Measuring Performance: A Strategy to Improve Programs[1,2]

Rebecca J. Stoltzfus*[3] and Gita Pillai*†

*Center for Human Nutrition, Department of International Health, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, MD 21205; and the †Department of Population and Family Health Sciences, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, MD 21205

ABSTRACT This article stresses the importance of evaluation as a tool for improving the design and implementation of effective programs to reduce iron deficiency anemia and to advocate for their continued and increased support. Current concepts in program evaluation are applied to the specific issues relevant to iron programs. Evaluations should be designed to meet the needs of specific audiences (e.g., community members, program implementers, policymakers, donors and researchers) and to answer specific questions. Evaluations might answer questions about provision, coverage or impact. The choice of indicators for evaluating impact of iron programs is discussed and illustrated with recent examples. Evaluation design can be broadly categorized as monitoring, plausibility evaluations and probability evaluations. These designs differ in cost and also in the strength of evidence that they provide; however, each has appropriate uses. It is important to document program costs in the evaluation process as policymakers and donors weight impact against costs. To be useful, evaluation findings must be disseminated, usually to multiple audiences. This requires allocation of time and resources and attention to the needs of various stakeholders. J. Nutr. 132: 845S–848S, 2002.

KEY WORDS: • *anemia* • *iron deficiency* • *program evaluation* • *research design* • *public health*

"Feedback is the breakfast of champions," a quote used to inspire personal effectiveness and leadership, is equally relevant to public health efforts. Knowing what is working and what is not working and considering corrective action make a healthy and necessary "breakfast." This breakfast may or may not taste good on all days but will nurture continuous improvements and drive systematic and evidence-based contributions to the common goal of improving health status.

"Any program worth implementing is worth evaluating" is cited as common sense but does not necessarily reflect common practice. The financial resources and technical expertise required for evaluation can be a barrier to planning and budgeting evaluations, as well as subsequent implementation. However, program implementation without evaluation is also costly. Programs without evaluations leave no lessons for future programs and perpetuate the wasteful expenditure of resources on untested strategies that may have limited added value to improving health. Moreover, they represent a missed opportunity to generate information needed to advocate for and leverage resources and attention for addressing iron deficiency anemia (IDA).[4]

In his presentation, Key Policy Issues for Iron Deficiency Control and Prevention, Ray Yip emphasized the importance of evidence-based rather than paradigm-based programming. In fact, several participants at this meeting have noted our tendency to design and implement programs based on beliefs or paradigms of sound programming rather than evidence. To achieve our goal of improving the health of women, children and families, we need to verify that our interventions and strategies are indeed evidence-based.

Habicht et al. (1) in an article on program evaluation suggest that we start with three central questions for every evaluation: Who will the evaluation inform?; What questions will the evaluation answer?; and How will you conduct the evaluation to generate useful answers? We address each of these questions in the context of iron deficiency control programs.

Who will review and use evaluation results and recommendations?

Different types of decision makers require different information, and in different forms, to guide their decisions and actions. Policymakers usually rely on digested and translated information on effectiveness and potential large-scale impact

[1] Presented at the Atlanta conference on Forging Effective Strategies to Combat Iron Deficiency held May 7–9, 2001 in Atlanta, GA. The proceedings of this conference are published as a supplement to *The Journal of Nutrition.* Supplement guest editors were Frederick Trowbridge, Trowbridge & Associates, Inc., Decatur, GA and Reynaldo Martorell, Rollins School of Public Health, Emory University, Atlanta, GA.

[2] This article was commissioned by the International Life Sciences Institute Center for Health Promotion (ILSI CHP). The use of trade names and commercial sources in this document is for purposes of identification only and does not imply endorsement. In addition, the views expressed herein are those of the individual authors and/or their organizations and do not necessarily reflect those of ILSI CHP.

[3] To whom correspondence should be addressed. Dept of International Health, School of Hygiene and Public Health, The Johns Hopkins University, 615 N. Wolfe St, Room W2041, Baltimore, MD 21205.
E-mail: rstoltzf@jhsph.edu

[4] Abbreviation used: IDA, iron deficiency anemia.

before advocating for policy change. Donors use evaluation results to justify expenditure and determine future allocations. Researchers depend on evaluation results to evolve scientific knowledge and future inquiry. Program managers need to know whether a particular strategy worked and the details of how and why it worked, when planning program adaptation and replication. Community groups and leaders involved in implementation use evaluation results to prioritize actions and responses. Last and most importantly, the woman who takes an iron supplement or feeds her child a fortified food uses the results to evaluate the benefits of consumption against the cost of accessing supplies and risks of non-consumption to sustain her healthy behavior. Self-monitoring tools that provide feedback and reminders for daily intake can reinforce supplementation behavior. At each of these levels, evaluation results are used to sustain and improve actions that contribute to the control of IDA.

What questions need to be answered by the evaluation?

Evaluation questions may be grouped into three categories: questions about provision, coverage and impact (1). Questions about provision refer to the availability, reliability and quality of services and supplies. Iron program examples include:

- Is there sufficient stock of iron-folic acid tablets?
- How many tablets were distributed in the last month?
- What proportion of providers are trained to explain and support correct intake?
- Is there a distribution site within 30 min of most households?

These questions are key for health facility and systems assessment. There is a caveat, however, to placing emphasis on provision indicators. Numerous programs implemented by private and government organizations successfully monitor and achieve improved provisioning but demonstrate no improvement in coverage or health impact. In most of these cases, staff focuses on achieving and recording a completed number of training programs or supplies distributed without sufficient attention to the utilization of services or practice of healthy behaviors most closely associated with health impact. Although indicators of provisioning can be useful to assess the process of implementation, overemphasis on provisioning indicators as a measure of performance can be detrimental to achieving program coverage and impact. During an assessment of programs implemented by CARE-India, Stan Foster commented: "You get what you inspect (provisioning), not what you expect (coverage and impact)," as a means of stimulating program staff to shift their attention to planning and achieving improved population-based coverage rates.

Coverage refers to the percentage of eligible people who receive the service or supply or who practice a promoted behavior. For example, in the case of iron supplementation, this might be: the percentage of women who consume iron tablets for at least 100 d before delivery or the percentage of pregnant women who had at least three antenatal care contacts.

Coverage rates may be facility-based, with the denominator being all those who came to the health center, or population-based, where the denominator is the entire eligible population that live within particular geographic catchments. The latter is more difficult to achieve because it requires a population-based denominator but is a better reflection of public health impact. Coverage rates are needed to find out whether successful provisioning in fact improved the practice of healthy behavior.

Questions of impact refer to changes in the prevalence of deficiency, disease, disability or mortality and usually involve biological markers. In the case of IDA, impact indicators include measures of iron deficiency, anemia or other measures of functional status, such as child development or adult productivity. In the context of programs, a logical progression is usually assumed, where supply and service provisioning is expected to improve coverage rates of healthy practice, which in turn improves health status. However, when evaluating outcomes, achievement is sometimes evident in one but not the other. For example, consider the finding where reported coverage of iron intake is high, but there is no measurable decline in the prevalence of anemia. This can occur with the over-reporting of intake, or if malaria or hookworm is endemic but not addressed by the program.

When evaluating impact, the best choice of indicator is often a function of feasibility, cost, required sample size and policy importance. As depicted in **Table 1**, anemia or hemoglobin assessment is usually the indicator chosen because of its affordability, ease and relatively low sample size and because we often assume that anemia will be highly responsive to iron fortification programs. However, as also depicted in Table 1, the potential impact of iron interventions on anemia is highly variable. This is a function of the multiple and varying etiologies of anemia in different environments.

Evidence from systematic reviews of the consequences of iron deficiency (2) suggest that indicators of iron deficiency are sometimes more sensitive and specific and, therefore, more powerful than hemoglobin in demonstrating the impact of iron interventions. A recent example from Zanzibar provides a good illustration. In a placebo-controlled trial of low-dose iron supplementation to preschool children (3), iron deficiency was significantly decreased, but anemia remained unchanged by the intervention. From this information alone, we might have concluded that the impact was, therefore, disappointing. However, we also assessed children's motor and language development and found that iron-supplemented children achieved significantly more milestones in the year-long intervention period, despite the lack of impact on anemia. This example highlights the importance of using multiple measures when possible and of connecting biomarkers of nutritional impact to cognitive developmental, productivity or health outcomes. It is possible that over-reliance on anemia to assess impact of iron interventions—especially in environments where anemia has multiple causes—leads us to significantly underestimate the positive impact of our interventions.

TABLE 1

Typical expectations of different types of impact measures for iron program evaluation

Impact measure	Potential iron impact	Ease	Sample size	Policy importance
Iron deficiency	100%	++	+	Low*
Anemia	20–80%	++++	++	Moderate
Functional outcome	5–20%	+	++++	High

* Potentially much higher if combined with advocacy.
\+ Increasing ease or sample size.

How to conduct the evaluation: evaluation designs

Habicht et al. (1) conceptualize adequacy, plausibility and probability evaluation on a continuum of increasingly robust designs and confident results (**Fig. 1**). Unfortunately, increasing confidence usually comes with increasing cost. The methods of evaluation should generate answers to questions with sufficient confidence to meet the needs of the decision makers.

Monitoring [also called adequacy evaluation (1)] involves the intermittent measurement of indicators over time. For example, the United States aims to reduce the prevalence of neural tube defects by increasing the median red cell folate to 220 ng/mL in non-pregnant women by 2010 with mandatory folate fortification of grain products. Recently published data (4) provide an excellent example of monitoring as a form of program evaluation. The distributions of red cell folate values in American women of reproductive age before and after the fortification policy was implemented in 1998 show that the goal has been achieved within 2 y of program implementation (**Fig. 2**). Although monitoring is depicted at the "low confidence" end of the continuum in Figure 1, this example illustrates the potential strength of monitoring evidence. In this example, we can conclude with reasonable confidence that the shift in the distribution was due to the program because: the assay used at both times was comparable; the United States population was relatively stable over the period; the sample was constructed similarly at both times; the program was implemented at a well-specified point in time between the two surveys; and there were no other obvious changes in dietary folate intake in this population during the same period. These are salient points when considering the potential strength of monitoring as an evaluation design.

A plausibility evaluation builds a reasonable argument for causality without a randomized trial. Many program evaluations fall into this category, where baseline and end-line measures are compared without a control group, or non-randomly allocated program-exposed and unexposed groups are compared. The inclusion of multiple measures of provisioning, coverage and impact allows for triangulation of data, which is very useful in building a plausible explanation for program performance.

A probability evaluation is less common and usually requires more expertise and funds but generates results and conclusions about program performance with the greatest confidence. In a probability design, program and non-program groups or areas are randomly assigned and cause and effect are established after controlling for potential confounding factors. A recent example of a probability evaluation of coverage and impact is a field trial of antenatal iron supplementation and anthelminthic (mebendazole) treatment implemented by CARE-India in Orissa, India (5). Iron and mebendazole intake were known to be low and the prevalence of anemia was high. In the program area only, CARE-India staff introduced innovations to the government of India anemia control program to improve receipt and reduce forgetfulness once pills were received, while the non-program area continued with existing government services and supplies. Surveys in the program and non-program areas before and after the program changes were implemented provided data to evaluate coverage and impact.

After controlling for background factors, there was significantly higher iron and mebendazole receipt and consumption and a significantly lower prevalence of anemia in the group exposed to the program strategy. These data also illustrate that the impact of iron tends to be more substantial on more severe anemia. The intervention impact on mild anemia (hemoglobin <110 g/L) was a 10% reduction (70.9% in the program area vs. 78.5% in the control area), but the prevalence of more severe anemia (hemoglobin <90 g/L) was reduced by 30% (24.5% vs. 34.9%). Moreover, there was a clear dose–response relationship between the milligrams of iron and mebendazole intake and anemia status (**Fig. 3**), adding plausibility to the argument that these interventions decreased the prevalence of anemia.

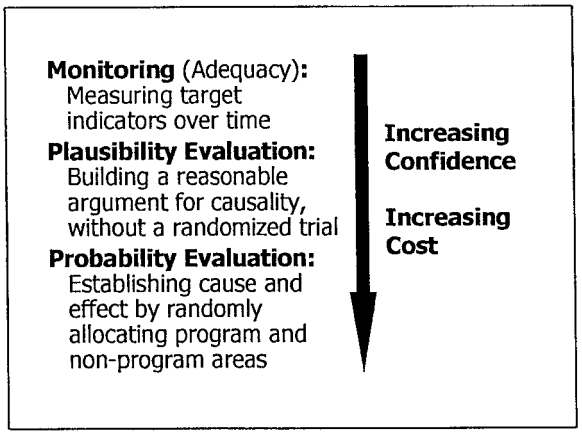

FIGURE 1 Designs for program evaluations. Adapted from Habicht et al. (1).

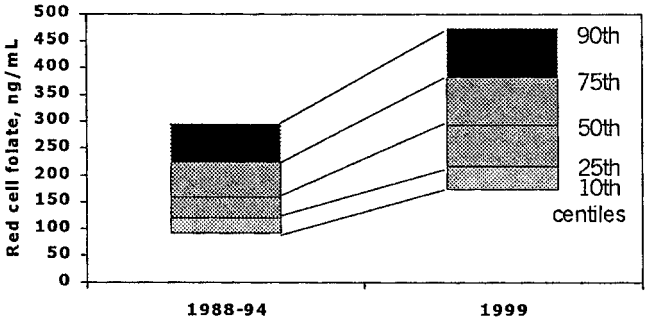

FIGURE 2 Distributions of red cell folate concentrations in American women of reproductive age before and after the start of the folate fortification policy in 1998. The lines indicate percentiles of the distribution, as labeled. These data were taken from Anonymous (2).

Importance of cost

An essential part of program evaluation is documentation of program costs. Program managers and policymakers need

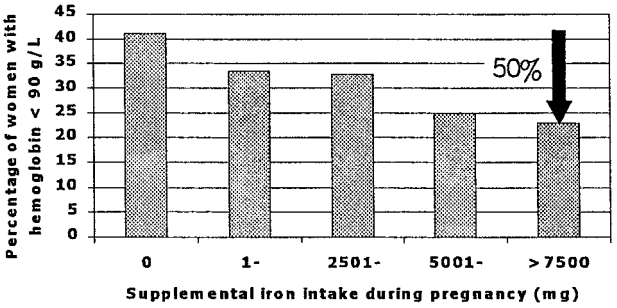

FIGURE 3 The proportion of pregnant women with hemoglobin concentration < 90 g/L by level of supplemental iron intake during pregnancy in Orissa State, India. These data were taken from Pillai (5).

information about the cost of achieving reported coverage or impact before they will endorse adaptation or wider replication. Knowing the incremental gains and associated costs of each intervention or strategy component can be especially valuable when selecting specific components of a strategy to adapt or scale-up.

Disseminating evaluation findings and recommendations

Communication of program evaluation findings is vital to improve programs and to advocate for continued or additional program support. There is value in involving stakeholders in data collection, analysis and interpretation if it fosters greater ownership and acceptability of findings and follow-up action. However, this benefit needs to be weighed against the risk of compromising objectivity or biasing results. Different constituencies of the program require information in difference formats and styles. Often multiple forms of communication are needed to facilitate use of the information by policymakers, donors, program staff, community-implementing partners and scientists. Dissemination of evaluation findings in a variety of appropriate formats is time-consuming, sometimes tedious and often underfunded, but vitally important to influence continuous improvement in our programs and their impact. Planning realistically for the time and support needed for dissemination of evaluation makes it a much more rewarding task.

LITERATURE CITED

1. Habicht, J.-P., Victora, C. G. & Vaughan, J. P. (1999) Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. Int. J. Epidemiol. 28: 10–18.

2. Stoltzfus, R. J. (2001) Iron-deficiency anemia: reexamining the nature and magnitude of the public health problem: summary: implications for research and programs. J. Nutr. 131: 697S–700S.

3. Stoltzfus, R. J., Kvalsvig, J. D., Chwaya, H. M., Montresor, A., Albonico, M., Tielsch, J. M., Savioli, L. & Pollitt, E. (2001) Effects of iron supplementation and anthelminthic treatment on motor and language development of Zanzibari preschool children. Br. Med. J. 323: 1389–1393.

4. Anonymous. (1999) Folate status in women of childbearing age–United States, 1999. Morbid. Mortal. Wkly. Rep. 49: 962–965.

5. Pillai, G. (2001) Improving Women's Health: An Assessment of Community-Based Strategies to Increase Iron and Mebenazole Intake and Reduce Maternal Anemia in Orissa, India. Doctoral thesis, The Johns Hopkins University, Baltimore, MD.