

Guidance Manual for Evaluators

Implementation Completion and Results Report Reviews for Development Policy Financing

Last Revision: February 22, 2023

Introduction

The Implementation Completion and Results Report (ICR) is one of the World Bank's main instruments for project- and operation-level self-evaluation. It is prepared by World Bank staff within six months of the close of every project funded by the International Development Association and the International Bank for Reconstruction and Development, or, in the case of a series of programmatic development policy operations (DPOs), within six months after closing of the final operation in the series.

The ICR Review (ICRR), conducted by the Independent Evaluation Group (IEG), is an independent, desk-based, critical validation of the evidence, results, and ratings of the ICR in relation to the project's design documents. It also assesses additional dimensions of the ICR to help promote staff learning. Based on the evidence provided in the ICR and an interview with the task team leader (TTL) at closing of the operation(s),¹ IEG validates the ICR findings and adjusts the ratings appropriately, based on the evaluation criteria agreed with Operations Policy and Country Services. IEG reviews all ICRs.

This manual provides guidance to evaluators preparing ICRRs on ICRs for development policy finance operations. It provides guidance for, and gives examples of, how to structure ICRRs with respect to content, presentation, and ratings. It also provides guidance on the preparation of ICRRs for development policy finance operations in fragility, conflict, and violence (FCV) situations to better reflect their particular characteristics and realities and make the ICRR a better tool for learning. Although this guidance manual does not focus on writing style, the ICRR should comply with IEG's writing style guidelines found in the [IEG Style Guide Fourth Edition](#).

¹ If the Implementation Completion and Results Report Review is for a programmatic series, there may be questions that can only be answered by previous task team leaders, who should then be interviewed.

Guidance Manual

Section 1: Information on Operation or Programmatic Series

Section 1 is filled in automatically by the system. Make sure your name appears as the evaluator. Note any missing fields.

Section 2: Objectives and Pillars or Policy Areas of Operation or Programmatic Series

2a. Objectives

Section 2a should *describe* the **project development objectives** (PDOs) of the operation or series.

Step 1: The formal PDO for the operation or series should be indicated in this section. The formal PDO is that which appears in the operation's Financing Agreement. If the PDO in the Financing Agreement differs in any way from that in the Program Document, the difference should be noted. If no formal PDO is stated in either the Financing Agreement or the Project Appraisal Document, this should be noted. In lieu of a formal PDO, the PDO identified in the ICR should be described. For a programmatic series, describe any changes or evolution in the PDOs across operations. There should be no assessment of the PDO in this section; it is purely descriptive.

Step 2: When necessary, it may be useful to "unpack" the PDO to arrive at the underlying de facto objectives.

Sometimes, the PDO consists of *several distinct objectives* (that is, it may contain subobjectives that either are loosely related or require policy actions in separate and distinct areas). If so, you should articulate the unpacked objectives for the purpose of the ICRR validation. It may be useful to review the prior actions (PAs) to inform the best articulation of unpacked objectives.

Example: The PDO "improve access to education and energy and foster financial inclusion," should be broken down into "improve access to education," "improve access to energy," and "foster financial inclusion."

Example: If the PDO is "promotion of fiscal consolidation," and the operation supports reforms on both the spending and revenue sides, the PDO should be unpacked into

revenue and expenditure components (for example, “control government spending” and “increase revenue mobilization”).

In some cases, you may find that the PDO is set at too high a level or has overly broad objectives (for example, “support inclusive growth”). In such a case, articulating a *credible* results chain linking the set of PAs to the associated PDO can be difficult. You may need to restate the PDO objectives as de facto objectives that better align with the scope and ambition of the PAs.

Step 3: After any unpacking, the ICRR text should state: “For the purpose of this ICRR, the objectives of the operation/series (against which outcomes will be assessed) are taken to be:” after this, the unpacked, de facto PDOs are listed (see box 1).

Box 1. An Example of Unpacking a Complex or Compound Project Development Objective

Project development objective: (i) strengthening the policy framework to support state effectiveness, private investment, and social inclusion; and (ii) improving the policy and institutional framework for public financial management.

For the purpose of this Implementation Completion and Results Report Review, the project development objectives of the operation/series (against which outcomes will be assessed) are taken to be:

- Strengthen the policy framework to support private investment
- Strengthen the policy framework to support social inclusion
- Improve the policy and institutional framework for public financial management

Source: Independent Evaluation Group.

Step 4: For section 3b, you will prepare a table that maps the full list of PAs associated with the operation(s) to the unpacked objectives from step 3.

2b. Pillars or Policy Areas

For the purposes of the ICRR, the terms *pillars* and *policy areas* have the same meaning and are used interchangeably. They refer to the area of reform required to support achievement of each objective. The text in section 2b is limited to describing the pillars of the operation as expressed in the Program Document.

2c. Comments on Program Cost, Financing, and Dates

This section describes the amount and source of financing of the operation or program (International Development Association grant, International Bank for Reconstruction and

Development loan, and so on), the approval date of the operation (or dates if a programmatic series), the date(s) it became effective, and the closing date. Specify the amount disbursed and explain any discrepancies between the amount approved and the amount disbursed. With development policy finance, since most operations are disbursed in a single tranche, differences are almost always due to exchange rate fluctuations between the approval and disbursement dates. If differences are large, you should seek additional information from the TTL during the standard ICRR interview. For a large movement in the exchange rate, the ICRR could note the movement between the approval and disbursement dates. This information can be found on the International Monetary Fund website at https://www.imf.org/external/np/fin/data/param_rms_mth.aspx.

Section 3: Relevance of Design

3a. Relevance of Objectives

Section 3a discusses the relevance of each objective (as unpacked and described in 2a).

The objectives of the operation (or series) are expected to contribute to country-specific development objectives and should reflect reform *priorities* as identified in diagnostic or analytical work.

The discussion of the relevance of objectives should address the following questions:

- Are the objectives relevant to tackling country-specific development constraints as identified in the Systematic Country Diagnostic or other relevant analytical work (for example, Financial Sector Assessment Program, Debt Management Performance Assessment, Public Expenditure and Financial Accountability, Public Investment Management Assessment)?
- Are the objectives relevant to the country's development strategy and the priorities set out in the Country Partnership Framework?
 - The discussion and assessment of relevance should go beyond simply noting that objectives are consistent or aligned with the Country Partnership Framework or the country's development plan. The text should assess the extent to which the objectives of the operation(s) would address *priority* country-specific challenges (for example, as identified in the Systematic Country Diagnostic or other diagnostic work). In effect, it should assess why the operation is a good use of scarce World Bank resources. An objective may be relevant if it responds to a significant shock or

development not foreseen when the Systematic Country Diagnostic or Country Partnership Framework was prepared.

- Are the objectives important enough to warrant direct World Bank involvement?
- Is the level at which the objectives are set appropriate, given the depth and scope of the reforms supported? (Generic objectives pitched at too high a level often lack specificity and extend well beyond the scope of the PAs.) If objectives are too high level and ambitious to be credibly achieved by the DPO's PAs, this should be noted.

Note: The ICRR does not evaluate the ambitiousness of the objectives. However, the ambition of objectives should be consistent with the scope and ambition of PAs. That is, it should be feasible for the reforms supported by the PAs to make a meaningful contribution to achievement of the objective(s), for example, by addressing important preconditions for reform progress. When PAs in support of a PDO are few and narrowly focused, the PDO should be similarly focused. For example, if PAs are limited to reforms in a single sector, a PDO that seeks "economic transformation of the economy" would be considered too broad or at too high a level.

For FCV countries, the discussion of the relevance of objectives may also cover the following points:

- The extent to which the objectives are realistic and achievable over the life of the operation or programmatic series given the FCV country context;
- The extent to which the objectives are consistent with the approach, strategies, and priorities identified in the Risk and Results Assessment or similar analysis. For example, in an FCV context, DPOs often have objectives that seek to strengthen a country's institutions or institutional capacity or build resilience. Where this is the case, it should be noted in the discussion of the relevance of objectives;
- Whether the focus of the operation or programmatic series is sufficiently narrow so as not to overtax the limited capacity of the country's institutions; and
- The extent to which the use of a DPO, rather than an investment project, is justified. For example, DPOs are seldom the best instrument for building technical capacity unless they are complementary to other efforts targeted at capacity building.

3b. Relevance of Prior Actions

Section 3b assesses the relevance of PAs in supporting achievement of the policy objectives (as unpacked in section 2a). The text should address the following questions:

- Does the PA (individually or in combination with other PAs) address constraints to achievement of the associated objective?
- Does the PA make a substantive and credible contribution to achieving that objective?

You should assess the credibility of the results chain that runs from each PA (or set of related PAs) to the relevant (unpacked) objective. Note that a PA may be relevant to more than one objective.

To facilitate understanding of the program’s design, PAs should be grouped by objective, and each PA should be listed as it appears in the Program Document(s) (that is, PAs should not be paraphrased). To help organize the discussion, each PA should be assigned a distinct number. Table 1 shows the recommended format for listing and numbering PAs.

Numbering is straightforward for a single-operation DPO. However, when the relevance of PAs for a programmatic series is being assessed, analysis can be facilitated by organizing PAs under each DPO, as in table 1. In this example of a programmatic series with two objectives, the first operation has four PAs and the second has three PAs. The PAs are numbered from 1 to 7 and listed in order, with PAs that are part of the same results chain next to each other.

Table 1. Numbering and Listing Prior Actions in a Programmatic Series: An Example from Mauritania

DPO 1	DPO 2
Objective 1: Improve Domestic Revenue Mobilization	
PA1: Minister of Finance has issued an order introducing the benchmark tax model for tax exemptions, and has published it in the official gazette, and has compiled a tax exemption registry for firms benefiting from tax exemptions under the 1982 Investment Code and the 1966 Free Zone Area law.	PA2: Ministry of Economy and Finance, based on a policy communique to the Council of Ministers, has notified the companies in full breach of their investment agreements that their tax and customs incentives, awarded under the 2012 Investment Code, will be revoked, effective January 1, 2018. PA3: The Ministry of Economy and Finance has adopted the legal provisions for a comprehensive transfer pricing documentation and disclosure requirements as well as an [sic] effective anti-abuse provisions, which limit an entity’s net interest deductions to a fixed percentage of its profit, measured using earnings before interest, taxes, depreciation and amortization (EBITDA).
Objective 2: Increase Efficiency of Public Spending	
PA4: The Council of Ministers has issued a decree creating an institutional framework for the evaluation, selection and execution of public investment projects, and has published it in the official gazette.	

DPO 1	DPO 2
PA5: The Council of Ministers has approved the budget law proposal for 2017 that includes an integrated public investment budget with combined domestic and foreign financed projects.	
PA6: The Minister of Economy and Finance has issued an executive circular requiring the expansion of the automated expenditure-chain system (RACHAD) to include all eligible EPAs in Nouakchott beginning January 1, 2017.	PA7: Minister of Economy and Finance has issued a policy communique instructing expansion of the treasury management system (RACHAD) to encompass revenues and expenditures of all eligible public agencies starting January 1, 2018, to reduce fiscal risks and enable budgetary savings.

Source: Independent Evaluation Group.

Note: DPO = development policy operation; EPA = administrative government agency; PA = prior action.

In assessing PA relevance, PAs are not expected to be sufficient in themselves to achieve objectives, but they are expected to move meaningfully along the results chain from the PA to the associated objective *in the specific country context*.

Assign a relevance rating for each PA based on a six-point scale, from highly unsatisfactory (1) to highly satisfactory (6; see table 2 and box 2). When PAs are clearly part of the same results chain (for example, complementary or subsequent steps in achieving the associated goal), you may assess them collectively. You should provide the following information to justify the assessment and the assignment of each rating, drawing on information contained in the Project Appraisal Document or ICR.

- **Results chain.** How the PA, in the country context (and considering known constraints), is expected to make meaningful progress toward the achievement of the relevant objective.²
- The rating for each PA should be noted in the paragraph in which its relevance is assessed (but numerical scores should not be included in the text). Where PAs are assessed together (that is, are part of the same results chain), the write-up can be consolidated into a single paragraph, but the distinct ratings for each PA should be articulated.

Ratings and justification should reflect the following points:

² For example, “By establishing detailed reporting on budget outcomes, PA1 is expected to support Uruguay’s implementation of a results-based budgeting framework to strengthen accountability and transparency in the budget process.”

- The clarity and credibility of the results chain linking the PA(s) to achievement of the relevant objective
- The extent to which the PA(s) is expected to
 - Address meaningful constraints to achievement of the objectives(s); and
 - Make a substantive and credible contribution to achieving the objective(s).
- The expected impact of a PA in making progress toward the achievement of the objective that is contingent on subsequent actions not contained in the programmatic series

The relevance of indicative triggers is not assessed.

Table 2. Assessing Relevance of a Prior Action or Set of Related Prior Actions

	Highly Satisfactory	Satisfactory	Moderately Satisfactory	Moderately Unsatisfactory	Unsatisfactory	Highly Unsatisfactory
Clarity and credibility of the results chain	There is an explicit, comprehensive, and convincing <i>results chain</i> linking the PA(s) to the achievement of the PDO, grounded in credible analytical work at the country level (and incorporating lessons learned from similar operations or experiences).		A credible <i>results chain</i> linking the PA(s) to achievement of the PDO is outlined but not explicitly described or grounded in credible analytical work.	Description of the <i>results chain</i> linking the PA(s) to achievement of the PDO is only partly convincing.	Description of the <i>results chain</i> linking the PA(s) to achievement of the PDO is unconvincing.	There is no reference to a <i>results chain</i> linking the PA(s) to achievement of the PDO.
Importance of PA to achievement of outcome	The PA(s) is the dominant factor in the achievement of the PDO.	The PA(s) makes a major <i>contribution</i> to the achievement of the relevant PDO.	The PA(s) makes a moderate <i>contribution</i> to the achievement of the relevant PDO.	The PA(s) makes a minor <i>contribution</i> to the achievement of the relevant PDO.		The PA(s) makes no discernible <i>contribution</i> to the achievement of any PDO.

Source: Independent Evaluation Group.

Note: PA = prior action; PDO = project development objective.

In an FCV context, the following should also inform the discussion and rating of the relevance of a PA (or set of related PAs):

- Is the PA consistent with the approach, strategies, and priorities identified in the Risk and Resilience Assessment or similar analysis? Does it show an awareness of

underlying fragility and conflict dynamics and the need to strengthen public institutions?

- Is the number of PAs (and policy areas) appropriate given the capacity and implementation constraints?

Determining the Overall Prior Action Relevance Ratings

To determine the overall relevance rating for PAs, first convert all PA scores to their numerical scores (see box 2). The default approach is to assign equal weight to each PA (that is, the overall relevance rating is the simple average of the individual PA relevance ratings). In some cases, one or more particular PAs may be considered more important than others. If so, you may use judgment to assign those PAs a higher weight, but the reweighting should be made explicit and a credible justification provided. Box 2 can be used again to convert that final score back to the HS–HU rating scale, with decimals rounded up or down as appropriate (see table 3).

Box 2. Numerical Scores for Prior Action Relevance Ratings

Highly satisfactory (HS) = 6

Satisfactory (S) = 5

Moderately satisfactory (MS) = 4

Moderately unsatisfactory (MU) = 3

Unsatisfactory (U) = 2

Highly unsatisfactory (HU) = 1

Source: Independent Evaluation Group.

Table 3. Ratings Methodology: Deriving the Overall Rating from Subratings

PA number	Rating on HS to HU scale	Rating on 6 point scale
1	S	5
2	MS	4
3	MU	3
4	HU	1
5	U	2
6	MU	3
7	U	2
8	S	5

Average	3.125
Converted back to HS to HU scale	MU

Source: Independent Evaluation Group.

Note: HS = highly satisfactory; HU = highly unsatisfactory; MS = moderately satisfactory; MU = moderately unsatisfactory; PA = prior action; S = satisfactory; U = unsatisfactory.

Section 4: Rating the Relevance of Results Indicators

Table 4 presents the criteria for judging the relevance of RIs

Table 4. Rating the Relevance of Results Indicators

	Highly Satisfactory	Satisfactory	Moderately Satisfactory	Moderately Unsatisfactory	Unsatisfactory	Highly Unsatisfactory
Likely impact of the PA in support of PDO(s)	The RI (alone or in conjunction with other RIs) fully and adequately measures the impact of the PA(s) on progress toward achievement of the targeted outcome, through reference to a clear and credible results chain.	The RI (alone or in conjunction with other RIs) is mostly adequate to measure the impact of the PA(s) on progress toward achievement of the targeted outcome, through reference to a clear and credible results chain.	The RI (alone or in conjunction with other RIs) is mostly adequate to measure the impact of the PA(s) on progress toward achievement of the targeted outcome, through reference to a clear and credible results chain.	The RI (alone or in conjunction with other RIs) partly measures the impact of the PA(s) on progress toward achievement of the targeted outcome, but its link to the PDO is unclear.	The RI (alone or in conjunction with other RIs) only peripherally measures the impact of the PA(s), is not clearly relevant to achievement of the PDO, or both.	The RI is not relevant to the impact of the PA toward to the achievement of the PDO.
Clarity of RI definition, data source, and data availability	(1) The definition and calculation of the RI is clearly explained in program documentation. (2) There are credible baseline data and a clear target; the sources of data to calculate the RI are clearly indicated. (3) The RI is used to <i>regularly monitor progress toward achievement of the target during implementation</i> of the programmatic series <i>and</i> at the time the ICR is produced.	(1) The definition and calculation of the RI are clearly explained in program documentation. (2) There are credible baseline data and a clear target; the sources of data to calculate the RI are clearly indicated. (3) Credible data are available to measure achievement of the target at the time the ICR is produced.	(1) The definition and calculation of the RI are explained in program documentation, but its calculation is unclear or not in appropriate units. (2) There are credible baseline data and a clear target; the sources of data to calculate the RI are clearly indicated. (3) Credible data are available to measure achievement of the target at the time the ICR is produced.	(1) The definition and calculation of the RI are not clearly explained in program documentation. (2) There are clear baseline data and a target, but sources for data to calculate the RI are vague. (3) The RI uses data that are either not credible or not available to assess achievement of the target at the time the ICR is produced.	(1) RIs are not defined in program documentation. (2) Data for either the baseline or target are missing, and data sources are not indicated. (3) The RI uses data that are not available to assess achievement of the target at the time the ICR is produced.	

Source: Independent Evaluation Group.

Note: the relevance of RIs is judged within the country context. In countries affected by fragility, conflict, and violence, the availability of regularly updated data for measuring progress may be limited, and you may need to augment the RIs with qualitative indicators. ICR = Implementation Completion and Results Report; PA = prior action; PDO = project development objective; RI = results indicator.

An RI that measures progress toward the objective but does not capture the impact of a PA is not considered relevant for the purposes of the assessment.

Example: In a case where the PDO objective was raising domestic tax revenues, the PA was an increase in the value-added tax rate and the RI measured the revenue to gross domestic product ratio, the RI would be considered moderately unsatisfactory since, although it captured the impact of that PA, it is also influenced by many other factors (for example, increases in other taxes, improved compliance). A better RI would be value-added tax collections.

Relevance also requires that each RI be clearly defined, including the associated data source and how the RI is calculated. Finally, RIs that capture the impact of PAs but are not connected to an objective through a coherent results chain are not considered relevant for the purposes of the assessment.

Example: In a case where the PA is increased funding for a program providing cash transfers to households conditional on children's school attendance, an RI measuring the increase in the number of beneficiaries of the cash transfer program would adequately capture one impact of the PA. However, if the relevant objective is to ensure better *funding and targeting* of programs for people living in poverty, the RI would not adequately capture the targeting element. Without another indicator capturing targeting, the relevance of the RI would be considered marginally unsatisfactory.

In an FCV context, institution building is critical. One or more RIs in this context would generally be expected to capture some aspect of this objective. The absence of indicators measuring progress to this objective (whether explicit or not) should be noted.

Required Table in Section 4

Section 4 of the ICRR should list the RIs as described in the Program Document. For ease of understanding the results chain (and for later assessing efficacy), group these by objective (as unpacked in section 2). Section 4 should include a table that contains information on both the relevance of RI and RI efficacy ratings (to be discussed in the following section; see table 5 for an example).³ The table should contain the following columns:

³ Results indicator baseline and target values (and associated dates) are included in the table, although that information is not discussed until the discussion of efficacy in section 5. The table should note the

- RI number and description
- PA(s) for which the RI is intended to capture impact
- Rating of RI relevance (see table 4 for guidance on rating RI relevance)
- The baseline and target values of the RI from the Program Document, including associated years
- Most recent data on RI (and date of observation)
- Assessment of actual change in RI relative to targeted change

Example: If the operation envisioned an increase in a particular RI from 40 to 100, the targeted increase is 60. If over the course of the operation, the RI increased to 70, the actual increase is 20. In the table you should note that only one-third of the planned change was achieved.

- RI achievement rating

In a programmatic series, list only the RIs and targets in place in at approval for the last operation of the series (RIs that are dropped should be excluded). An RI used in several operations but for which the RI target value changed should focus on the RI target for the last operation in the series. You should still make note (in the text) of RIs that were dropped or changed during the life of the series (this should also be noted in the section on Bank Performance—Implementation in discussing the adaptation of the series over time), but the assessment of relevance (and efficacy) should be based on only the final set of RIs and targets.

The criteria for assigning relevance ratings to RIs are described in table 5. These ratings and their justification are discussed in the text. The *overall* relevance rating for RIs is determined as for PAs, mapping individual ratings to numerical scores and then taking the unweighted average of the scores. This average is then mapped back to the associated rating after rounding up or down as appropriate. Record the overall relevance rating at the end of section 4.

status of the indicator at the target date in the last column. Often this information is contained in a table in the Implementation Completion and Results Report and can be directly imported, although the information may need to be reorganized.

Table 5. Sample Table on Results Indicators (Required)

RI Description [assigning a number to each RI]	Associated PA(s)	RI Relevance	Baseline [including units and date]	Target [including units and date]	Actual Value as of Target Date ^a	Actual Change in RI Relative to Targeted Change	Most Recent Value Available [if not target date]	RI Achievement Rating
Objective 1: Increase domestic revenue mobilization								
RI1: Tax revenue (percentage of GDP)	PA1	MS	17 (2015)	18.2 (2019)	Actual 18.8 (2019)	Over 100% of targeted change	19.0(2020)	High
RI2: Public enterprises' and agencies' extra-budgetary spending and carry-forwards (percentage of GDP)	PA2	HU	1.2 (2016)	0.2 (2018)	Actual 0.5 (2019)	70% of targeted change; (no data for superior indicator available)		Negligible
Objective 2: Increase private sector participation in nonextractives sector								
RI3: Executive PPP Unit has reviewed and assessed PPP projects according to new regulatory framework	PA4	S	0 (2016)	Half of PPP portfolio (2018)	Actual : 100% of proposed projects reviewed by PPP unit (2018)	Over 100% of targeted change	100% (2020)	High
RI4: Increase in the number of formal properties titled	PA5	S	27,168 (2015)	31,000 (2018)	Actual 29,275 (2018)	55% of targeted change	32,130	Modest

Source: Independent Evaluation Group.

Note: GDP = gross domestic product; HU = highly unsatisfactory; MS = moderately unsatisfactory; PA = prior action; PPP = public-private partnership; RI = results indicator.

a. For a programmatic series, if the RI has been dropped before the final approved operation in the series, use "Dropped" in place of "Actual."

Section 5: Achievement of Objectives (Efficacy)

Section 5 evaluates the extent to which the objectives of the operation or series have been achieved or are expected to be achieved in the near future. Efficacy is defined as the extent to which the objective has been achieved as a result of the PAs supported by the operation(s).

Begin by assessing achievement of the target for each RI.

Step 1. Assign an achievement rating to each RI using the four-point rating scale in table 6. The rating is based on the *change* in the RI relative to the *targeted change* (not relative to the RI's target value). If you determined in the RI relevance section that an RI *does not adequately capture* the impact of a PA, progress toward the associated objective, or both, or if data for the RI are not credible, you should adjust the achievement rating downward (unless other relevant evidence is produced). If data for the RI are not available, the RI targets should be considered not achieved (that is, negligible).

Example: Consider an objective to increase agricultural productivity in citrus fruits and corn, and a PA to give fertilizer vouchers to producers of these two products. The RI was "bushels of corn produced," with a targeted increase of 2 million bushels per year. The targeted change was achieved. However, the evaluator identified two shortcomings of the RI: (i) the RI focused only on the output side of production (whereas productivity has both an input and output dimension) and (ii) the RI only captured corn production. Since the RI did not adequately measure progress toward the productivity objective or capture the intended impact of the PA on citrus fruit production, the evaluator should downgrade the achievement rating unless additional information can more satisfactorily verify the intended PA impact toward the objective.

Table 6. Step 1: Assigning Achievement Ratings to Each Results Indicator

Rating	Description
High	The RI target was met or exceeded for the indicator. The assessment can be informed by additional evidence.
Substantial	At least two-thirds of the targeted change in the RI was realized by the target date. The assessment can be informed by additional evidence.
Modest	Less than two-thirds but more than 25 percent of the targeted change in the RI was realized by the target date. The assessment can be informed by additional evidence.
Negligible	No more than 25 percent of the targeted change in the RI was realized by the target date. When there is insufficient evidence to assess the achievement of the target and no credible additional evidence is presented, the target is considered not verified, which is equivalent to negligible.

Source: Independent Evaluation Group.

Note: RI = results indicator.

If the ICR or the TTL provides additional relevant evidence of progress toward achievement of a particular objective as a result of a PA,⁴ you may consider this in assessing achievement. You may choose to include additional evidence in the assessment, although you are under no obligation to expend significant effort in locating it.

Record these ratings in the final column of table 4 (Achievement Rating).

Step 2. Determine objective-level efficacy. Create a separate section for each objective. Under each objective, summarize the intended outcomes from the objective (the changes expected in the RIs, where RIs are relevant), noting results achieved relative to targeted results and highlighting where RIs were not appropriate for capturing progress. If other relevant evidence is available, describe it here. For each objective, look at the set of RI achievement ratings and compute the objective-level efficacy score using the rating methodology shown in table 7 (a six-point scale, from highly unsatisfactory to highly satisfactory).

Report the objective-level efficacy rating at the end of the section.

Table 7. Step 2: Rating Efficacy at Objective Level

Rating	Description
Highly satisfactory	Achievement of all RI targets is rated high.
Satisfactory	Achievement of most ^a RI targets is rated substantial or better; achievement of no RI target is rated negligible.
Moderately satisfactory	Achievement of at least half of RI targets is rated modest or better; fewer than one-third of RI targets are rated negligible.
Moderately unsatisfactory	Achievement of most ^a RI targets is rated modest or lower; at least one RI target is rated negligible.
Unsatisfactory	Achievement of most ^a RI targets is rated negligible; remainder rated no higher than modest.
Highly unsatisfactory	Achievement of all RI targets is rated negligible.

Source: Independent Evaluation Group.

Note: RI = results indicator.

a. *Most* is defined as two-thirds or greater. These rating definitions should cover the majority of situations. In the rare situation where the achievement of RI targets fits into more than one category, you should exercise judgment, taking into account the relevance of the RIs, existence of additional relevant evidence, and the extent to which there are gaps in the results framework measuring progress to the project development objectives as a result of the prior actions.

Step 3: The overall efficacy rating draws on the efficacy ratings for each objective. To calculate the overall efficacy rating, convert the efficacy scores for each objective to numbers using the mapping in box 2 (if scores were rounded up or down, please revert to the original scores up to two decimal places). Average the efficacy scores across objectives and map back to the

⁴ See Conducting the Task Team Leader Interview as Part of the ICRR Exercise later in this manual.

ratings (rounding up or down as appropriate). The overall efficacy rating is an unweighted average of the objective-level efficacy ratings.

Note: In an FCV context, flexibility may be needed in assessing efficacy, particularly for a situation of conflict. The level of uncertainty and volatility in the underlying context may make it unrealistic to expect all RI targets to be achieved. However, it may be difficult to anticipate, ex ante, which RI targets or pillars will be achieved. Moreover, the availability of credible and timely data may be limited. This may suggest the need for greater attention to qualitative data, lower-level outcomes, and proxies in assessing progress toward objectives.

Section 6: Outcome

The rating for overall outcome is determined using figure 1. The write-up should *briefly* summarize the findings on relevance of PAs and on efficacy. It should note the main strengths and shortcomings that contributed to those two ratings. For example, you could point out that the overall outcome rating was brought down by the low relevance of PAs.

Figure 1. Calculating the Overall Outcome Rating

		HS	S	MS	MU	U	HU
		Achievement of Objective (Efficacy)					
		HS	S	MS	MU	U	HU
Relevance of Prior Actions	HS	HS	S	MS	MU	U	HU
	S	HS	S	MS	MU	U	HU
	MS	S	S	MS	MU	U	HU
	MU	MU	MU	MU	MU	U	HU
	U	MU	MU	MU	U	U	HU
	HU	U	U	U	HU	HU	HU

Source: Independent Evaluation Group.

Note: HS = highly satisfactory; HU = highly unsatisfactory; MS = moderately satisfactory; MU = moderately unsatisfactory; S = satisfactory; U = unsatisfactory.

Section 7: Risk to Development Outcomes

The discussion of the risks to development outcomes should highlight the risks to *sustaining the development outcomes achieved*. It should *not* highlight the ex ante risks to the achievement of

the PDO as noted in the Program Document.⁵ Identify which outcomes are at risk of *not being sustained* and explain the nature of the risks that threatened their sustainability. Discuss developments or actions taken that could mitigate risks of policy reversal or erosion of progress achieved. If a subsequent supporting World Bank operation or International Monetary Fund program is in place, for example, discuss whether (and how) it supports the sustainability of the outcomes achieved.

Section 8: Assessment of Bank Performance

Bank performance is assessed for (i) the design and preparation of the operation or series (that is, up to approval of the operation or the first operation in series) and (ii) implementation of the operation or series (that is, post approval of the operation or the first operation in a programmatic series).

8a. Design and Preparation

Section 8a should cover the following points:

- The extent to which World Bank staff have drawn on lessons learned from prior experience in design of the operation or series. These lessons should be clearly identified and could be either from the country in question or from similar operations or activities in other countries.
- The adequacy of the analytical underpinnings of PAs and RIs (including their role in articulating the underlying results chain). For example, are the assumptions underpinning the theory of change based on sound and rigorous analysis that is relevant to the country context? Is the theory of change based on clearly identified diagnostic findings?
- The extent to which the Program Document identified the main risks and constraints to achieving PDOs and the quality and depth of the discussion of the main risks. The assessment should also include consideration of the credibility and coherence of the mitigating measures identified to reduce the risks. For example, where institutional capacity constraints in a government posed risks to implementation, was technical support from the World Bank or other development partners envisaged?

⁵ The assessment of the adequacy of the identification and discussion of the ex ante risks in the Program Document is covered in the Implementation Completion and Results Report Review section on Bank Performance: Design and Preparation (section 8a).

- The extent to which operation drew on consultations with relevant major stakeholders and development partners or envisaged collaboration, as necessary (for example, where other development partners were involved in similar support).

For FCV countries, the assessment should also cover the following factors:

- The extent to which *lessons learned* from prior experience in *FCV contexts* informed program design.
- The adequacy of *analytical underpinnings* of operation *in the specific FCV situation* in which the operation is being implemented, including with respect to the key drivers of fragility. This could include work done by both the World Bank and other development partners.
- The extent to which the operation identified possible negative impacts on drivers of fragility and conflict. For example, did evaluators draw on a Poverty and Social Impact Analysis of the reforms supported by the PAs to identify risks that could increase instability or violence?
- The extent to which the World Bank *proactively* supported efforts to mitigate or reduce risks identified *ex ante*. In FCV situations, weaknesses in technical and institutional capacity may pose particularly important risks to the ability of the authorities to implement supported reforms. Where this is the case, the World Bank should have had a strategy to address these shortcomings, through parallel technical assistance, training, or project support, provided directly or by development partners.
- The extent to which design of the operation drew on *consultations and cooperation* with major stakeholders and development partners (when necessary). In an FCV context, this may extend beyond traditional development partners (for example, United Nations agencies or humanitarian, diplomatic, and security actors may be critical partners).

8b. Bank Performance—Implementation

Implementation refers to the period following approval of the operation or the first operation in a programmatic series.

Consider the following questions:

- Is there evidence of ongoing monitoring of progress toward achievement of targets using the results framework (for example, aide-mémoire, notes to file)? This is particularly important for a programmatic series, in which progress toward RI targets

should be monitored regularly. To enable this, the selection of RIs should take into account the availability of data during the implementation of the series (not just at closing).

- In the case of a programmatic series, were triggers, targets, or RIs adapted appropriately to lessons learned or changes in underlying conditions, risks, operational priorities, or unexpected events, post approval?
- Were planned mitigating measures implemented (for example, technical capacity constraints, ownership concerns) to address risks to achievement of PDOs identified in the Program Document?
- Was there stakeholder and donor coordination where needed? In FCV situations, this might include (where appropriate) humanitarian, diplomatic, and security actors.
- Was there an effort to identify *new and emerging risks* to the achievement of the PDOs?

The ratings guidance for **Bank performance** is shown in table 8.

Table 8. Rating Bank Performance

	Highly Satisfactory	Satisfactory	Moderately Satisfactory	Moderately Unsatisfactory	Unsatisfactory	Highly Unsatisfactory
Prior experience and lessons learned	The design of the operation or series explicitly drew on prior experience and lessons learned.		The design of the operation or series referenced prior experience and lessons learned.	The design incorporated limited prior experience and analytical and diagnostic work, if relevant.		The design made no reference to the incorporation of prior experience or lessons learned.
Identification and mitigation of risks to achievement of PDOs	The operation contained a meaningful discussion of the major risks to achievement of PDOs, articulated credible mitigating measures, and incorporated them in the design of the operation.	The operation discussed some of the major risks to achievement of PDOs and articulated credible mitigating measures.	The operation discussed specific risks to achievement of PDOs, but only a subset of the mitigating measures were credible and substantive.	The operation contained a discussion of risks to achievement of PDOs at a general level, but key risks were missed. Mitigating measures were discussed but were largely superficial or not implemented.	The operation contained a superficial and incomplete discussion of risks to achievement of PDOs. Mitigating measures were not discussed.	There was no discussion of risks to achievement of PDOs or of mitigating measures.
Consultation with major stakeholders	The operation was informed by consultation with all major stakeholders.		The operation was informed by consultation with most major stakeholders.		The operation was informed by consultation with only some of the major stakeholders.	Few stakeholders were consulted in the design of the operation.
Coordination with development partners	There was close cooperation and coordination with all major development partners.	There was close cooperation and coordination with major partners.		There was limited cooperation and coordination with major development partners.	There was minimal cooperation or coordination with major development partners.	There was no cooperation or coordination with major development partners.

	Highly Satisfactory	Satisfactory	Moderately Satisfactory	Moderately Unsatisfactory	Unsatisfactory	Highly Unsatisfactory	
Monitoring	There is credible evidence (for example, reports, aide-mémoire) of regular monitoring of progress toward achievement of targets for all results indicators.	There is evidence (for example, reports, aide-mémoire) of periodic monitoring of progress toward achievement of targets or most results indicators.		There is evidence (for example, reports, aide-mémoire) of periodic monitoring of progress toward achievement of targets for a few results indicators.		There is no evidence of monitoring of progress toward targets for results indicators before series completion.	
Adaptation	Circumstances and priorities changed, and the series was adapted appropriately and explicitly to lessons learned.	Circumstances and priorities changed, and some elements of the series were adapted to lessons learned.		Changed circumstances or lessons learned resulted in modest adaptation of the series.	Changed circumstances or lessons learned resulted in insufficient adaptation of the series; the rationale for changes was not explained.	Changed circumstances or lessons learned resulted in minimal adaptation of the series, with little explanation for the changes.	Changed circumstances or lessons learned did not result in any meaningful adaptation of the series.

Source: Independent Evaluation Group.

Note: PDO = project development objective.

Section 9: Other Impacts

Frequently, operations will have significant impacts, both positive and negative, in addition to those explicitly identified in the Program Document. These include social, gender, poverty, climate, environmental, and conflict-related impacts. It is important that actual observed impacts be identified in the ICR.

Note this section is *not a description of expected impacts identified in the Program Document*, but a discussion of *actual* impacts. You should draw on the ICR to identify these other impacts, noting when evidence is absent or inconsistent. Where no such assessment appears in the ICR, note this in the ICRR. Failure to identify and discuss other impacts should negatively influence IEG’s rating of the quality of the ICR. This is particularly the case when social, gender, poverty, climate, and environmental impacts were *expected* (for example, they are identified in the Program Document) and are not discussed in the ICR.

For FCV countries, “other impacts” may include disproportionate impacts on aggrieved, excluded, or vulnerable groups; gender-based violence; and possible implications for fragility and conflict drivers. It is important to assess possible FCV risks that may be exacerbated by policy actions (for example, reforms to subsidies or tariffs).

Section 10: Quality of ICR

Because the ICRR is largely based on the information found in the ICR, the reliability of IEG’s ratings depends critically on the accuracy and quality of the evidence it provides. For this reason, IEG rates the quality of the ICR, taking into account the following criteria:

Internal consistency. Does the ICR present a coherent narrative of the program that flows logically?

- **Quality of evidence.** Does the ICR, including in annexes or appendixes, present an adequate and robust evidence base to support the achievements reported? Does the evidence come from credible sources, and is it appropriately referenced and presented in a concise fashion?
- **Quality of analysis.** Has there been sufficient and balanced interrogation of the evidence and clear linking of evidence to interventions and outcomes through a coherent results chain?
- **Quality of lessons learned.** Are the lessons formulated in the ICR supported by the evidence and findings of the ICR? Are they operationally relevant (that is, be drawn on

to concretely influence future behavior)? Are they focused on what can be derived from experience with the operation, or have they been overly generalized? In general, lessons based on evidence from a single country could not be extended to other countries or groups of countries.

- **Outcome orientation.** Is it clear how better results could have been achieved or what should be done differently in the future to improve impact?
- **Consistency with guidelines.** Has the report followed the ICR guidelines and methodology (for example, with regard to structure and ratings)?
- **Conciseness.** Does the ICR focus on critical information and evidence, or is it overly descriptive and contain information unnecessary for self-evaluation?

Section 11: Ratings

The ratings summary table lists and compares World Bank staff (ICR) and IEG (ICRR) ratings for outcome, Bank performance, relevance of results indicators, and quality of ICR (table 9). The IEG ratings are automatically generated from those entered in earlier sections of the ICRR. Wherever ICR and IEG ratings for Outcome or Bank Performance differ, you should briefly note the source of the difference.

Table 9. Example of a Ratings Summary Table

Ratings	ICR	IEG	Reason for Disagreement or Comments
Outcome	Satisfactory	Moderately satisfactory	Weak relation between some PAs and outcomes and some unclear results indicators reduced efficacy rating and hence rating for overall outcome
Bank performance	Satisfactory	Satisfactory	
Relevance of results indicators	—	Moderately unsatisfactory	
Quality of ICR	—	Substantial	

Source: Independent Evaluation Group.

Note: ICR = Implementation Completion and Results Report; IEG = Independent Evaluation Group; PA = prior action.

Section 12: Lessons

Each ICR presents lessons to inform future efforts. ICRs for programs that do not achieve their objectives often produce some of the most valuable lessons.

IEG, in the context of the ICRR, reviews the lessons articulated by staff and assesses them for clarity, coherence, and value added. You should identify the most pertinent lessons from the ICR and redraft them for clarity or to better reflect the finding of the ICRR. You should note where lessons do not appear well grounded in the evidence and analysis presented in the ICR.

You may also include lessons that emerge from the ICRR that are not identified in the ICR. These should meet the same standard of quality, specificity, and rigor as is expected in the ICR. Avoid identifying generic lessons.

Conducting the Task Team Leader Interview as Part of the ICRR Exercise

As part of the ICRR drafting exercise, you will conduct an interview with the last TTL of the operation. The purpose of the meeting is twofold: (i) to gain a better understanding of the project experience to improve the accuracy and quality of IEG's ICRRs and (ii) to ensure due process by providing the project TTL and the IEG ICR reviewer an opportunity to discuss the project experience. *The meeting is explicitly not intended to discuss any possible ICRR ratings.*

This meeting is conducted before IEG sends the draft ICRR to the Global Practice. The meeting with the TTL is different from the meeting that the Global Practice might request to discuss the draft ICRR after receiving it from IEG (please see point 4 below for further details on the timing of the meeting).

The meeting should be held with the last TTL of the project or, in the case of a programmatic series, the TTL of the final project. The meeting should not be held with the ICR author alone, unless the last TTL and the ICR author are the same person, or unless the last TTL specifically delegates to the ICR author responsibility for the meeting on behalf of the Global Practice. If the last TTL of the project has left the World Bank, on consultation with the ICRR coordinator, you should contact the concerned sector manager for an alternative suggestion. It would be up to the project TTL to invite other Global Practice staff at their discretion.

The meeting should be conducted only after you have prepared an advanced draft of the ICRR, and after the feedback on the first draft is received from the panel reviewer. You are expected to indicate in the relevant sections of the draft ICR that information will be sought to substantiate the assessment when submitting the draft to the panel reviewer, along with the list of questions that you intend to ask.

You should inform the meeting participant(s) that additional information obtained during the meeting and their comments may be used in the ICRR. You should focus on missing or ambiguous information in the ICR that is necessary to answer IEG's evaluative questions, including any additional evidence that may be needed to substantiate the ratings. For

example, often an ICR states that an RI target will be achieved by a specified date that is later than the ICR's publication date. In the TTL interview, you should ask for confirmation and evidence that the target was achieved. The ICR may have contradictory data in different sections. If so, the TTL interview is a chance to ask for the correct data. Finally, the ICR may mention that other development partners supported the reform agenda, without providing detail. The TTL interview is an opportunity to ask for details. You should use the meeting to confirm your understanding of the project context, gain a better understanding of the factors that might explain the project's (good or bad) performance, and probe what the project TTL might have done differently had they had the option.