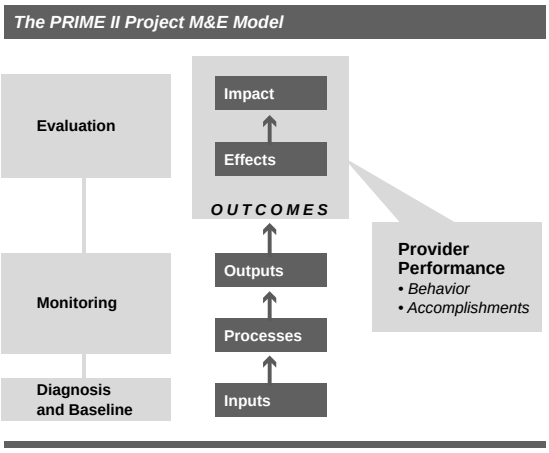
- Infrastructure/Sector: measuring systemic functions; political support; linking quality to outcomes.

Final selection of indicators depends on the activities and goals of the project or program. Careful selection requires that indicators be laid out rationally within the specific program design (e.g., log-frame or results framework) and that clear operationalization of their measurement along with requisite data are spelled out for each.

As Elkins emphasized, applying and interpreting indicators can pose a number of challenges, among them avoiding subjectivity, lack of adequate regard for relevant "local conditions or assumptions," and unclear yardsticks for measurement. Elkins proposed that some performance-related indicators might be inherently subjective, such as "leadership," "quality of care" and "improvement." This does not preclude their use, but the subjectivity needs to be addressed through very careful and precise application. An example of a relevant local condition or assumption would be an evaluation indicator that relies on facility records that may not exist, or may be available only at variable levels of completeness.

Figure 1

Provider performance within the PRIME II Project M&E model



An unclear yardstick occurs when an indicator's measurement has not been spelled out in sufficient detail and thus indicator values are not replicable or comparable over time or across health interventions. For example, an indicator for contraceptive costs might be insufficiently clear if it does not define such factors as whether the cost is an overall average, whether and how the calculation takes into account seasonal variations, local cost versus cost in foreign currency, fluctuating exchange rates, and so on.

Elkins concluded her presentation with a classic pedagogical exercise providing "tips" to ensure that data adequately capture the intended indicators of performance. These reminders include making sure that the necessary data are collected or retrieved (*accessibility*), encompass all areas of interest (*coverage*), are fully reported or collected (*completeness*), and are obtained through reliable sources or tested instruments (*accuracy*) as often as is required (*frequency*) to reflect the time period of interest (*impact and reporting schedule*).

Adapting the HPT definition to the FP/RH field: Examples and measurement issues from the PRIME II Project

To demonstrate how the HPT definition of performance can be applied in a "real-world" FP/RH context, Candy Newman and Alfredo Fort presented experiences and results from the design and implementation of the PRIME II Project. They began by stating that PRIME II uses the HPT definition of "behavior and accomplishments" for performance, and that measuring provider performance is one of the main mandates of the project. In order to put PRIME's measurement of provider performance into perspective, they also presented the place in which provider performance falls within the project's overall M&E model. Figure 1 illustrates how measuring both provider behavior and accomplishments fits neatly into the *Effects* dimension of PRIME's evaluation paradigm.

Though the Project is not mandated to measure the *Impact* of its interventions, measuring provider performance has allowed PRIME to move from output-oriented evaluation (e.g., people trained or supported) to outcome- or *results*-oriented evaluation.

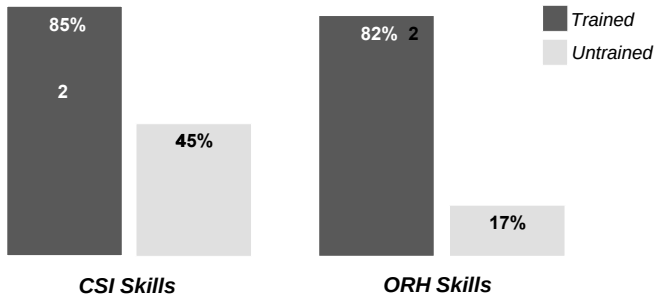
Newman then offered a few examples from PRIME program evaluations to illustrate how measuring both behavior and accomplishments is valuable to achieving project goals. In an evaluation of a vaccination initiative with community midwives in Yemen, the measurement of performance combined the ability of the midwives to administer the vaccines

with the number and type of clients (mothers and infants) that were served. Similarly, in a pilot program to establish adolescent-friendly services in primary health facilities in Uganda, performance was established as both the ability of providers to offer FP counseling to adolescents according to agreed upon standards *and* the number of adolescents in the project clinics who actually received such counseling (a measurement that could also be extended to the fraction of adolescents counseled who initiated the use of a contraceptive method). In evaluating the training of paramedics to implement an Essential Service Package in Bangladesh, performance measurement incorporated provider knowledge and skills in child survival interventions (CSI) and other reproductive health (ORH)[1] services (*Behavior*) and the number of ORH and CSI clients seen over a 12-month period (*Accomplishments*).

Fort went on to describe the instruments used to measure behavior and accomplishments in the Bangladesh example. An evaluation conducted more than a year after PRIME's intense training interventions found the skills of trained paramedics at significantly higher levels than untrained providers (see Figure 2).

Figure 2

Differences in CSI and ORH skills between trained and untrained paramedics *Bangladesh Training Study, 2001*



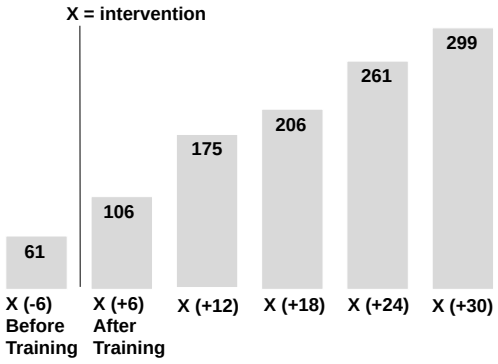
[1] ORH was differentiated from family planning services and was defined as a combination of Antenatal Care (ANC), Postnatal Care (PNC), Newborn Care (NBC) and Reproductive Tract Infection/Sexually Transmitted Diseases (RTI/STD) services. CSI was defined as a combination of Acute Respiratory Infection (ARI), Control of Diarrheal Diseases (CDD), Immunization, Breast Feeding and Interpersonal Communication.

The study also found steady increases in clinic attendance after trained paramedics returned to their facilities (see Figure 3).

Figure 3

Average monthly clinic attendance before and after paramedicsí training

Bangladesh Training Study, 2001



Thus, the Bangladesh report's conclusion—that provider performance had improved—reflected an integrated concept of performance in line with the HPT definition.

Measurement issues

Measuring behavior

Fort also outlined some of the issues that arise when measuring behavior, notably the lack of consensus on definitions of commonly used terms. For example, while some projects evaluate *proficiency*, others might use the terms *competency*, *quality* or even *performance* to describe similar measurements of behavior.

David Nicholas, of the Quality Assurance Project/University Research Corporation, described *competence* as "possession of the skills and knowledge to be *able* to comply with predefined standards" and *quality* as "performance according to standards." For Brethower, however, the definition of quality is always "customer satisfaction." Kai Spratt, from JHPIEGO, defined competency as a "task done to standard," emphasizing that competency leads to improved job performance. *Proficiency*, on the

other hand, is often used to refer to more skilled performers who achieve competence through carrying out steps in a more "unconscious" way because of their extensive experience. One participant referred to Carl Binder, the HPT methodologist, who defines proficiency as the *fluency* with which an experienced worker carries out a task. The content of evaluations using these terms varies widely, with some measuring technical competence (e.g., of clinical examinations or infection prevention) and others measuring counseling and informational or interpersonal skills or even the fulfillment of administrative tasks (e.g., registering data on a clinical history or keeping track of client appointments).

The methodology to obtain these data varies too, ranging from self-assessments and direct observations to mystery or simulated clients and client exit interviews. The Quality Assurance Project asserts, according to Nicholas, "that the most effective performance assessment approach is that of *self-assessment* carried out by teams as part of their quality improvement activities."

Similar self-assessment principles are employed in the *Quality Measuring Tool* (QMT) developed by EngenderHealth and presented to the meeting by Erin Mielke. She described the QMT as a ten-section instrument corresponding to the seven client rights and three provider needs as formulated by Huezo and Díaz (1993) (see Appendix 1 for the full list of rights and needs).[2] Although self-assessment exercises, in principle, can motivate and empower workers concerned with improving the services they deliver, these internal methodologies have an intrinsic subjectivity that may be regarded as insufficient or even inappropriate when measuring program effectiveness.

While recognizing the limitations of direct observations, Nancy Fronczak, from MEASURE *Evaluation*/DHS, finds them preferable to self-assessments, which are "valuable for process, but not as telling as data." For instance, data on the number of IUD clients who return with infections provide a more objective measure than provider self-assessments on IUD insertion. In another example, the question "Do staff give information on breastfeeding and infant care to all postpartum clients?" is likely to be answered more accurately through observation of real or simulated exchanges between providers and clients than by asking the providers themselves. Purporting to assess fulfillment of clients' rights solely by interviewing providers is another weakness of self-assessment tools like the QMT.

External evaluation bears its own disadvantages—for example, the inability to understand critical internal processes or seek out collateral or unexpected activities that may have contributed to program results. Still, this method provides an increased assurance of objectivity and, thus, is used more frequently, as Nicholas pointed out, when assessments are being made for accreditation or certification purposes. An ensuing discussion centered on the relative merits of external and internal assessment. For Nicholas, it would be a "wrong philosophy if people relied more on external than internal evaluations." However, other participants noted the value of external evaluations as a source of helpful information to providers. The need to ensure that qualified people conduct external evaluations was duly noted.

Analysis and data management can also be used to report results in different ways, depending on whether answers are coded dichotomously (e.g., yes/no) or on a scale (e.g., five-point, from "excellent" to "unacceptable"). Scores of indicators (e.g., tasks) can be tallied separately or combined. Summary or holistic scores, such as indices, can be obtained in several ways. Furthermore, ratings and scales depend on limits and thresholds, and these, in turn, rely on standards set before data collection. In the clinical arena, such standards need to be based on the most recent scientific evidence, usually determined through internationally recognized bodies such as WHO.

Measuring accomplishments

To complete the performance equation, accomplishments must be measured: What are the results of improved provider behavior (and other clinic enhancements)? Typically, these are measured through reviewing client records and clinic statistics

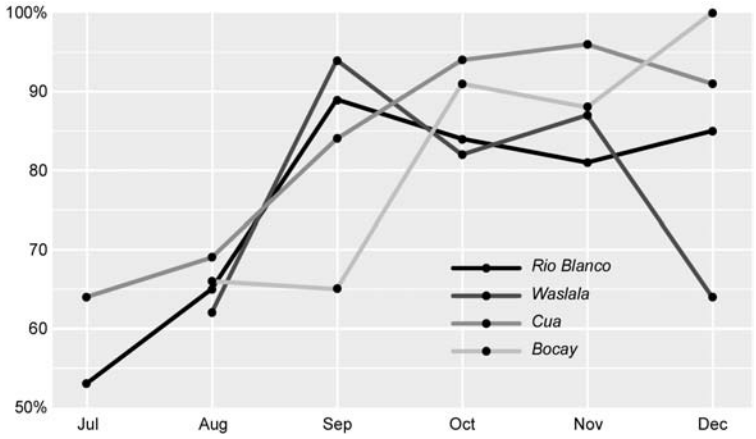
[2] Unfortunately, the provider needs section was not designed with a "performance improvement" paradigm in mind, hence lacks important recognized performance factors such as *Incentives*, clear *Job Expectations* and appropriate and regular *Feedback*, which also constitute provider needs.

to reveal the number of new or total clients receiving various services and the characteristics of those clients (e.g. civil status, age, sex). Nicholas, who defined performance as "the actual output and quality of work performed by organizations, teams or individuals," provided a few examples of such statistics from the QAP in his presentation. He advocated the use of *run charts* or "graphs of certain indicator results over time," which allow for rapid analysis of trends in relation to periods before and after interventions. A run chart used to track the correct use of a partograph among clinic staff in Nicaragua is presented in Figure 4.

Figure 4

Performance in the process of monitoring labor, 2000

Percentage of deliveries in which partograph was used



Obviously, the most important challenge in compiling such statistics is the accuracy and completeness of data. Attribution also presents problems if increased clinic attendance might be due to factors other than an individual provider's improved performance. This can be the case when clinics have more than one provider, or when concomitant internal or external interventions affect a clinic's accessibility, public image or capability to offer services (e.g., IEC campaigns, improved logistics and supply chains).

Toward a single indicator combining behavior and accomplishments

In an effort to come up with a systematic way to ensure that both components of performance are measured, Fort presented a suggestion for combining separate behavior and accomplishments measurements into a single indicator. Using PRIME II's study of Bangladeshi paramedics to illustrate the proposition, he elaborated on how to construct an *index* of performance. In essence, *standards* would need to be set beforehand, as well as the methodology, analysis and interpretation of findings. A summative index *relative* to the standard would be produced for the behavior measurement. In similar fashion, an average or summative figure would be obtained from the percentage increase in client attendance as a result of the intervention(s), and compared to the desired objective. The relative effectiveness (i.e., achieved/desired result) found in each component would also be averaged to arrive at the single indicator of "average performance." (See Appendix 2 for the mathematical calculation used in the example.) Clearly, a number of issues arise in considering this simple approach, notably the validity of assigning equal weights to each of the components, given that they stand at different levels of measurement (i.e., behavior precedes accomplishments). This and other issues surrounding measurement were discussed further in the meeting's concluding session.

Examples of performance-related measurements used by organizations and projects

Subsequent presentations expanded the discussion of the various methods and instruments used by different organizations and projects to measure provider performance or its related components. Most FP/RH organizations have concentrated their efforts on measuring aspects of the "competence" of the provider, the "quality" of services offered, and the "readiness" of clinics to offer such services.

As evaluation efforts are concentrated at the clinic or facility level, instruments have been developed to assess the care given at these facilities. Ruth Bessinger, from MEASURE *Evaluation*/DHS, presented the Quick Investigation of Quality (QIQ) tool, which "was created in response to the need for a low-cost, practical means to routinely measure quality of care of family planning services." The QIQ is a checklist made up of 25 items or indicators that collect information using three different methods:

1. a facility audit
2. observation of client-provider interactions and selected clinical procedures
3. exit interviews with clients leaving the facility (who had previously been observed).

Figure 5

Percentage of health facilities meeting basic infection prevention standards *Istanbul,Turkey*

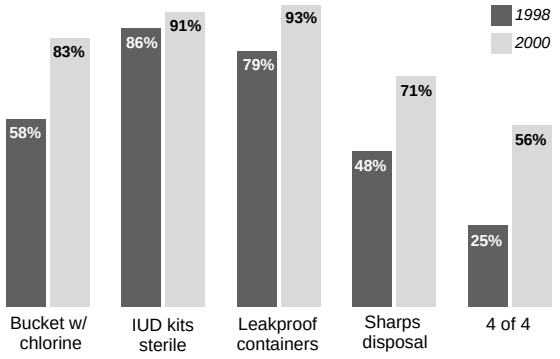
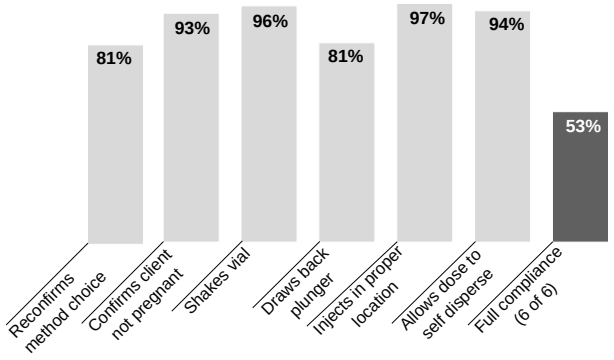


Figure 6

Provider compliance with clinical guidelines when administering injectable contraceptives *Uganda*



A condensed form of the QIQ can be examined in Appendix 3. The facility audit determines the "readiness" of the facility to provide quality services. The audit collects types of services provided, commodities in stock, availability of equipment and supplies, the operating condition of the facility, and the types of records kept. Observations assess the technical competence of providers in counseling and performing clinical procedures. Exit interviews collect information on the clients' experience at the facility and their perspective on the care received. MEASURE has applied the QIQ in a number of countries, expanding its scope to other reproductive health services such as prenatal care (Uganda) and postabortion care services, and using the tool to compare quality of care between intervention and control groups (Uganda) and different cadres of health care provider (Ecuador). Figure 5 and Figure 6 show the types of results obtained through the QIQ inventory and observation modules.

Fronczak presented a similar instrument, the Service Provision Assessment (SPA) tool, which also measures the "readiness or capacity" of a facility to provide services according to standard. The SPA uses the same three methods of data collection as the QIQ (facility inventory, observation of client-provider interaction and client exit interview), though its content goes beyond family planning and includes other reproductive and child health services. The SPA facility assessment encompasses systems that support provider performance, such as training and supervision, job aids and management practices for quality (contents of the information collected can be seen in Table 1).

Fronczak described some of the results of MEASURE *Evaluation*'s full application of the SPA in Kenya in 1999. A total of 512 observations were carried out in 89 facilities, complemented by facility-centered information gathered from 332 facilities. Among the results presented there was a good degree of consistency between findings from observations and those from inventories. Table 2, for example, shows the high degree of correlation between observation and facility audit assessments of the percentage of facilities reaching certain standards in prenatal care.

Erin Mielke from EngenderHealth presented a tool used by providers to assess quality of care based on client rights and staff needs. This Quality Management Tool (QMT) was tested in Tanzania in 1996 and 1999, and the changes in scores were used to show increases in "quality" of care (see Table 3).

However, as Bessinger pointed out, quality-related instruments such as the QIQ, SPA and QMT do not measure the entirety of provider performance because they do not address the outcomes of quality services. She offered the example of a family planning program that would "wish to improve the

Table 1

Information collected for each service assessed by the SPA

Capacity: Resources and Support Systems	
Availability of essential items for each service:	Basic equipment
	Advanced diagnostic equipment
	Staff and level of training
	Basic medications
	Higher level medications
	Protocols
	Client teaching materials
	Basic HIS register/records
	Client records
Type and functioning of support systems related to:	Supervision
	Equipment maintenance
	Infection control and disposal of hazardous waste
Provider Performance	
History: content, use of client record	
Basic examination	
Advanced examination	
Counseling content; use of visual aids	
Client recall related to history and examination	
Client recall related to counseling points	

quality of family planning services in order to reduce rates of method discontinuation by clients." For such measurements, "other data collection methods such as client follow-up or record reviews" would be needed.

After reviewing the main measurement and interpretation problems encountered with both direct observations of providers and client exit interviews (see Appendix 4), Federico León, from the Population Council (Perú office), presented the Service Test, a standardized method based on using simulated clients to evaluate provider behavior. After careful selection, training and supervision, simulated clients are asked to observe and recall provider behavior for between 46 and 72 different tasks. By using this service test to diagnose the quality of counseling and services to a simulated client asking for injectable contraception, a Population Council study was able to show contrasts between the two areas of service delivery (see Table 4).

Table 2

Comparison using observation vs. facility audits as unit of analysis for service provision practices *Kenya SPA 1999*

Item Assessed	*Observation*	*Facility Audit (compliance to standard)*	
	% (n=511)	50%	75%
Provide iron w/ folic	35	45	31
Provide antimalarial	3	7	3
Advice on nutrition	38	42	28
Take weight	98	96	94
Measure BP	83	85	82
Assess for edema	81	86	76

(adapted from Fronczak, 2001)

Table 3

Changes in the mean scores upon application of the QMT *Tanzania, 66 sites*

Items	*Mean Score (%)*		*Percent Increase*
	1996	*1999*	
Clients have a right to:			
Information	55.9	76.3	36.5
Access to Care	59.9	74.6	24.5
Informed Choice	52.5	71.4	36.0
Safe Services	64.7	83.1	28.4
Privacy and Confidentiality	74.9	92.3	23.2
Dignity, Comfort and Opinion	69.8	82.5	18.2
Continuity of Care	62.7	74.6	19.0
Staff need:			
Facilitative Supervision and Management	61.7	78.5	27.2
Information and Training	54.9	81.4	48.3
Supplies and Infrastructure	74.4	88.0	18.3
Total	62.7	79.4	26.6

(from Mielke, 2002)

León went on to explain how checklists and other job aids can be used to ensure standardization of desired quality of care. At the trial stage, instruments can also be assessed for their internal robustness and reliability. He gave as an example the *Balanced Counseling Strategy* job aid, which is being tested in Perú to ensure consistent high quality services (León, 2002).

Table 4

Example of the use of the service test for diagnostic purposes *León, 1999*

Provider's Expected Behavior	*Percentage Accomplishing*
General Questions Asked	
Date of last menstruation / suspicion of pregnancy	97
Does client want more children	7
Method Options Offered	
4 to 7 methods offered	84
Client asked to choose method	96
DMPA Use Instructions	
Following doses given every 3 months	93
Allowable window surrounding quarterly injection date is 2 weeks	4
DMPA Side Effects	
Menstruation might stop altogether	84
Temporary infertility of 6-12 months might follow stopping method	15
Instructions on Barrier Methods	
Condoms/vaginal tablets should be used while waiting for injection	18
Vaginal tablet must be inserted 15 minutes before coitus	54

(Utilization: Peru MOH reformulates counseling strategy in 2000. Simulated clients ask for counseling and choose DMPA, n=144 client-provider interactions in Peru.)

Defining and measuring provider performance

Because of the inherent difficulty of summarizing behavior and accomplishments in a single indicator, participants began by debating whether it is acceptable to measure just one component as representative of *performance*. Some participants suggested measuring only behavior when available data demonstrate that certain behavior leads to accomplishments, arguing that measuring activities and costs might be less expensive than collecting outcome data. A review of the methods and indicators used by organizations working in the FP/RH field makes it clear that they have indeed relied more on behavioral measurement. As one participant put it, "We've been better at measuring behavior. Perhaps we need to get better at measuring accomplishments if we're going to promote using the word performance."

Behavior-related data appear to be conceptually easier to obtain and interpret since these data are usually collected at the facility level. In cases where one provider is responsible for delivering FP/RH services at a facility, observing his/her behavior and relating it to interventions seems reasonably uncomplicated. This is not to say such measurements do not require much effort. On the contrary, methods such as direct observation and use of mystery or simulated clients demand careful selection and intensive training of observers, plus close supervision and quality control during data form completion. On this note, one participant contested the suggestion that behavioral measurement is inexpensive. In his view, "collecting data about behavior is incredibly expensive."

Participants also disputed the assertion that there was evidence linking behavior with outcomes. Though a number of recent studies demonstrate that improved quality can lead to increased use and continuation of use of services, these studies have incorporated a variety of elements of quality and have also shown other concomitant factors affecting client outcomes.

On the other hand, some participants advocated measuring and using only accomplishments to represent performance. This stance was quickly rebutted, however, as it was acknowledged that FP/RH services are social in nature, and that the mere measurement of outcomes might not capture how well these services were offered and delivered to the public. Also, an assertion of the relative low cost and ease of obtaining outcome data was challenged by a participant who argued that "adoption of [FP] methods is reasonably easy to measure; [but] continuation or discontinuation and contraceptive prevalence is more difficult." In effect, collecting data on discontinuation and actual use of contraceptives, if these are to be included among data to qualify program effectiveness, usually requires population-based tools and methods (e.g., careful application of questionnaires and sophisticated methods of data analysis such as life-table charts, disaggregated by reasons of discontinuation).

At this point one participant cautioned the group about the limitations that FP/RH training and service delivery projects experience regarding measuring outcomes, noting that their mandates typically do not allow them to collect population-based data (e.g., contraceptive prevalence). Thus, evaluators of these projects have had to find ways of collecting more easily accessible data, such as facility-based data. Even though there are issues surrounding the quality of these data, which limit comparison within and between countries, participants expressed the increasing need to find ways to tap into facility records. A side benefit of collecting such data might be improved quality of records as providers realize the importance of good record-keeping. Having and using good records will also help resolve the controversy on the relative merits of internal and external evaluation.

Participants reviewed the issue of *causality* when measuring accomplishments, and the difficulty of assigning a cause-effect relationship between behavior and accomplishments. Attribution of causes to effects presents another troublesome aspect of measuring accomplishments. The difficulty arises when evaluators suspect or know that there might be other concomitant interventions in the area of study. In such cases, measured outcomes may not be due to a single program's intervention but rather to a combination of interventions. Until research aids evaluators on how to assign proportional shares of attribution, it may only be possible to report the existence of these other interventions in the field and attempt to describe qualitatively their contribution to the measured outcomes. Another important difficulty arises when outcomes pertain to facilities with more than one provider. In such service delivery points, the actions of the other personnel may affect client accessibility and use of services in unknown ways. Attributing clinic outcomes to single providers may not be appropriate in such circumstances.

The participants agreed that agencies and projects have tended to measure behavior and accomplishments separately, without meaningful attempts to relate them. Because of the universal adoption of the Quality of Care paradigm, in recent years increased attention has been paid to measuring provider behavior and clinic conditions. A number of facility-based tools have been developed and applied; these have been used, however, mostly at an aggregate level. Similarly, measures of client and clinic outcomes have been documented, but not used in conjunction with or relation to the competency or skills shown by individual providers.

On the indicators selected to measure performance

Based on Elkins' presentation, participants discussed the best combination of types and numbers of indicators to measure provider performance. It was noted that performance indicators are often crafted in complex ways, making their interpretation and measurement difficult. Participants also stressed that organizations would benefit if they found ways to develop and promote indicators that are "useful on the ground" and originated by the users of services—i.e., less complete reliance on "top-down" indicators. For example, there may be occasions where the program's desired performance of providers or their facilities may not match the expectations held by local clients or potential consumers. These aspects need to be considered in future renditions of interventions directed at performance improvement and the monitoring and evaluation of those interventions.

Other participants expressed concern over the constraints that donor agencies impose on organizations and projects when frameworks are changed frequently, resulting in demands for new arrays of indicators. An emphasis was placed on the need to strive for fewer rather than ever-increasing numbers of indicators in evaluation plans while also acknowledging contexts in which the evaluator should take advantage of an opportunity to gather more data in an efficient way. For example, in regard to the SPA tool, one participant pointed out the benefit of having the team of interviewers present at the facility and able to ask additional questions of the providers.

Several participants expressed the idea that it is a mistake to use an indicator as a target. As one pointed out, "when you set a target, you don't know if a system has the capacity to meet it." Other drawbacks cited were the likelihood of low morale when targets are missed and the potential for falsification of data to meet a target. Inappropriate use of targets can lead to an excessive interest in "gaining in numbers but [resulting in] inappropriate care." Caution should also be exercised in instances where lower-level workers are asked to perform

higher-level tasks in the interest of meeting targets. Participants stressed that the appropriateness of who sets the targets is another important consideration. As one participant noted, the use of targets derives in part from the pioneering work of psychologist Ed Locke at the University of Maryland, and from the literature that ensued from his theory that institutions can motivate workers to improve performance by giving them specific, clear and ambitious goals (or targets).

Study design and methods for measuring provider performance

Provider performance can be measured using any conventional design method of evaluation. Several of these were mentioned during the meeting including the commonly used pre/post intervention design. Some participants advocated the use of time series designs using run charts and trend analysis, the preferred method of the QAP. Time series offers the particular strength of allowing project stakeholders to see changes in outcomes of interest associated with interventions within the clinic environment. Trend analysis, comparing measurements at different points in time, permits a clearer attribution of changes due to an intervention. However, a participant noted that time series design does not resolve the issue of contextual factors that may affect results in intervention areas. For example, thanks to the existence of a control group, evaluators of a project in Ecuador learned that a trend in IUD sales between July and December was not due to their intervention but was rather a seasonal effect. Other participants underscored the value of control groups in delimiting net effects. One participant argued, however, that the true nature of contextual factors affecting an intervention will, in fact, be revealed if run charts are continued over a long enough period.

It was also stressed that experimental or quasi-experimental designs are costly and more difficult to organize for projects that do not have an operations research mandate. There is also the issue of comparability of control groups and, increasingly, their availability in contexts of interventions involving multiple actors. Sample size is often a constraint, in particular when projects attempt to demonstrate small changes between groups. One participant mentioned the possibility of employing alternative evaluation designs, such as a long series of experiments in which different variables could be manipulated over time to assess degrees of change. However, these designs would likely be even more difficult to organize and fund.

There was also discussion of using other techniques, such as multivariate analysis, within evaluation designs to "sort out" factors associated with results. Participants noted the challenges that evaluators face at design and analysis stages when attempting to distinguish performance from its environment. Discussion centered on whether analysts should evaluate individual competence in the clinical setting or individual competence relative to the performance of the clinic and/or the system. Other challenges include controlling for system effects when evaluating provider performance; determining whether all critical steps should have equal weight; ascertaining the comparability of [performance] measurement when standards and guidelines for service providers are not uniform across countries; and deciding whether one-day observation is an accurate measure of provider performance, or, if not, how performance should be aggregated from individual providers to districts and systems. A couple of participants even maintained that it was impossible or irrelevant to measure individual performance, since "nothing worthwhile can be accomplished by a single individual."

The "contamination factor" was also brought up as an issue related to measuring provider performance. Trained and supported providers can affect the environment and those around them, including other providers, possibly diminishing the validity of comparisons and creating uncertainty about apparent differences among providers and attributions of effect to the interventions themselves. To avoid such potential contamination, one participant advocated the use of different districts when comparing providers. Another participant stressed that the ideal is to focus measurement on facilities, not providers.

Participants noted that oftentimes donors are unrealistic about the time and funding allocated to an intervention in relation to their expectations of measurable results. One participant related that "donors [are] looking for tremendous behavior change in two-year projects." These unrealistic expectations are particularly evident in the case of content areas such as maternal health for which

accomplishment indicators (e.g., maternal mortality) are not amenable to change except over a long period, and even then necessitate a large sample size to obtain valid measurement of indicators.

Regarding interpretation of observations, participants noted that most tools measure the existence of materials or whether or not a certain task was completed at the facility level. Often, however, they do not assess *per se* the quality or standard of care provided, which needs to be either interpreted from the data (e.g., adherence to standards) or assessed separately by better trained or "expert" observers who can judge the proficiency of each procedure or interaction. There was debate around this idea, since quality could be assessed from two different angles: whether critical tasks were completed or not, or how proficiently completion of tasks was carried out.

On the proposed construction of a single indicator to measure performance, the group acknowledged that insufficient research had been done to come up with a consensus on the issue. One participant suggested the possibility of multiplying rather than adding the components of performance into a single product in order to stress the necessity of the presence of both components for good performance (i.e., if one component was missing or "zero" then performance would also be null). However, other participants thought such a criterion or mechanism was too strict. One participant argued that there is a danger in multiplying quantity, since its product does not necessarily mean a valued result. Yet another participant noted that although a single measure was appealing it meant losing information, to which another participant replied that as long as the original factors are included, having a summary indicator should not pose a problem.

Reaching consensus

For the purposes of measuring FP/RH provider performance, the group adopted the HPT definition of performance—i.e., *both* the behavior and accomplishment components. Participants called for more research into ways of using both components of performance in a balanced way. They advocated for collaboration among agencies to achieve complementary goals. For example, some of the group members expressed the idea that service delivery projects measuring the quality or behavior aspect of performance could collaborate with research-oriented agencies that would measure the accomplishment component. One participant suggested that collaborating agencies could draft a proposal to test the components empirically through a joint initiative. If funded, such an initiative could lead to considerable advancement in the theoretical construct of performance. Participants expressed the need for increased dissemination of the issues discussed at the meeting in order to bring them to the attention of donor and policy-setting agencies. The group also agreed that methodologies must counteract the "insane" tradition of carrying out performance appraisals by drawing purely on the impressions of supervisors.

In summary, the meeting was considered to be a successful initiation of a much-needed dialogue between program and research-oriented agencies working in the field of FP/RH development assistance. The use of a standard definition of performance is seen as a critical step to avoid confusion and provide sounder and more comparable measurement in the future. Increased and targeted funding, improved interagency collaboration and more consistent dissemination were also deemed essential to improved measurement of provider performance in FP/RH programs.

Appendix 1

Client rights and provider needs

Client Rights
Information
Access to Services
Informed Choice
Safe Services
Privacy and Confidentiality
Dignity, Comfort, and Expression of Opinion
Continuity of Care
Provider Needs
Facilitative Supervision and Management
Information and Training
Supplies and Infrastructure

as per Huezo and Díaz, 1993

Appendix 2

Calculation of performance: Bangladesh example

Stakeholders decide paramedic performance should increase by 200% at the end of year one (i.e., triple their performance)

Performance: defined as combination of CSI/ORH skills and # clients seen in clinics

Assumptions:

- One paramedic per clinic, supported
- Equal weights to all components
- No concomitant interventions

Calculation of Overall Performance

$$\text{(Average) Performance} = \frac{\text{\% change behavior + \% change accomplishments}}{2}$$

$$\text{\% change behavior} = \frac{\text{\% change CSI + \% change ORH}}{2} = \left[\left(\frac{\%\Delta 85}{45}\right) + \left(\frac{\%\Delta 82}{17}\right) / 2\right]$$

$$= \frac{88.9 + 382.4}{2} = \mathbf{235.7\%}$$

$$\text{\% change after accomplishment} = \text{\% change client attendance at 12 months after intervention}$$

$$= \frac{\%\ \Delta 175}{61} = \mathbf{180.3\%}$$

$$\text{\% change performance} = 235.7 + \frac{180.3}{2} = \mathbf{208\%}$$

(where Δ = "change," CSI = Child Survival Interventions, ORH = Other Reproductive Health services – see main text for full descriptions)

Conclusion: The Bangladesh paramedics <u>did</u> improve their performance, as expected

Appendix 3

Short list of QIQ indicators

Indicator Number	*Indicator*	*Client Exit Interview*	*Observation*	*Facility Audit*
Provider				
I-1	Demonstrates good counseling skills (composite)	X	X	
I-2	Assures client confidentiality		X	
I-3	Asks client about reproductive intentions (More children? When?)	X	X	
I-4	Discusses with client what method she would prefer	X	X	
I-5	Mentions HIV/AIDS (initiates or responds)	X	X	
I-6	Discusses dual method use	X	X	
I-7	Treats client with respect/courtesy	X	X	
I-8	Tailors key information to the particular needs of the specific client	X		
I-9	Gives accurate information on the method accepted (how to use, side effects, complications)	X	X	
I-10	Gives instructions on when to return	X	X	
I-11	Follows infection control procedures outlined in guidelines		X	
I-12	Recognizes/identifies contraindication consistent with guidelines		X	
I-13	Performs clinical procedures according to guidelines		X	
Staff				
I-14	Treat clients with dignity and respect	X		
Client				
I-15	Participates actively in discussion and selection of method (is "empowered")	X	X	
I-16	Receives her method of choice	X	X	
I-17	Client believes the provider will keep her information confidential	X		
Facility				
I-18	Has all (approved) methods available; no stockouts			X
I-19	Has basic items needed for delivery of methods available through SDP (sterilizing equipment, gloves, blood pressure cuff, specula, adequate lighting, water)			X
I-20	Offers privacy for pelvic exam/IUD insertion (no one can see)	X	X	X
I-21	Has mechanisms to make programmatic changes based on client feedback			X
I-22	Has received a supervisory visit in the past ___ months			X
I-23	Adequate storage of contraceptives and medicines (away from water, heat, direct sunlight) is on premises			X
I-24	Has state-of-the-art clinical guidelines			X
I-25	Waiting time is acceptable	X		X

Source: Adapted from the QIQ instrument. See MEASURE Evaluation website.

Appendix 4

Main measurement and interpretation problems encountered with methods to assess the quality of care

Direct Observations	*Client Exit Interviews*
Diversity of client profiles (how to control for it)	Clients are unprepared to observe
Provider will be at maximum, not typical performance	Clients are asked to judge over abstract/unfamiliar categories
Client behavior will always be affected in unknown ways	There are memory and recollection problems for specific provider behaviors

PRIME II

The PRIME II Project
Intrah
School of Medicine
The University of North Carolina at Chapel Hill
1700 Airport Road, Suite 300
CB #8100
Chapel Hill, NC 27599-8100
Tel: 919-966-5636 Fax: 919-966-6816
Intrah@intrah.org www.prime2.org

PRIME II Partnership: Intrah, Abt Associates, EngenderHealth, Program for Appropriate Technology in Health (PATH), Training Resources Group, Inc. (TRG) with supporting institutions, the American College of Nurse-Midwives and Save the Children.

This publication was produced by Intrah at the University of North Carolina at Chapel Hill for the PRIME II Project and was made possible through support provided by the Center for Population, Health and Nutrition, Global Bureau, U.S. Agency for International Development, under the terms of Grant No. HRN-A-00-99-00022-00. The views expressed in this document are those of the authors and do not necessarily reflect the views of the U.S. Agency for International Development.

This publication was also made possible through the support of MEASURE Evaluation, funded by 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.