IMF Publication

New Common Evaluation Framework for IMF Capacity Development

INTERNATIONAL MONETARY FUND

INTERNATIONAL MONETARY FUND

April 2017

NEW COMMON EVALUATION FRAMEWORK FOR IMF CAPACITY DEVELOPMENT

IMF staff regularly produces papers proposing new IMF policies, exploring options for reform, or reviewing existing IMF policies and operations. The Report prepared by IMF staff and completed on April 7, 2017 has been released.

International Monetary Fund
Washington, D.C.

INTERNATIONAL MONETARY FUND

NEW COMMON EVALUATION FRAMEWORK FOR IMF CAPACITY DEVELOPMENT

April 7, 2017

Approved By
Sharmini Coorey

Prepared by the Institute for Capacity Development in consultation with other departments. The note was prepared by ICD's evaluation team: Ms. Gust (team lead), Ms. Bal Gündüz, and Messrs. Million and Warner, under the guidance of Messrs. Desruelle and Powell.

EXECUTIVE SUMMARY

This document outlines a new common evaluation framework for the Fund's capacity development (CD) activities. The new common evaluation framework is intended to streamline current practices and increase comparability and use of results by adopting for all CD evaluations a common four-step process that includes use of the OECD Development Assistance Committee (DAC) evaluation criteria. Around this common approach, the proposals allow flexibility to adapt evaluations to reflect the wide range of CD activities. Key elements of the framework, shown in the table below, are grouped around the objectives of:

- producing shorter, more focused, and more comparable evaluations;
- improving the information supporting evaluations;
- spending the same level of resources on evaluations while allocating these scarce resources more efficiently; and
- using the information from evaluations to alter practices or shift the targeting of CD resources.

Produce shorter, more focused, and more comparable evaluations:

- Specify clearly the objectives of all CD activities to be evaluated
- Apply a four-step common framework (thought process) to all evaluations
- Ensure clarity and consistency in definition and application of internationally-accepted OECD DAC evaluation criteria
- Use standardized Terms of Reference templates to reduce workload and sharpen the focus of evaluations.

Improve the information supporting evaluations:

- Consistent with the new results focus of all IMF CD, ensure that all CD activities have a data collection plan in place before the start of the CD activity
- Obtain ex-post information from TA providers and recipients and other stakeholders
- Extend the pre- and post-course tests to all IMF training courses
- Revise the end-of-course and follow-up survey questions for training to be more results oriented
- Maintain a channel for regular reviews of training course content
- Seek information from IMF country teams on the extent to which they observe tools learned in IMF training being applied at the country level

Maintain the current level of resources allocated to evaluation but use these scarce evaluation resources more efficiently:

- Establish reasonable accountability standards without evaluating everything
- Introduce a CD evaluation work plan with a rolling three-year horizon, revised annually
- Decide what and how to evaluate based on potential value of the information, cost, and achievement of accountability standards
- Maintain the flexibility to conduct mid-term or rapid evaluations of TA projects

Use the information from evaluations to alter practices and/or shift resources:

- Introduce quantitative scoring to complement qualitative information in evaluations to facilitate comparisons and aggregation
- Ensure evaluations are easily accessible upon completion and findings are broadly shared among Fund staff
- The November CCB meetings would be expected to use evaluation results when setting CD priorities

I. CONTEXT

1. **Regular evaluation is a crucial component of a sound capacity development strategy to foster learning from past experiences and enhance accountability.** At the Board discussion of *The Fund's Capacity Development Strategy—Better Policies through Stronger Institutions* (2013), Executive Directors endorsed strengthening the Fund's monitoring and evaluation framework to better incorporate feedback from evaluation results into the prioritization and delivery of technical assistance and training (collectively called capacity development, CD). They also saw merit in a unified approach to evaluation that would help distill lessons, including through independent external assessments as appropriate.

2. **Since then, ICD has been working with departments to develop proposals to improve the Fund's evaluation framework for capacity development.** ICD has discussed current practices with the technical assistance (TA) providing, area, and other departments. Strengths of the current system include regular evaluation of both TA and training activities, regardless of funding source, and the opportunity for these results to influence future CD delivery. However, there are some important weaknesses. TA evaluations are often long, unfocused documents, done as a matter of routine requirement, and many believe that important questions are sometimes left unevaluated in the current system. For training, there is more support amongst staff for current practices, but many believe the current evaluations still do not provide enough information, again leaving important questions unanswered.

3. **The changes summarized here are proposed with an eye to feasibility and cost considerations.** Pointing out gaps in the current system does not necessarily mean that action must be taken. Consideration must also be given to time, costs, feasibility and whether any proposed change will have a high enough probability of yielding better, actionable, information. The aim of the proposals is to preserve the current strengths of the Fund's evaluation activities while addressing weaknesses.

4. **This document reviews the current evaluation framework and presents a new common evaluation framework for Fund CD.** Section II describes the existing evaluation framework. Section III considers the strengths and shortcomings of the current framework, while Section IV presents the new common framework. A companion guidance note, also being prepared, will describe the methodology in more detail and provide specific operational guidance on how to apply the proposed four-step common evaluation framework to TA and training.[1]

[1] See Annex 1 for a summary of the four steps in the Common Evaluation Framework.

II. THE EXISTING FRAMEWORK: WHAT IS CURRENTLY DONE?

A. TA Evaluations

5. **The Fund undertakes regular internal and external evaluations of its TA, which include:**

- Fund-wide evaluations;
- External evaluations of Regional Technical Assistance Centers (RTACs), Topical Trust Funds (TTFs), and some bilateral accounts; and
- Self-assessment evaluations conducted by TA departments.

6. **Fund-wide evaluations are done to assess overall capacity development (CD) policies and activities.** Some are done on an *ad hoc* basis, for example, the IEO evaluation of TA in 2005, which paid special attention to the relevance and effectiveness of Fund TA and how to enhance ownership. At present, the main vehicle for the Executive Board to assess overall CD policies and activities is the regular review of the Fund's Capacity Development strategy, expected to take place every five years.[2]

7. **Evaluations of RTACs, TTFs, and some bilateral subaccounts are conducted by external evaluators.** Most field delivery of TA is financed by external donors, and is subject to periodic evaluations. For RTACs, these evaluations are usually conducted midway through each RTAC's five-year cycle; an evaluation subcommittee made up of members from the Steering Committee (including Fund staff) determines the terms of reference and chooses the evaluator. The process is similar for the TTFs which support the Fund's TA on thematic areas, and for the country-specific trust funds. External evaluations are also done for some bilateral subaccounts (e.g. Japan, Switzerland).

8. **External evaluations of the RTACs and TTFs normally include the following common elements:** (i) assessments of specific TA and training activities; (ii) assessments of the RTAC's planning and execution and the center overall's operations; (iii) case studies including in-depth studies of individual TA projects; and (d) desk-based research on training workshops delivered or sponsored by the RTAC. They also typically include: (i) interviews with all relevant departments at IMF HQ; (ii) reviews of documentation (e.g. briefing papers and BTOs) (iii) detailed reviews of published and internal reports; (iv) field trips to the RTAC to discuss the Center's strategy, operations, and TA interventions; (v) surveys of TA recipients; and (vi)

[2] At the Board discussion of the 2013 Capacity Development Strategy paper, Directors asked for the next review of the strategy to take place in 2017. Subsequently, the Board agreed that reviews of the Fund's surveillance, lending, and CD activities should take place every five years.

online surveys of beneficiaries of the TA projects reviewed, participants in selected workshops, and Steering Committee members. Most recent evaluations use the OECD's Development Assistance Committee (DAC) evaluation criteria (Relevance, Effectiveness, Impact, Efficiency, and Sustainability).

9. **TA departments also conduct their own evaluations.** While there are occasional independent evaluations of departmental TA delivery, these evaluations are generally undertaken by the department delivering the TA with the goal of assessing the impact of TA advice and extracting lessons learned. For example, the Fiscal Affairs Department (FAD) undertakes regular inspection missions to intense TA users and distills policy lessons from topical evaluations, some of which conducted with external participation. These lessons feed back through FAD guidance notes, internal seminars, and workshops. The Monetary and Capital Markets Department (MCM) conducts ad-hoc ex post evaluations of multiyear projects through its TA evaluation program**.** It also conducts regular assessment visits to countries with intensive TA projects with a view to reporting to donors, and assessing how to adjust activities. MCM holds TA forums and seminars to share knowledge and lessons from these evaluations to help guide future TA work. The Statistics Department (STA) evaluates the most intensively delivered TA to member countries, averaging about one review per year since 2005 and incorporating RBM elements since 2011, to enhance the effectiveness of its CD activities and the Legal Department (LEG) also conducts regular evaluations.

10. **The IEO published in 2015 an assessment of self-evaluation at the IMF.** The IEO recently completed an evaluation of the IMF's systems and practices for self-assessment of its own work, including TA and found that considerable self-evaluation takes place and that many activities and reports are of high technical quality.[3] Executive Directors concurred on the importance of distilling and disseminating self-evaluation lessons in ways that highlight their relevance for staff work and facilitate learning. They saw scope in developing products and activities and revamping knowledge management practices aimed at better distilling and sharing lessons, as recommended by the report.[4] The Managing Director also supported this recommendation.[5]

B. Training Evaluations

11. **The Fund also undertakes regular evaluations of its training.** Regular internal and external evaluations of Fund training have been in place for some time. These evaluations follow the four-level Kirkpatrick (1976) model for training

[3] See *Self-Evaluation at the IMF: An IEO Assessment* (2015).
[4] See The Acting Chair's Summing Up - Independent Evaluation Office—Self Evaluation at the IMF—An IEO Assessment - Executive Board Meeting, September 18, 2015.
[5] See Statement by the Managing Director on the IEO Evaluation.

evaluations. The first level (Level 1) is reaction, which measures how well participants liked the training based on self-reported satisfaction. The second (Level 2) is learning, which measures knowledge acquisition. The third (Level 3) is behavior, which measures whether behavior changes following the course (e.g. whether the knowledge learned during the course is effectively applied). The final level (Level 4) is results, which measures the final results or outcomes occurring as a result of attendance and participation in the training. Currently, Fund training evaluations include:[6]

- End-of-course surveys of participants (Level 1);
- Pre- and post-course tests (Level 2);
- Follow-up surveys of participants and sponsors (Level 3) and;
- The triennial survey of sponsoring agencies (a combination of Levels 1, 3, and 4).

12. **End-of-course surveys are given to each participant in class on the final day of every course, with a response rate close to 100 percent.**[7] These surveys measure participants' perceptions of the training just received. Reactions and suggestions are collected through both a formal questionnaire and further probing during the closing session of the course. In ICD, this information is then transmitted to the department's senior managers in the back-to-office report, and eventually to ICD's Curriculum Development Committee (CDC). The TA departments that offer training under the ICD training program include FAD, the Finance Department (FIN), LEG, MCM, and STA. These and other departments also offer courses captured in the non-ICD training program data. Of the training offered separately from the ICD training program, MCM conducts end-of-course surveys but their results stay with the teachers. STA conducts surveys on most of its courses. LEG conducts training in the regional technical assistance centers and participant survey results are sent to ICD.

13. **Pre- and post-course tests to measure learning are currently administered by all ICD training divisions with varying coverage across courses, time and division.**[8] Tests are not yet standardized across training divisions (for example, by using a core set of common questions for the same course topic across

[6] Most of the training evaluations currently done are for ICD courses, though the changes in training evaluations described later in the paper are proposed to be applied to all Fund training, to the extent possible.

[7] All participants in online courses through both SPOCs (private courses for government officials) and MOOCs (massive online open courses for both officials and the general public) are asked to complete an end-of-course survey that is open for about a week, targeting a 50 percent response rate (relative to the number of active participants). The survey has some common questions with the one given for face-to-face courses.

[8] ICD's African Division started this practice, with a pilot program in 2010 to assess learning by administering a test at the beginning and at the end of selected courses. The pilot found that learning took place (i.e. the average test results after the course were better than the test results before the course and the differences between the before-and-after scores were statistically significant) in the courses that were sampled.

various offerings) but ICD's Curriculum Development Committee is discussing plans to do so. For the online courses, tests are conducted at the beginning and end of each course.

14. **Follow-up surveys were introduced in ICD (then INS) in 2006 in response to a donor request.** These are sent one year to eighteen months after a course to the participants and to the managers who sponsored their participation in the training. The survey asks whether the courses help participants do their job better and help them in their careers, and whether the knowledge gained in the course is used and shared with others. These surveys are conducted by an independent external firm to ensure anonymity of the responses. Coverage, even though expanded over the years, has remained selective, focusing on ICD courses. Since FY11, eight courses per year have been surveyed, with more attention being paid in recent years to ensuring a representative sample of courses by region, topic, and language of delivery.

15. **A triennial survey of training has been conducted in INS/ICD every three years since 1995.** The most recent survey was completed in 2015, covering the period 2012–14.[9] The survey, conducted by an external firm to maintain confidentiality, is sent to sponsoring government agencies with the objective of gathering their views on the effectiveness of the Fund's ICD training program and information about future training needs. The triennial survey does not target information about a specific course nor does it solicit the views of participants, setting it apart from the other two surveys. Rather, it seeks an overall evaluation from sponsoring agencies about the effectiveness of the training program and seeks indications of future demand for courses. The results of the triennial survey are summarized in a memorandum to management, and posted on the intranet.

III. STRENGTHS AND WEAKNESSES OF THE CURRENT SYSTEM

A. Technical Assistance

16. **Regular evaluation of TA activities is a key strength of current practices.** As mentioned in Section II, there is already quite a bit of TA evaluation, regardless of funding source. Over the past three years, an average of about seven evaluations per year have been prepared, the majority of which were done for externally-financed CD activities. Procedures also exist to allow lessons learned from evaluations to influence future TA delivery, e.g. staff responses and action plans to respond to mid-term evaluations for RTAC and TTF evaluations.

[9] The 2015 Triennial Survey concluded in November 2015.

17. **However, there is a lack of consistent and comparable methodologies used for evaluations.** A repeated comment by departments regarding current evaluations of technical assistance is that while a lot of reports are generated, the definition of terms, methodology and substantive focus vary so much that there is little scope for comparison of performance across different kinds of TA (and training). This is in part a by-product of the fact that many evaluations are conducted to respond to the legitimate interests and criteria of donors and partners, which may not always be consistent with each other. Some staff working on evaluations see the current system as too ad-hoc, and thus costly to implement with limited benefits. In wide-ranging discussions with TA, area, and other departments on developing a common evaluation framework, there were repeated calls for the Fund to have a single evaluation framework and methodology, and more standardized practices. Even though many of the external evaluations and some internal evaluations done by TA departments (e.g. STA) use the OECD DAC criteria, the definitions of the terms have varied, particularly for past external evaluations, and the quality of the evidence gathered has been disappointing in many instances.

18. **In the past there has been insufficient agreement and clarity about CD objectives, indicators, and milestones.** As noted in the 2013 Board paper, there have been no common benchmarks for success for Fund TA and this has inhibited systematic assessment. The discussions with departments pointed to the need for agreed indicators and a more standardized and rational process for collection and aggregation of that information. The IMF's Results Based Management (RBM) system is based on an agreed catalog of objectives, outcomes, and indicators for the Fund's main TA work streams. A key innovation in the IMF RBM system is the systematic articulation of baselines, as well as standardized objectives and outcomes (results).

19. **Good evaluation will require better information and data.** The minimum requirement includes baseline and ex-post information on the key indicators and control variables. Evaluations can also help answer whether the right TA was delivered through the right modalities and whether it was delivered at the right time. Some departments have suggested that it would be useful to keep records of whether a TA request was initiated by the authorities or IMF staff working on the country. The currently-used survey information, while helpful, is also seen as having a tendency toward upward bias in ratings. More use of face-to-face interviews might be preferable to bring out important issues and complement surveys. Interviews could be conducted with senior officials at 2-3 year intervals. Inputs from staff in area departments could also be helpful. "Flash" information, drawing on RBM data, was also suggested to facilitate fast identification of problems with TA, while time for remedies still existed.

20. **A significant part of the current evaluation effort is devoted to the evaluations of the Regional Technical Assistance Centers (RTACs).** About one-third of the evaluations done in the past three years have been for RTACs. Analysis of these evaluations reveals that although many ostensibly use the DAC criteria, the definitions of these criteria,

the indicators used to assess them, and other aspects of the evaluation vary so widely that there really is no common methodology being applied. Key weaknesses identified include:

- little standardization in terms of structure of the documents or the data presented and inconsistency in key evaluation questions across reports, hampering comparison;
- unnecessarily long evaluation reports;
- the objectives of the entity being evaluated are not stated clearly or not stated at all, especially ex ante;
- evaluations sometimes attempt to evaluate and also provide views on RTAC management and operational advice even when not relevant to TA delivery; and
- case studies tend to be costly and are often of dubious merit, particularly when taken individually.

21. **Inadequate use of counterfactuals.** A final critique is that any assessment of impact and efficiency requires information and/or judgment regarding the counterfactual – i.e. what would have happened if the CD had not been delivered, or it had been delivered in a different way, or at a different time. This is rarely incorporated in Fund TA evaluations. Step two of the four step process for the common evaluation framework provides scope for evaluators to share their assessment of what likely would have happened in the absence of the CD delivery. For example, when an evaluator determines that the IMF TA was unique and there was no alternative available, then this can be communicated in the evaluation. The counterfactual in this case would be no TA delivered. In other cases, an alternative organization may have offered similar TA.

B. Training

22. **Use of participant surveys is a key strength of current practices.** The current evaluation system for training uses participant surveys, administered at the end of all courses. In addition, lecturers and supervisors solicit feedback and clarification from students, and frequently alter the course in response. Ratings by participants influence the selection of course counselors and guest lecturers; lectures are sometimes dropped from the next offering of the course or significantly revised on the basis of lower participant ratings.

23. **However, the information gathered is inadequate for answering some questions.** The shortcomings of the current system are twofold. First, not much information is collected as to whether participants use tools taught when they are back on the job. Second, since most of the participants are not academic experts on the course topic, reliance on participant surveys as the prime source of information is inadequate for certain issues, for example, whether it was appropriate to teach a particular topic, whether evidence was presented fairly or comprehensively, or whether lectures were technically correct (i.e.

without errors or incomplete or distorted explanations).

24. **Current training evaluations lack the information to evaluate behavior change and results.** The information currently gathered in the participant surveys appears to be valuable mostly for measuring reaction, and may be less valuable for learning, behavior change, or results, although as noted in Section IIB some of the current evaluations do touch on these topics.

25. **Evaluations do not currently employ a control group or counterfactual.** Analysis using a control group could be useful to isolate the impact of IMF training.[10] Some course tests are administered before and after the course, which provides scope for a before/after comparison to assess how much learning took place during the course, and this is increasingly being done.

IV. GOING FORWARD: BUILDING ON CURRENT EVALUATIONS WHILE ESTABLISHING A CONSISTENT FRAMEWORK

26. **To strengthen the current framework, several new elements are outlined below.** They are grouped around the following topics:

- Producing shorter, more focused, and more comparable evaluations;
- Improving the information supporting evaluations;
- Maintaining the current level of resources used for evaluation but allocating these scarce evaluation resources more efficiently; and
- Using the information from evaluations to alter practices and/or shift priorities.

A. Producing Shorter, More Focused, and More Comparable Evaluations

27. **Specify clearly the objectives of all CD activities to be evaluated.** For TA, objectives and indicators of success should be consistent with those in the RBM catalog, which can be used for reference, as it lists objectives, expected outcomes, indicators and milestones for technical assistance. For training, ICD's new course curriculum includes objectives for each course and log frames for training are expected to be developed and included in the RBM catalog. Experience developing and implementing RBM will help to further develop the common evaluation framework and vice versa.

[10] For example, the control group could be constructed from the pool of applicants who have already been accepted to take a course but have not yet taken the course.

28. **Apply a four-step common framework (thought process) to all evaluations.** The four-step process is explained in Annex I and will be further articulated in the forthcoming companion note on methodology. The four steps are: (i) define the log frame or causal chain from inputs to outcomes; (ii) to the extent possible, indicate what is likely to have happened if the IMF did not deliver the CD (e.g. assess the counterfactual) to help assess impact;[11] (iii) assess outcomes using the OECD DAC evaluation criteria; and (iv) discuss why achievement of the DAC criteria were low/high, what factors explain it, and whether alternative interventions might have provided better results. Around this common core approach, the framework will allow flexibility to adapt evaluations to reflect the wide range of capacity development (CD) activities. Use of a common approach is intended to provide cross-activity comparability, permit aggregation, and enable an overall assessment of performance. Additional, satellite questions could be added as appropriate to enable evaluations to respond to requests by donors and to address issues that are specific to particular activities. In other words, while using a similar evaluation framework, different aspects of evaluations may be appropriate for different products and may be produced by different evaluators.

29. **Ensure clarity and consistency in definition and application of internationally-accepted OECD DAC evaluation criteria.** A key shortcoming of current evaluations is that while use of the DAC criteria may be specified in the terms of reference for the evaluation, different evaluators interpret the definitions of the criteria differently. The key to remedying this is to apply a common approach to defining the DAC criteria, how to assess their achievement, and what kinds of questions and indicators to use. Annex I provides definitions of the criteria and examples of typical evaluation questions for each criterion. The forthcoming companion note on evaluation methodology will provide greater detail on the application of the DAC criteria to TA and training activities.

30. **Use standardized Terms of Reference (ToR) templates (e.g. based on using the four-step process and including sample questions to assess each of the DAC criteria) to reduce workload and sharpen the focus of evaluations.** Evaluations themselves should have clear objectives. Standardized ToR templates along with the four-step process would help achieve this, especially in keeping the evaluations focused, and also assist with ensuring comparability of evaluations. Such templates will be included in the forthcoming companion note on evaluation methodology. Additional material desired by specific donors could be added to the ToR as required. Setting page limits (e.g. no more than 25-30 pages) to help ensure more focused evaluations may also be useful. ICD's SE division will review ToRs (not just for evaluations conducted by external evaluators but also for those

[11] According to the classic definition, the impact of a project is the difference in outcomes that occurred with the project compared to what would have occurred without the project. For example, simply looking at the revenue/GDP ratio before and after a country received fiscal TA would be insufficient to capture the impact of that TA. An evaluator would also need to look at what might have happened if the TA was not delivered (e.g. in an oil-exporting country, the revenue/GDP ratio could have fallen due to a drop in oil prices but the fall would have been even larger had the country not received the TA).

conducted internally) to ensure consistency.

B. Improving the Information Supporting Evaluations

31. **Consistent with the new results focus of all IMF CD, ensure that all CD activities have a data collection plan in place before the start of the CD activity.** Wherever possible, baseline data should be obtained before technical assistance and training begins. Through the IMF's RBM system, there will be agreement on the indicators, the data sources, and the responsibilities for data collection. For example, this is already being done for training (e.g. based on data from the tests administered prior to the start of training). For TA, this data could come in the form of the baseline indicators identified in the RBM system and clear ex ante identification of outcomes and indicators that are the focus of the CD intervention.

32. **Obtain ex-post information from TA providers and recipients and other stakeholders.** An "ex-post" survey to TA recipients would include questions about the perceived usefulness of the TA received and be administered immediately after the TA activity concludes. FAD already conducts surveys of this sort and a revised version should be extended to all TA. In order to gather information for the last step of the four-step common evaluation framework (i.e. evaluate why achievement of the DAC criteria was low/high, and whether alternative activities or modes of intervention would have achieved better results – see Annex 1), a questionnaire/survey should be developed to collect information from both the provider(s) and the recipient(s), as well as staff in area departments, on the outcomes of the TA provided within 12 months after the TA has concluded, to the extent that this information is not already being collected elsewhere (e.g. in CD-PORT). This information would then be stored, organized and used for future evaluations.

33. **Extend the pre- and post-course tests to all IMF training courses.**[12] Tests should be standardized for each course across offerings (e.g. the test for any given course should be the same regardless of where it's being offered). This standardization need not be 100 percent, but should include a standard core list of questions for each course that permit comparison over time, allowing scope also for revisions and to customize questions based on region-specific content. Randomized control trials (e.g. administering quizzes to groups of participants and non-participants to evaluate the impact of training on learning across different training modalities, for instance online vs. face-to-face) could also be considered.[13]

[12] It is not envisaged to administer tests for activities targeting high-level government officials and academia such as high-level policy dialogue forums/seminars.

[13] Another suggestion is to use quantitative techniques to explain differences in quiz results based on key participant attributes (e.g. agency worked at, prior online course taken, etc.) in order to inform the selection process for who attends training courses.

34. **Revise the end-of-course and follow-up survey questions for training to be more results-oriented:**[14]

- Revisions to the **end of course survey** will include new questions on the perceived usefulness of training for on-the-job performance (e.g. to what degree the training will influence the participants' ability later to perform his/her job, whether the training was job relevant and critical to success, whether the training contained new information, and whether the participant intends to use the skills and knowledge acquired during the training).

- The **follow-up surveys** will be revised to introduce a systematic and course-specific follow-up survey three to six months (rather than the current 12-18 months) after training is received. This shorter time horizon should help to facilitate follow-up with participants, which was rated as important by the agencies responding to the 2015 Triennial Survey of Training. New questions could include to what extent participants applied the knowledge/skills learned during the course, whether the participants used the tools/techniques taught at the course (for tool-focused courses and if primary functions of participants' work units involve use of these tools), and what factors were barriers or enablers to application of skills and knowledge acquired during the training.

35. **Maintain a channel for regular reviews of training course content.** For example, reviews could be done at the curriculum level, as was done recently by ICD which is in the process of restructuring its curriculum based on review by an independent expert.[15] Reviews could also be done at the course level, with independent experts attending lectures on an occasional basis to provide substantive feedback that course participants are not able to provide (primarily on substantive course content and flagging possible omissions). Such regular reviews would help to ensure that course curricula and content remains of high quality and relevant for the training needs of participants.

36. **Seek information from IMF country teams on the extent to which they observe tools learned in IMF training being applied at the country level.** To better assess country-specific training needs, a short annual questionnaire, will be introduced to gather the input of country teams and resident representatives. Questions could focus on the use of some tools taught at IMF training courses (e.g. Financial Programming and Policies (FPP), Debt Sustainability Analysis (DSA), External Balance Assessment (EBA), Early Warning Exercise (EWE), and some forecasting techniques, etc.) at the country level, and the level of expertise in using them.[16] Keeping track of this country-specific information

[14] The Triennial Survey of Training for 2015, which surveyed sponsoring agencies that sent participants to training under the ICD training program during 2012-14, was revised to introduce new questions on country-specific training needs and the perceived integration of TA and training activities.
[15] ICD has also introduced new topic networks to keep courses under review and up to date.
[16] Time to complete the questionnaire is expected to be no more than 15 minutes.

over time will provide a basis for assessing changes in trained skills and associated enablers or barriers to applying those skills.

C. Allocating Scarce Evaluation Resources More Efficiently

37. **Establish reasonable accountability standards, without evaluating everything.** A sufficient degree of accountability requires that there is a significant chance that a particular activity will be evaluated. The appropriate way to achieve this is a judgment to be made when developing the rolling three-year evaluation work plan (described below). In recent years, the IMF has been doing about 8-10 evaluations per year, the majority of which were external evaluations of donor-financed vehicles (e.g. RTACs, TTFs, and some bilateral activities). Rather than doing more evaluations, use of the new common evaluation framework will ensure that future evaluations will provide more comparable and useful information than past evaluations. Thus, implementing the common evaluation framework is expected to be budget neutral as the resources currently being spent are used more effectively. Putting more systematic RBM information into CD-PORT, with monitoring and evaluation in mind, could facilitate evaluations and lower their costs.

38. **Introduce a CD evaluation work plan with a rolling three-year horizon, revised annually.**[17] A rolling three-year evaluation work plan will be developed to include all evaluations either currently underway or planned. The evaluation work plan will include not only a list of planned evaluations, but also plans to collect better data and information for future evaluations, including through the implementation of the IMF's RBM system. Topics for evaluation could include not only specific projects but also broader issues such as whether risks identified in surveillance are being sufficiently addressed in TA work. The proposed work plan would be discussed and endorsed by the CCB at its July meeting, and updated annually.[18]

39. **Decide what and how to evaluate based on potential value of the information, cost, and achievement of accountability standards.** Several considerations should influence the decisions as to which evaluations to include in the work plan. The first is which evaluations must be done (e.g. RTAC/TTF/some bilateral subaccount evaluations, which are mandated by partners) and which additional evaluations would be beneficial to have. The work plan should achieve an acceptable degree of accountability with sufficient coverage of evaluations across the range of the Fund's CD activities. The second consideration is the potential for answering important questions with the data; there is little point in collecting data for an evaluation that is likely to be too inconclusive even with the most perfect data. A third consideration is the perceived importance of the question to be

[17] The three-year TA evaluation program in MCM is an example of this kind of work plan.
[18] In future years, the previous year's proposal for years 1 and 2 would be revised, with new proposals added for year 3.

addressed with the evaluation. The final consideration is cost.

40. **Maintain the flexibility to conduct mid-term or rapid evaluations.** It would be desirable to allow for quick mid-term evaluations or evaluations of topics that become urgent during the year as part of the evaluation work plan. For example, if it is suspected that a TA program is running into problems, a quick evaluation of causes and possible remedies could be conducted. Similarly, if the results from certain TA activities are very favorable, it would be worthwhile to conduct a quick evaluation to determine causes of success and any lessons learned which could be applied to current and future CD activities.

D. Using the Information from Evaluations to Alter Practices and/or Shift Resources

41. **Introduce quantitative scoring in evaluations to facilitate comparisons and aggregation**. To obtain an aggregate score for evaluations of activities with multiple objectives or outcomes, weights must be assigned (e.g. Objective or Outcome 1 gets a 30 percent weight; Objective or Outcome 2 a 70 percent weight)**.** For TA, such weights are expected to be included in the RBM system and these will be used in evaluations. For training activities, weights could be assigned and agreed as a prelude to any evaluation, such as in the evaluation ToR. The five criteria in the DAC framework will be scored on a 1-4 scale and averaged. For the overall score of the evaluation, a weighted average of these scores will be computed with the weights given to the objectives. This ensures that each evaluation will have a score attached to each objective or outcome and a summary score for the whole evaluation.

42. **Ensure evaluations are easily accessible upon completion and findings are broadly shared among Fund staff.** The IEO evaluation of TA in 2005 along with its 2014 update recommended that evaluations would be widened and disseminated more systematically, and aligned with knowledge management best practices aimed at better distilling and sharing lessons. To address this, ICD will produce a summary report of evaluation findings each year.[19] This report would summarize performance scores from several evaluations, as well as substantive lessons learned.[20] The report could be discussed at the July CCB meeting, where the outcomes for the previous fiscal year are discussed, and draw lessons for future delivery.[21] This report, along with the underlying evaluations would be posted on the CCB's intranet page. The evaluation summaries will also help the CD strategy by feeding into the regular review of CD activities done every five years, with the objective of improving future CD delivery. These activities would be part of the regular

[19] As noted in paragraph 37, recently there have been about 8-10 CD evaluations produced per year on average.

[20] A similar report had been done previously by OTM, with the last one produced in 2010.

[21] Responses to findings from evaluations can range widely, including recommending different kinds of interventions, different modes of interventions on a pilot basis, and terminating interventions.

work of the CCB and will not involve additional resources.

43. **The November CCB meeting would be expected to use evaluation results to assist with setting CD priorities.** CD priorities and activities for the coming three years are discussed annually at the November CCB meetings. The annual evaluation report, to be discussed during the July CCB meeting, can provide additional information to be used to help determine the medium-term CD priorities in November

Annex 1. The Four Steps in the Proposed Common Evaluation Framework

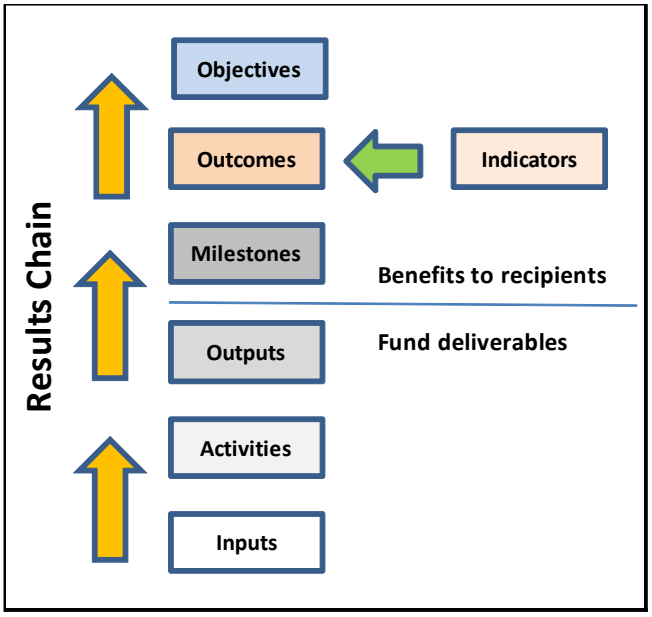
STEP 1. The Causal Chain: Defining the Logical Framework from Input to Outcomes

The logical framework is a device for organizing the causal relationships linking CD activities to desired outcomes and objectives (also called a log frame or the results chain). A typical example of a log frame describes how *inputs* (e.g. financial and human resources) are translated into *activities* (e.g. missions, backstopping, delivering a training course) in order to produce the *outputs* (e.g., a **T**A report). It then describes what *outcomes* (i.e. the actual capacity improvements) are expected along with any *milestones* (interim steps) that will be completed en route, in order to meet the *objectives* of a specific CD activity. The term *interventions* is sometimes used to conveniently summarize all of the *inputs* and the *activities* of a CD activity.

Multiple activities may be required to achieve an outcome, and exogenous variables may also affect the likelihood of achieving the outcome. Furthermore, activities range from being helpful for achieving an outcome to being necessary and/or sufficient for achieving an outcome. All of this should be clarified, for example, in the terms of reference for the evaluation and made available to the evaluator(s) prior to the evaluations. For completeness, it would be helpful to explicitly state the ultimate objective for an activity even if achievement of these objectives will not be evaluated in all evaluations.

The results based management (RBM) catalog in CD-PORT contains log frames and thus will normally be the most appropriate source for the log frames for TA and training.

Figure 1. An Example of a Results Chain for CD Activities



STEP 2. Describe the counterfactual

The impact of a specific CD activity should be assessed against a counterfactual, i.e. what would most likely have occurred without the intervention, either quantitatively (preferred, where feasible) or qualitatively. This step simply asks the evaluator to describe the counterfactual explicitly to the best of his/her ability based on information available. [1]

In the event of several "without-intervention" scenarios the evaluator should provide information about the likelihood of the major scenarios, since one must be selected for the impact assessment. Ideally, the evaluator would provide an informed assessment of the probabilities of such without-intervention scenarios while being candid about the degree to which he or she has confidence in such assessments. It is expected that this step will be based on a-priori reasoning, client interviews, and/or other evidence.

STEP 3. Assessing Outcomes Using the OECD DAC Criteria

The OECD DAC criteria, developed and endorsed by OECD members through the DAC, are a widely-accepted set of five criteria against which to assess public sector interventions. As such they constitute a convenient standard for assessing IMF CD activities, regardless of whether those activities were funded by IMF internal resources or using donor funds.

The five DAC criteria are relevance, effectiveness, impact, efficiency and sustainability. Before proceeding with definitions, it is important to explain how these criteria will be applied. The first step is to decompose the activity being evaluated into the interventions and the objectives. The art of evaluation consists partly in conceptualizing the interventions and objectives in a way that is broad enough to represent an important question but narrow enough to be precise. Any intervention with an objective can be evaluated. For training the intervention might be "teaching course X over a two week period" and one of the objectives, "to increase learning by the participant by at least X percent". For technical assistance an intervention might be "place a long term advisor in Ministry X" and the objective might be "to teach skills Y so that law X is adopted and implemented by date Z".

The second step is to pose the following sequence of questions (in the case of two objectives and one intervention):

Was objective 1 relevant?

Did intervention 1 achieve objective 1:

(a) Effectively?
(b) With impact?
(c) Efficiently?
(d) Sustainably?

Then this sequence of questions is repeated for every intervention and objective. Since there are two interventions in this example, the second set of questions would be as follows:

[1] Note that a counterfactual is not the same as a baseline, which is a special case of a counterfactual in which the pre-intervention variables are unchanged. In fact, a counterfactual could be a deterioration in the baseline circumstances.

Was objective 2 relevant?

Did intervention 1 achieve objective 2:

(a) Effectively?
(b) With impact?
(c) Efficiently?
(d) Sustainably?

When an overall performance rating is desired for an entity or a series of interventions with different objectives, weights will have to be assigned for achievement of those objectives. For technical assistance projects, such weights are expected to be included in the Results Based Management system. It is worth clarifying that this methodology, in which evaluations focus on the degree to which interventions achieved objectives, means that an evaluation of an entity such as a Regional Training Center (RTC) or a Regional Technical Assistance Center (RTAC) would really be an evaluation of the extent to which all interventions of that specific entity achieved their objectives. Information from individual evaluations of interventions could then be aggregated into an overall performance assessment of the entities providing these interventions.

Definition of the DAC criteria:

It is important to maintain a common understanding of what the DAC criteria mean and what kinds of questions are appropriate for each. The table lists each of the five DAC criteria, and provides definitions and examples of typical evaluation questions that might be asked under each category. These definitions of the DAC criteria should be specified in the Terms of Reference (ToR) to guide the evaluation.

DAC Criteria	Example Evaluation Questions
Relevance The extent to which CD activities (TA or training) served important objectives. Alternatively, an assessment of the importance of the objectives pursued.	• How high did the national authorities rank the objectives of the CD activity on their list of priorities (scale 1-10)? • Provide your own assessment of the importance of these objectives and support with evidence (if your assessment is that the CD activity was low priority, provide examples of higher-value alternatives). • To what extent were the objectives of the CD activity derived from capacity gaps identified by others (e.g. national authorities, country teams) or international standards? • To what extent did the objectives of the CD activity come from surveillance or program priorities for the country?
Effectiveness The extent to which CD activities attained their objectives. (This is not necessarily an assessment against a counterfactual.)	• To what extent were the objectives of the CD activity achieved or are likely to be achieved?
Impact	• List all changes that can be attributed to the CD activity, whether intended or not, compared to the counterfactual you believe would have been most likely.

Measures the positive and negative changes brought about by CD activity, compared to the counterfactual. The impacts can be direct or indirect, intended or unintended. (The relevant counterfactual is what *most likely* would have happened in the absence of the CD activity.)	• Provide quantitative estimates of these impacts, if possible.

Efficiency
 	
Measures the monetary value of the outcomes or benefits of CD activities compared to the monetary value of the inputs or costs incurred to achieve them. (An *impact* assessment is required in order to assess efficiency)
	• Provide estimates of the costs of the CD activity, to the fullest extent possible.
	• In light of what was concluded above under impacts, estimate the value of those impacts and compare them to the costs incurred, if possible.
	• Provide estimates of the costs of alternative ways of delivering the CD activity, if possible.
	• If no estimates can be provided for monetary value of impacts, assess the extent to which CD activity delivered is minimum cost, as assessed by
	  ○ Comparison of costs with other similar CD activity, or
	  ○ Examination of the process and implementation, including evidence of excessive staff turnover, unnecessary delays, inefficient organization etc.

Sustainability Measures the extent to which the outcomes or benefits achieved by the CD activity are likely to continue or last.	• For CD activities, assess the degree to which the transfer of knowledge is likely to be further disseminated (through CD recipients delivering CD to others) • Assess the extent to which the skills and knowledge gained will be retained and not forgotten. (Related question, extent to which skills will be used on job) • Assess the extent to which funding for CD will continue (note that there is no presumption that continued funding is necessarily desirable). • If the objective of the CD was to change behavior, assess the extent to which any achieved behavioral change will persist. • If the objective of the CD was to support new policies or laws, assess the extent to which these will persist.

Step 4. Results and Alternatives

This final step provides scope to examine two questions: (i) why achievement of the DAC criteria were low/high, what factors explain it; and (ii) whether alternative interventions would have provided better results.

For the first question, a list of explanatory factors might include:

- The intervention was based on a sound diagnosis of the key problem
- The mode of delivery was innovative
- Background conditions required for success were implemented simultaneously
- The original rationale for the intervention was poor
- Unforeseen exogenous factors changed and undermined the success of the project

The purpose of the second question on whether alternative interventions would have provided better results is to provide scope for the evaluator to share useful information or observations acquired during the evaluation. When suggesting alternative interventions that may achieve better results, the evaluator is expected to describe the reasoning and the supporting evidence.

When offering recommendations, the evaluator is expected to be mindful of the DAC questions as a group rather than a la carte. Suggested interventions that improve some of the DAC criteria, (achieving effectiveness for example) but fail on others (failing efficiency for example, due to high costs) are not viable recommendations.

Box 1. Theoretical Underpinnings of Training Evaluations

The logical framework for IMF training builds on Kirkpatrick's (1976) four-level model, which continues to be the industry standard for training evaluations. The most widely-used existing models take his basic approach of considering evaluation as steps (labeled as levels) of measuring reaction, learning, behavior, and results. Recent work has either expanded it or pointed out weaknesses, such as the need to develop more diagnostic measures.[1]

The four levels of the Kirkpatrick model essentially describe a chain of impact that guides the evaluation process as well as the *ex-ante* data collection plan. Generally, the value of information increases as progress is made through these levels, which are briefly described as follows:

- **Reaction** measures how well the participants liked a particular training program based on their feelings or satisfaction. Subsequent models added participants' planned action as well, i.e. a written plan for implementing what they have learned. A positive assessment at this level does not indicate that participants have learned new knowledge or skills.[2]

- **Learning** determines objectively the amount of learning that occurred, typically using formal tests before and after the training program. Where practical, a control group should be used and results should be analyzed statistically. Furthermore, the test should accurately and comprehensively cover the material presented. A positive assessment at this level does not guarantee that participants will apply what they have learned once they are back on the job.

- **Behavior** measures changes in on-the-job behavior resulting from the training. A systematic appraisal should be made of on-the-job performance on a before-and-after basis. Ideally such an appraisal should be made by a comprehensive group possibly comprised of participants themselves, their supervisors, their subordinates and peers, or other people thoroughly familiar with their performance. When feasible a statistical analysis using a control group to compare before-and-after performance and attribute changes to the training program would be the best practice. The post-training appraisal should be made three months or more after the training so that participants have an opportunity to practice what they have learned.

- **Results** are ultimate specific objectives of training programs and from an evaluation standpoint it would be best to evaluate training programs directly in terms of results desired. Key challenges are establishing causality and controlling for other factors, i.e. how much of the improvement is due to training as opposed to other factors. The literature acknowledges that for most training programs it is extremely difficult, if not impossible, to evaluate programs at this level.

[1] Salas and Cannon-Bowers (2001) present a comprehensive survey of literature. In 1980s, Phillips added a fifth level labeled as the return on investment (ROI), comparing monetary benefits from the program to program costs. Kraiger et al (1993) proposed a multi-dimensional view of learning, implying that learning refers to changes in cognitive, affective, and/or skill-based outcomes.

[2] Alliger et al (1997) noted that utility-type reaction measures (i.e. questions measuring the perceived utility value, or usefulness, of training for subsequent job performance) were more strongly related to learning and performance (transfer) than affective-type reaction measures. Surprisingly, they also found that utility-type reaction measures are more predictive of transfer than learning measures.