Science-Based Evaluations for Program Improvement

The Challenge
Capacity development and learning are key to the World Bank's success in poverty reduction, and the Bank supports both. In FY04, 1,016 capacity development activities were delivered to 78,500 clients, and 3,300 learning activities were delivered to 8,250 Bank staff. These numbers represent significant growth over previous years. Despite the size and importance of these activities, evidence is limited regarding their intermediate- or long-term results. Before 2001, evaluations typically focused on participants' satisfaction, and program suppliers themselves conducted most evaluations of their programs. Given the importance of capacity development in achieving the Bank's goals, rigorous and objective evaluations of these learning programs are both prudent and desirable.

World Bank Institute Intervention
The World Bank Institute Evaluation Group (IEG) evaluates capacity development and learning, with reporting responsibility to both the World Bank Institute's (WBI's) Vice President and the Bank's Chief Learning Officer. With increased focus on results in the development community, IEG began an intensive effort in late FY01 to introduce more varied evaluation techniques, which led to greater scientific rigor. Evaluations expanded from measures of learning satisfaction and knowledge gain to include assessments of the impact of the activities, and evaluation findings are more broadly available now than in past years. IEG evaluations are formally considered self-evaluations because IEG is situated within WBI and the Bank. However, because these evaluations use science-based methods, they yield valid and reliable assessments of program effectiveness and outcomes and provide information to stakeholders about the features of effective learning programs.

Underlying Assumptions
Four assumptions drive the emphasis on increasing the scientific rigor of the evaluations:

1. The use of science-based methods improves the validity of evaluation studies—that is, the correspondence between the evaluation findings and program outcomes.
2. Science-based methods improve reliability, which means that quantitative results can be reproduced when repeated measurements are taken, or that different experts can look at the same information and arrive at approximately the same judgments.
3. Conducting science-based evaluations and publishing their findings promotes transparency.
4. Greater validity, reliability, and transparency are important dimensions of independent evaluations that are transferable to self-evaluations.

Evaluation Methods and Instruments
The report summarized in this brief describes the evolution of science-based self-evaluation carried out by IEG staff from FY02

---

2. The Organization for Economic Co-operation and Development (OECD) defines self-evaluation as "an evaluation by those who are entrusted with the design and delivery of a development intervention." Institutional research and evaluation therefore fall into the category of self-evaluations because these evaluation offices report to those who are also responsible for the implementation of learning programs, even though they are not the direct implementors.
to FY04, in five areas: (a) using mixed methods, (b) measuring learning, (c) establishing counterfactuals, (d) sampling, and (e) analysis.3

**Mixed Methods.** Evaluation experts recommend using multiple methods and multiple data sources to confirm findings and to expand the types of evaluation questions that are addressed. All major IEG evaluations relied on mixed methods combining two or more of surveys, interviews, focus groups, product assessments, retrieval of archival data, document review, observations of training, and tests. As a result, the share of evaluations using mixed methods increased sharply in FY03-04 compared with the previous two years (see Figure 1).

**Measuring Learning.** Beginning in FY02, IEG evaluators have employed professional standards for test development. They started with a two-dimensional test specification matrix with content topics and estimated level of difficulty. This matrix guided the development of test questions to ensure that pretests and posttests for the same course are comparable. Evaluators then began helping WBI course providers improve test questions through the use of statistical item analysis, increasing the number of items to make the test more representative of course content than it would be with fewer items, and offering expertise in test development. These efforts measurably increased the internal reliability of WBI’s learning measures.4

IEG also worked to improve tests for World Bank staff learning and to develop a “high stakes” certification test for staff managing trust funds. As a result, the reliability of tests used in evaluating staff learning also improved, to meet professional standards.5

**Counterfactuals.** The new approach to evaluation also established counterfactuals—that is, “control” groups—to assess the effectiveness of learning programs by comparing the results of an intervention with results of a similar situation where the intervention had not occurred. IEG used two quasi-experimental techniques for creating control groups: (a) postlearning matching of participants with nonparticipants using either propensity-score matching or other types of trait-by-trait matching techniques,6 and (b) prelearning matching of participants with similar nonparticipants scheduled for subsequent participation (see Figure 2).

**Sampling.** As the reach of both WBI and staff learning programs grow, the previous practice of conducting tracer studies of total populations of participants in specific courses becomes infeasible. IEG developed sampling strategies to support effective evaluation, as illustrated by the three examples in Figure 3. Most approaches

---

4. From Cronbach’s alpha of 0.41 for tests developed without IEG assistance to 0.56 for tests developed with this assistance.
5. Cronbach’s alpha of 0.83 for the certification test and an average of 0.82 for other tests; professional standards call for internal consistency reliability to be higher than 0.80.
6. Propensity score matching estimates the probability of an individual’s participation in a learning event; propensity score matching can be done when data exist for a large number of individuals, only some of whom have participated in a learning event.
Figure 3. Examples of the WBIEG’s Sampling Strategies

Case I: Evaluation of Staff Learning

Courses Meeting Eligibility Criteria
- In Learning Catalogue
- At least half of attendees are Bank staff
- Not e-learning
- Not yet evaluated in the Fiscal Year

Random selection on a biweekly basis
Random Sample 242 Staff Courses

Case II: Evaluation of Six Sectoral and Thematic Programs

Six Programs Random samples of learning activities selected from each program
Random Sample 25 Learning Activities 50 participants randomly selected from each activity
Random Sample 1225 Participants

Case III: A Country-Focused Evaluation

Five Countries Participants selected from each country.
Random samples selected when n>200.
All participants selected when n<200.
Sample 1200 Participants

used stratified random sampling, and most required the use of weights in statistical analyses.7

Analysis. With more sophisticated analytic techniques, evaluations moved beyond the reliance on (a) descriptive statistics for single learning events, (b) summaries of informal interviews and focus groups, and (c) assessments of learning. The use of counterfactuals allows for simple tests of program effectiveness: were there measured differences in outcomes for participants as compared with nonparticipants? In addition, better measurement instruments and scientific sampling permit evaluators to use multivariate analyses to explore features of learning programs that are associated with positive outcomes (e.g., learning or perceived effectiveness) or that allow for testing of hypotheses regarding effectiveness for groups of participants.

Evaluation Results:
Program Effectiveness
Science-based evaluations collectively yield robust findings across staff learning and WBI activities. Comparison groups enable objective, detailed assessment of program effectiveness. A review of five IEG evaluations in which a comparison group had been constructed yielded relatively consistent results:

1. Independently evaluated measures of performance—scores on tests and blind ratings by independent observers—are more sensitive to changes caused by courses than are performance ratings by program participants’ supervisors. Manager ratings of participants are not significantly different from manager ratings of nonparticipants.

2. Self-reports combined with counterfactuals reveal program effects. Participants report higher use and sustained use of knowledge imparted by the course they attended than did matched nonparticipants (who may have attained the knowledge elsewhere), with effect sizes of 0.25 or more.

3. Multiple methods often revealed comparable findings. Interviews and focus groups with participants in higher-rated courses generally reveal positive assessments of the same courses for relevance, usefulness, and application. Interviews also detected differences between the views of participants and nonparticipants.

7. Stratified random sampling divides the population into subgroups or strata and draws random samples from each stratum.
Features of Effective Learning Programs
Findings from five meta-analyses show remarkable consistency regarding the features of effective learning programs for adults:

4. Four dimensions of course quality are consistently associated with positive immediate or mid-term outcomes: longer course duration, professional design, the use of active learning methods, and follow-up. These findings are aligned with current wisdom on course outcomes. Time for learning is recognized as an important element of all training. Professional design is associated with courses that are well conceived in advance. Action learning, with participants encouraged to develop action plans, is another positive feature in both client and staff learning. Follow-up provides reinforcement as participants implement what they learned.

5. Dimensions of the course related to how closely a course fits participants’ needs are also positively associated with outcomes: alignment of the course to the participant’s job, homogeneity of participants, and alignment with country needs. Again, these results are not surprising, given that when the content of a course matches learners’ needs, learners are more likely to benefit.

Implications
The use of science-based methods in the evaluation of staff learning and client capacity development programs at the World Bank contributes to program improvements in three ways:

- **Introduction of counterfactuals allows for the direct assessment of program effectiveness.** Attribution of effects to a program is plausible when counterfactuals are available. Effective programs can be distinguished from less effective programs on the basis of their value-added to participants’ knowledge, skills, and behaviors. Differences between participants and comparison groups were observed on some indicators only. However, on indicators most closely associated with the program—learning and team products—differences typically were observed.

- **Use of multivariate analytic techniques helps to identify features of effective programs.** Improvements in program design, based on evidence, are therefore possible. In some cases, program designers have used early indications of effective program features to inform the design process.8

- **Multivariate techniques also help target capacity development and learning to appropriate individuals, based on evidence.** Alignment of the course to the overall participant group, the participants’ jobs, and to their country builds a closer match between the course content and the needs of adult learners.

Acknowledgement
The evaluation activities summarized in this brief were conducted with the financial support of the World Bank Institute and the World Bank’s Learning Board.

---

8. For example, the share of WBI activities with “action planning” has increased to 56 percent in FY04 from 38 percent in FY00-FY01.