IEG
Independent Evaluation Group

IFC International Finance Corporation
World Bank Group

Technical Note:

A Review of IEG's Methodology for Assigning Development Outcome Ratings

December 2008
Number 3

IEG analyzed the development outcomes of 672 projects evaluated since 1996, to determine their relationship with the ratings of the four underlying indicators: project business success; economic sustainability; environmental and social effects; and private sector development. Although there has been no formal mechanism for deriving development outcomes directly from the four indicator ratings, IEG found a strong, logical relationship between them. In practice, therefore, the combination of guidance and peer reviewing has resulted in a robust and consistent approach to assigning development outcomes. Nevertheless, IEG believes that improved guidance would make the process more transparent and avoid the possibility of rating anomolies occuring in the future. IEG's recommendation is to introduce a system whereby the underlying indicator ratings are averaged and the most appropriate development outcome rating is selected from a lookup table.

About IEG-IFC

IFC's Independent Evaluation Group (IEG-IFC) independently evaluates IFC's investment and advisory services operations and reports its findings to IFC's management and Board of Directors.

About Technical Notes

Technical Notes present the findings and recommendations of IEG's reviews of evaluation processes and methodologies. The views expressed here are those of IEG-IFC and should not be attributed to IFC. The findings here do not support any general inferences beyond the scope of evaluation, including any references about IFC's past, current or prospective overall performance.

Online Access

http://www.ifc.org/ieg

Introduction

As part of a wider review of IEG-IFC's evaluation methodology, this paper reflects on the approach used by IEG to assign project development outcome ratings as part of its independent validation of Expanded Project Supervision Reports (XPSRs). It analyzes eleven years of evaluation results to determine the relationship between development outcomes and the ratings of its four underlying indicators, and how consistently guidance has been applied in the past. Finally, it addresses the question as to whether a more formulaic approach is merited going forward, i.e., whether development outcomes should be derived automatically from the ratings for the four underlying indicators, or whether this should be left to the judgment of the evaluator.

Current Practice

In an XPSR, a project's development outcome is measured across four indicators: project business performance; economic sustainability; environmental and social effects; and contribution to private sector development. Each of these measures a distinct aspect of the operation's performance in fulfillment of IFC's Article 1 purpose and mission. The development outcome rating is a bottom-line assessment of a project's results on-the-ground, and not an average of these four indicators. Taking into account the four indicators, a project's overall impact on the development of its host country is rated on a six-point scale based on the following evaluation standards:

Highly Successful: A project with overwhelming positive development impacts, with virtually no flaws.

Successful: A project without material shortcomings, or some very strong positive aspects that more than compensate for shortfalls.

Mostly Successful: A project which may have some shortcomings, but with a clear preponderance of positive aspects.

Mostly Unsuccessful: A project with either minor shortcomings across the board, or some egregious shortcoming in one area which outweighs other generally positive aspects. **Unsuccessful:** A project with largely negative aspects, clearly outweighing positive aspects.

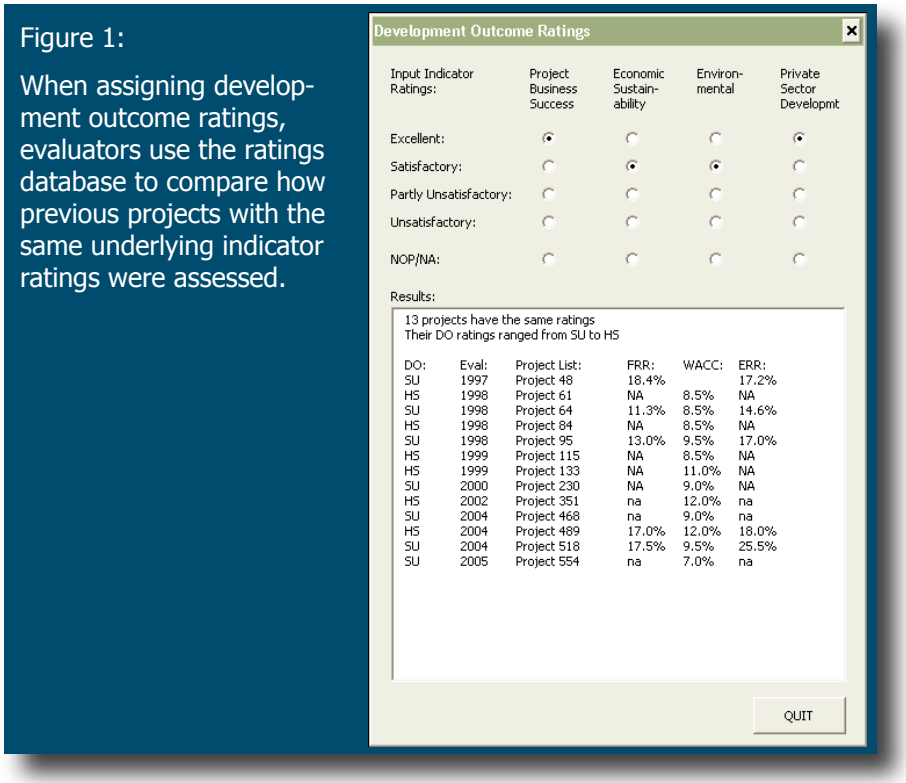
Highly Unsuccessful: A project with material negative development aspects with no material redeeming positive aspects to make up for them.

As an additional frame of reference, for any rating of mostly successful or better, IFC should be able to explain convincingly (without embarrassment) to a public audience why it rates a project a success. The guiding principle should be: if all of IFC's projects were mostly successful, it should just be able to justify its existence as development institution.

Under the guidance, therefore, the evaluator is at liberty to assign a development outcome rating that best reflects the balance of positive and / or negative attributes as measured by the four underlying indicators. There is no predefined weighting system: depending on the nature of the project (location, sector, size, objectives etc.), the evaluator may attribute more relevance to one or other indicator in determining overall development outcome. For example, a project that is commercially successful but which has egregious environmental impacts may (and indeed should) be considered less than successful in terms of its overall development outcome regardless of the "average" rating implied by the four underlying indicators.

Figure 1:

When assigning development outcome ratings, evaluators use the ratings database to compare how previous projects with the same underlying indicator ratings were assessed.

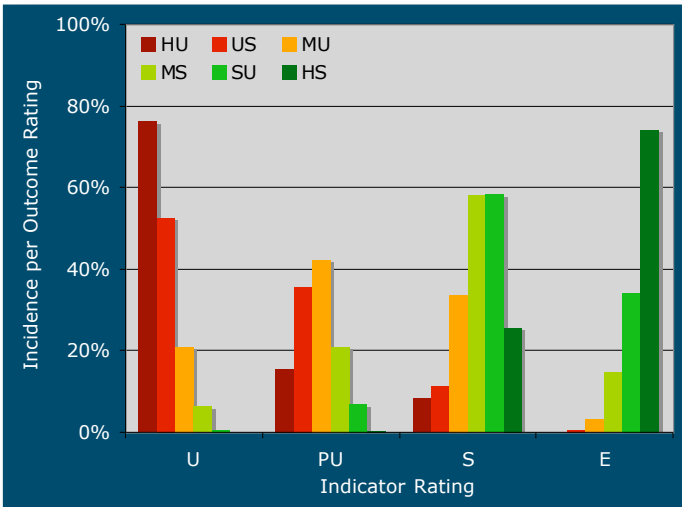


The lack of a weighting system or formula for rating development outcome gives rise to the risk of inconsistency between the approach of different evaluators. IEG controls for this risk two ways. Firstly, evaluators can refer to the database of XPSR ratings to see how previously evaluated projects with the same combination of indictor ratings were rated for overall development outcome (see Figure 1). The database provides important context such as the project's financial and economic rates of return (for real sector projects). Secondly, the Head of Micro reviews each XPSR and Evaluative Note to check for inter-rater consistency and correct application of the guidance by evaluators in assigning ratings.

Results Based on Current Methodology

IEG has analyzed the pattern of development outcome ratings and their relation to underlying indicator ratings for 672 projects in the XPSR database evaluated between 1996 and 2007. Figure 2 shows for each development outcome rating, the incidence of ratings of unsatisfactory (U), partly unsatisfactory (PU), satisfactory (S), and excellent (E) among the four underlying indicators.

It shows: (i) the clear preponderance of U and PU indicator ratings for projects with highly unsuccessful (HU) or unsuccessful (US) development outcomes; (ii) at the opposite end of the scale, the high incidence of S or E indicator ratings for projects with highly successful (HS) or successful (SU) development outcomes; and (iii) the greater dispersal of indicator ratings for projects with mostly unsuccessful (MU) or mostly successful (MS) development outcomes, albeit with their distributions skewed to the left and right respectively as one would expect.



Development Outcome	Indicator Ratings				
	U	PU	S	E	Total
HS	0 0%	1 0%	70 25%	204 74%	275 100%
SU	4 1%	54 7%	451 58%	263 34%	772 100%
MS	35 6%	118 21%	327 58%	83 15%	563 100%
MU	98 21%	198 42%	158 34%	15 3%	469 100%
US	210 53%	143 36%	45 11%	2 1%	400 100%
HU	108 76%	22 15%	12 8%	0 0%	142 100%

Figure 2:

Showing the distribution of indicator ratings for each development outcome rating. High development outcomes mostly comprise indicator ratings of S or E, whereas low development outcomes have a preponderance of PU and U indicator ratings.

Extending this analysis, IEG assigned numeric scores[1] to each indicator rating as follows: U=0; PU=1; S=2; E=3. Then, for each project, IEG calculated the average score (s_{avg}) of the underlying indicator scores. Figure 3 summarizes the results by grouping projects according to their development outcome ratings. There is a clear delineation between different development outcome ratings based on the average underlying indicator scores s_{avg}. This suggests that even though the approach currently used by IEG to arrive at development outcome ratings is judgmental as opposed to formulaic, it nevertheless results in an aggregate picture which is both intuitive and remarkably undistorted.

Pursuing this analysis further, IEG looked at the distributions of s_{avg} to check that the aggregated results displayed in Figure 3 were not obscuring wide variations in the range of s_{avg} for each development outcome rating. Figure 4 shows the distribution of underlying s_{avg} among projects, categorized according to their development outcome rating (the means of each distribution are equivalent to the values shown in Figure 3). What can be noted about the distributions in Figure 4 is: (i) they are single-peaked and resemble standard distributions (with the exception of those for HU and HS development outcome ratings, which are curtailed due to the boundaries imposed by the scoring sys-

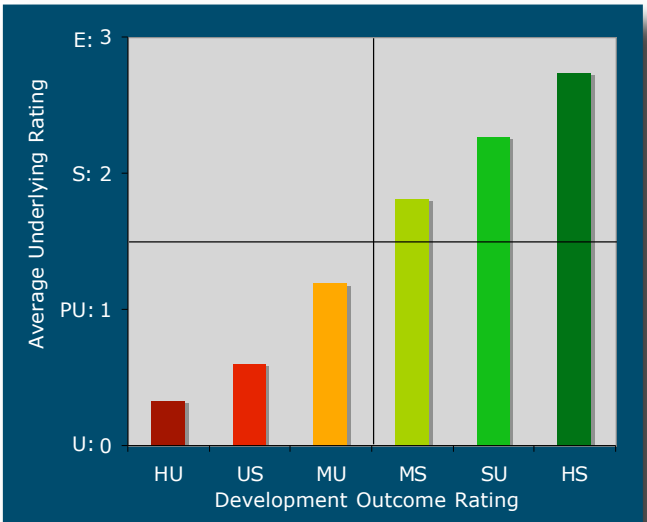


Figure 3:

Showing the average indicator rating s_{avg} for each development outcome rating, based on a score of 0 to 3 for indicator ratings of U to E respectively. The chart shows the clear delineation between outcome ratings based on their s_{avg}.

[1] Note that such a scoring system assumes that indicator ratings are distinctly granular and exhibit a linear relationship in terms of the "quality of impact" from a rating of U through to E. In reality, neither assumption is correct: the indicator ratings disguise a continuum of "quality of impact", and the relationship is unlikely to be linear, i.e., the difference in "quality of impact" between an E rating and a S rating (or between an U rating and a PU rating) may be much greater than between a PU and a S rating.

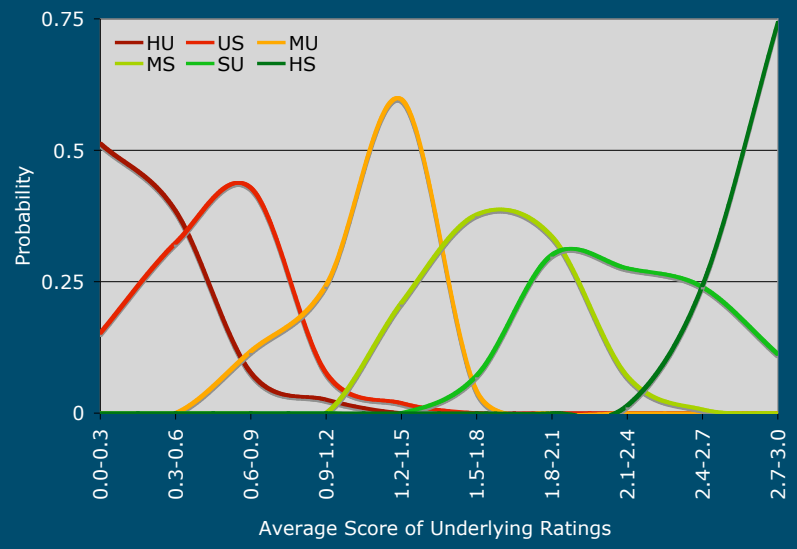


Figure 4:

Shows the distribution of underlying s_{avg} among projects, categorized according to their development outcome rating (the means of each distribution are equivalent to the values shown in Figure 3).

The distributions are distinct with no overlap in the case of HU and HS outcome ratings or US and SU ratings. There is some overlap between the middle outcome ratings of MU and MS.

tem); (ii) there is no overlap between the distributions for HU and HS development outcome ratings; (iii) there is virtually no overlap between the distributions for US and SU development outcome ratings; (iv) there is overlap between distributions for MU and MS development outcome ratings.

Clearly then, IEG's assignment of development outcome ratings appears to reflect the underlying indicator ratings in a consistent and logically defensible way. There is clear differentiation between projects that are rated HU or US from those rated SU or HS. For projects that fall into the middle ground where judgment is more difficult, the difference between MU and MS ratings is less marked – the s_{avg} distributions overlap – however, they are still distinct if not mutually exclusive.

For reporting purposes, most of IEG's analysis is based on a binary simplification of indictor and outcome ratings, i.e., high ratings or low ratings. In the case of development outcomes, high ratings are mostly successful or better, whereas low ratings are mostly unsuccessful or worse. IEG analyzed the incidence of each underlying indicator rating for projects with either high or low rated development outcomes. Again, projects with high development outcome ratings tend to have underlying indicator ratings of S or E (an 87 percent incidence), whereas projects with low development outcome ratings have a greater proportion of PU or U ratings among their underlying indicators (a 77 percent incidence).

In terms of project s_{avg} scores, IEG found that all projects with an s_{avg} below 1.2 had low development outcomes, and all projects with an s_{avg} above 1.8 had high development outcomes. Only within the s_{avg} range of 1.2 to 1.8 were binary development outcomes mixed: for 71 percent of those projects with s_{avg} of 1.2-1.5 the development outcome was low; for 93 percent of those projects with s_{avg} of 1.5-1.8 the development outcome was high.

Merits of a Rating System Based on Weighted Scores

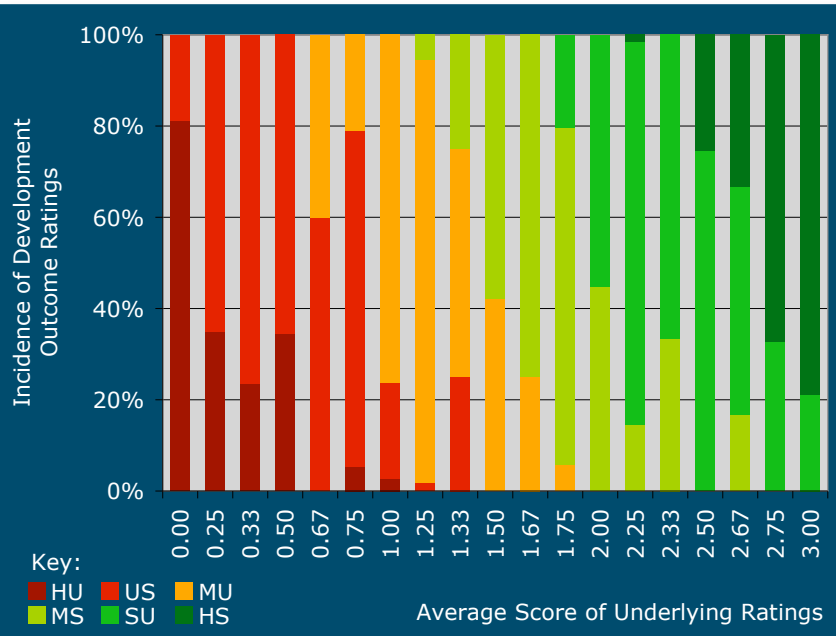
IEG has considered whether there is merit in changing its approach to assigning development outcomes to a more formulaic method based on a scoring and weighting of the underlying indicators. An overall score would then be calculated and compared against predefined benchmarks. On the one hand, this would remove the possibility of inter-rater inconsistency in determining development outcomes on the current judgmental basis, and therefore could appear more robust from a methodological standpoint. It would also eliminate any debate between IEG and IFC over the appropriate outcome rating, a subject in which departments have an increasing interest since development outcome ratings now feed into scorecards.

However, IEG believes the disadvantages outweigh these advantages. Firstly, such a rigid system does not recognize the spectrum of "quality of impact" within each indicator rating, but instead imposes an artificial granularity. Under the current approach, IEG can take into account the difference between a "just satisfactory" and a "not quite excellent" indicator rating in its assessment of overall development outcome. A scoring system would remove this flexibility unless it was based on a scale with significantly more than four categories. While that might be possible, it would necessitate defining additional benchmarks, probably a subjective process in its own right given the underlying impacts are largely qualitative in nature.

Secondly, it introduces the possibility of gaming the system. For example, a department may choose to emphasize certain aspects of the project to ensure underlying indicator ratings are sufficient to yield a positive overall development outcome. IEG's hands would be tied. Of course, departments may already choose to overstate the positive and understate the negative, but the flexibility of the current system encourages them to be candid and present a balanced rationale for the overall development outcome.

Thirdly, IFC's portfolio of private sector projects is not a homogenous population. The range of different sectors, countries, project types, economic circumstances, sponsors, instruments etc. mean that no two projects are identical. Attributing the appropriate weight to development indicators is therefore an intuitive process reflecting the nature of the project on a case by case basis. For example, environmental and social impacts hold far greater significance in judging the development outcome of a mining operation than they would for an insurance company. For a rating agency, its profitability and direct economic contribution are relatively minor considerations compared to its potential impact on private sector development. In contrast, financial and economic returns are critical in judging to what extent an import substitution project is benefiting from protection. It is unlikely, therefore, that a single scoring or weighting system would be appropriate for all projects; instead it would need to be tailored for many different types of IFC intervention.

Lastly, IEG would need to consider whether such a system was compatible with the Evaluation Cooperation Group's Good Practice Standards for the evaluation of private sector projects. These emphasize the use of "summary qualitative performance judgments based on the underlying indicator ratings" in assigning outcome ratings, rather than using a simple average.[1]



s_{avg}	Guideline Development Outcomes
0.00	**HU**, US
0.25	**HU**, US
0.33	HU, US
0.50	HU, **US**
0.67	**US**, MU
0.75	US, MU
1.00	US, MU
1.25	US, **MU**
1.33	**MU**, MS
1.50	MU, MS
1.67	MU, **MS**
1.75	**MS**, SU
2.00	MS, SU
2.25	MS, SU
2.33	MS, **SU**
2.50	**SU**, HS
2.67	SU, HS
2.75	SU, **HS**
3.0	SU, **HS**

Figure 5:

The chart shows how projects with different s_{avg} were rated for overall development outcome, based on the 672 projects evaluated since 1996. From this historic data, IEG has constructed a lookup table to indicate either of two possible development outcome ratings for each s_{avg} score. Where one is shown in bold, this indicates a preference for that particular rating over the other. Had such a system been employed in the past, only 2 percent of outcome ratings would differed from those actually assigned (1 percent on a binary basis).

[1] See GPS third edition performance standard 4.2.4.

A Hybrid Rating System

Given these limitations, IEG considered how, if at all, its current approach could be improved. One option it considered is to use a hybrid of its current methodology and a rigid scoring system. Under such a hybrid, IEG would use a simple scoring system to indicate not one but a range of possible development outcome ratings. This has the advantage of providing some direction to the evaluator on the appropriate development outcome, but leaves them sufficient flexibility to attribute their own weightings to the underlying indicators given their importance to the project in question. It would also reduce the incidence of rating anomalies (i.e., high development outcomes for projects with very low s_{avg}, or visa versa, low development outcomes for projects with very high s_{avg}).

The table in Figure 5 shows how such a system might work. Based on the scoring system of 0, 1, 2 and 3 for indicator ratings of U, PU, S and E respectively, the evaluator would calculate s_{avg} and use the lookup table to determine the guideline development outcome rating. For each s_{avg} there are two possible outcome ratings; where they are shown in bold this indicates a preference for one rating over the other. IEG has developed these benchmarks based on historic rating patterns. Consequently, if such a system was used to re-rate past projects, only 14 of 672 development outcome ratings (2 percent) would differ from those actually assigned; and on a binary basis only 7 (1 percent) would differ with negligible net effect on the overall success rate (since these 7 would include 3 upgrades and 4 downgrades).

Conclusions

From its analysis of the development outcome ratings for 672 projects, IEG has made the following observations:

(i) Based on IEG's current methodology, where evaluators are at liberty to assign development outcome ratings based on their own qualitative assessment of the four underlying indicators, there has been a strong and consistent relationship between outcome and indicator ratings with very few incidences of clear rating outliers.

(ii) In practice, therefore, the combination of guidance and peer reviewing has resulted in a robust and consistent approach to assigning development outcomes.

(iii) A more formalized approach, whereby development outcomes are determined automatically based on a scoring and weighting of the underlying indicators, would be too constraining and would likely fail to differentiate the many and varied types of project supported by IFC.

(iv) There is, however, a case for using a hybrid approach whereby a simple scoring system is used to indicate a range of possible development outcome ratings. This would increase the transparency of IEG's methodology and eliminate the possibility of obvious rating anomalies going forward.

Resources

Lead Author:
Nicholas Burke

Director, IEG-IFC:
Marvin Taylor-Dormond

Head, Knowledge, Communications & Quality, IEG-IFC:
Sid Edelmann

IEG Help Desk:
(202) 458-2299
askIEG@ifc.org