i-eval THINK Piece, No. 3

Rating Systems in International Evaluation

Kanika Arora

William Trochim

International Labour Office
Evaluation Unit

January 2013

Rating Systems in International Evaluation

Kanika Arora
Maxwell School, Syracuse University

William Trochim
Cornell University

Prepared for the Evaluation Unit of the International Labour Organization

CONTENTS

PURPOSE OF STUDY

The late 1990s marked the onset of the "results revolution" in international aid evaluation. Broadly characterized as "Results Based Management (RBM)," this approach emphasized the establishment of performance management systems at higher, more strategic, organization levels, such as the country program and the agency. Rating practices have evolved as a key component of reviewing and reporting performance data in this specific context.

There are several reasons why ratings systems have become so popular. First, they are relatively easy to use. Giving a rating to some object is one of the most common measurement approaches in contemporary societies. Once a set of criteria has been developed, it is relatively simple for organization professionals or their consultants to rate a program on one or more dimensions of interest. Second, rating systems are easy to present to key stakeholder audiences. A simple rating system can summarize a broad range of key features of a program in just a few numbers. Third, rating systems are understandable by key audiences of interest. Most professionals these days have had considerable experience with the types of quantitative summaries that rating systems produce. Fourth, because they are quantitative in nature, ratings can be easily aggregated, making it possible to summarize across a portfolio of related programs. Finally, the numerical nature of rating systems gives the appearance of precision and suggests that there is an empirical or scientific basis for the results.

However, rating systems also pose considerable challenges. The quantitative and presumed scientific basis of ratings and their apparent ease of use, presentation and comprehensibility can mask a range of methodological issues. We cannot assume that just because a quantitative value is used that the value is either consistently obtained (reliability) or accurate (validity). In order for a rating system to work well it needs to be well-tested and carefully applied.

The primary purpose of this document is to provide a general overview of key design and implementation elements that influence the validity and reliability of the ILO's rating practices.

QUALITY OF RATINGS

There are a number of critical issues involved in the construction and implementation of a rating system, any of which can have a determinative effect on the accuracy and consistency of the ratings that are ultimately produced. To contextualize the analysis of these issues, it is important to begin by laying out a more fundamental question – what are important elements of a high-quality rating system and how can they be measured? In our analysis, the two key attributes that contribute to the quality of a rating system are "validity" and "reliability." We discuss these in turn.

Validity

The idea of the validity of ratings generated through rating systems is taken from the idea of construct validity in measurement. The key issue is the degree to which it is reasonable to conclude that a rating reflects the criterion that it is intended to measure. For instance, if a satisfactory rating is given for a project with respect to a specific criterion, we would say that the rating is "valid" (i.e., has construct validity) if there is independent evidence that the rating really represents satisfactory performance. There are a variety of ways one could assess the construct validity of ratings, each of which has its own advantages and disadvantages. One simple way is to have independent judges examine a project (either directly or through project reports) and make an "expert" judgment of whether the assigned ratings are valid. Another would be to compare two independent, presumably valid rating systems applied to the same projects and determine whether the ratings are correlated. A third would be to correlate ratings with actual controlled project evaluation outcome assessments as a means of validating that the ratings reflect measureable results. A fourth would be to demonstrate that a set of ratings of multiple criteria behave as one would theoretically expect. For instance, multiple ratings of indicators of the same criterion should be more highly correlated with each other (convergent validity) than are multiple ratings or indicators of different criteria (discriminant validity).

Assessing the validity of a rating system is a complex challenging problem that is seldom approached and even less seldom successfully accomplished. Rating validity is affected by a wide range of factors including the quality of the criteria, the clarity of the indicator descriptions, and the degree to which evaluators sensibly translate or interpret evaluation results - that are based on actual program data, stakeholder interviews, desk review of documents - into ratings.

Reliability

Reliability refers to the consistency with which ratings can be done across raters, projects, and times. Probably most important for rating systems is determining the degree to which independent ratings of the same project produce consistent results. This is known as inter-rater reliability. A high-quality rating system will demonstrate that independent raters would give consistent ratings for the same projects when done at the same point in time. Reliability can be negatively affected by a number of factors. For instance, different raters may interpret the same criteria or indicator specifications differently. Or, raters can be affected by situational factors like the time of day ratings are done, types of evidence examined, mood of the raters, or even their demographic characteristics. High-quality rating systems provide empirical evidence that ratings can be consistently obtained across independent raters.

DESIGNING AND IMPLEMENTING RATINGS SYSTEMS

Organizations that attempt to construct valid and reliable rating systems face a number of important choices. For many of these choices, there is no empirical or other research base to guide decision-making. In the remaining part of this paper, we review some of the key issues organizations confront when developing, using and managing high quality rating systems.

Choice of Criteria

The rating criteria (or the dimensions on which performance is rated) are the broad standards by which the ratings are made and vary considerably from agency to agency. In general, agencies base performance ratings on some combination of "cross-cutting" and "sector-specific" criteria.

Cross-cutting criteria are general enough to be applied in the evaluation of any aid intervention. These include OECD/DAC's five evaluation criteria of relevance, efficiency, effectiveness, impact and sustainability. Many organizations frequently supplement the OECD/DAC criteria with three additional generally-applicable criteria put forth by the Active Learning Network for Accountability and Performance in Humanitarian Action (ALNAP). These include: connectedness, coherence and coverage. Other examples of general criteria used by agencies include "partner performance" and "innovation." Criteria such as these are "cross-cutting" because they are not subject or area-specific and consequently could be used

to evaluate virtually any project or program. ILO staff may wish to complement the "cross-cutting" criteria with ones that are specific to a particular sector. For example, the International Program on the Elimination of Child Labour (IPEC) might include criteria related to performance of a Time Bound Program (TBP) in one country or in a region.

Choice of Indicators

Rating criteria are judged using one or more pre-defined indicators. At the ILO, indicators are developed in the form of guidance questions or as specific pointers that are accorded to each rating criterion. The purpose of these questions is to enhance the objectivity of an evaluator's judgment in the process of assigning ratings. Presumably because any single indicator will necessarily be a fallible reflection of the typically complex criterion, using multiple indicators should enhance both the validity and reliability of the system. However, multiple indicators increase the burden of rating and may introduce a false sense of precision which would be justified only by a more thorough empirical examination.

Agencies often attempt to develop criterion-based indicators that are specific, measureable and achievable. Articulation of indicators of this type allows evaluators to assign ratings to each indicator as a sub-criterion. The overall rating for the criterion is then computed as an arithmetic average of its indicator ratings. Again, whether the multiple indicators can be reasonably aggregated is something that should be tested empirically prior to application of the rating system.

Choice of Response Categories

Response categories are a key component of the overall performance of a rating system. Response categories can be either even or odd in number. For example an even-numbered scale might be "highly satisfactory", "satisfactory", "unsatisfactory", and "highly unsatisfactory." Alternatively, a five-point, odd-numbered scale might be used by adding a middle category such as "neither satisfactory nor unsatisfactory." The advantage of using an even-numbered scale is that it forces the rater to make a choice and eliminates the tendency to choose the neutral middle category when they are uncertain or undecided.

Response categories at various organizations also differ in the number of categories included on the scale. Some prefer a four-category descriptive response scale. Others prefer six-category descriptive response scale as it presumably allows for a more nuanced assessment of intervention results. Additionally, a six-point scale also helps to overcome the reluctance of raters to attribute the best (4) or worst (1) score to interventions and this consequently tends to result in the clustering of ratings in the mid-range (2-3). Response categories across agencies are usually unidirectional, where the lowest numerical score reflects the most negative outcome and the highest reflects the most positive outcome.

Regardless of the number of points on the scale, clear definitions of response categories is essential because it allows for more clear differentiation and thus better facilitates the choice-making process. Agencies provide varying degrees of detail in defining response categories. While the World Bank provides general descriptions of each response category (for example, "highly satisfactory" means there were "no shortcomings," while a "satisfactory" reflects "minor shortcomings"), other agencies seek to contrast response categories by providing actual examples of projects receiving different overall ratings.

Nature of Measurement Scale

The nature of the measurement scale also needs to be considered. Ratings are made at one of four scales of measurement: nominal, ordinal, interval or ratio. The table found below explains these four levels using examples from daily life. Scales of measurement are important because they determine the statistical techniques that can legitimately be used to analyze the results of the rating (see far right column).

Table 1 Level of Measurement

Level of measurement	What it measures	Permissible Statistics
Nominal	Gender is an example of a variable that is measured on a nominal scale. Individuals may be classified as "male" or "female", but neither value represents more or less "gender" than the other.	Percents, Mode, Chi-square
Ordinal	An example of an ordinal scale would be the results of a horse race, reported as "win", "place", and "show". We know the rank order in which horses finished the race. The horse that won finished ahead of the horse that placed, and the horse that placed finished ahead of the horse that showed. However, we cannot tell from this ordinal scale whether it was a close race or whether the winning horse won by a kilometer.	Percents including mode and median
Interval	A perfect example of an interval scale is the Fahrenheit scale to measure temperature. The scale is made up of equal temperature units, so that the difference between 40 and 50 degrees Fahrenheit is equal to the difference between 50 and 60 degrees Fahrenheit. With an interval scale, you know not only whether different values are bigger or smaller, you also know how much bigger or smaller they are. For example, suppose it is 60 degrees Fahrenheit on Monday and 70 degrees on Tuesday. You know not only that it was hotter on Tuesday, you also know that it was 10 degrees hotter.	Mean, standard deviation, correlation, regression, analysis of variance
Ratio	The weight of an object would be an example of a ratio scale. Each value on the weight scale has a unique meaning, weights can be rank ordered, units along the weight scale are equal to one another, and there is an absolute zero (weightlessness).	All statistics permitted for interval level plus analysis of ratios.

At the level of rating individual criteria, most agencies tend to adopt an ordinal (ranked) scale. Typically, each criterion is scored separately and then averaged to give an overall project rating. Often, when project ratings are aggregated at a higher organizational level, such as the country program or agency level, donor agencies tend to assume an interval rating scale. An interval scale is assumed to be reasonable at higher levels where averaging often leads to final project ratings that include a decimal point. However, donor agencies seldom provide descriptions about threshold levels and explanations on the process of designating rating intervals. Other rating systems assume that ratings are ordinal and cannot be sensibly averaged. Such systems typically rely on percentages (such as the percent of projects having a satisfactory or highly satisfactory rating) when aggregating results.

The issue of whether ratings should be treated as ordinal or interval level has both methodological and interpretive implications. If one assumes the ratings are only ordinal, it makes no statistical sense to average them. For instance, assume that two projects receive a rating of 2 (unsatisfactory) and 3 (satisfactory) respectively on a four-point scale. Is it reasonable to assume that the average across these two projects of 2.5 is a meaningful number? And, if it is, how would one interpret this average? If one believes such an average can be meaningfully interpreted, then it may be justifiable to treat the ratings as interval-level and use averaging to aggregate them. If not, then the data would probably be considered ordinal and some form of percentage-based tabulation would be used for aggregation.

Who Rates and When are Ratings Conducted

Ratings are conducted at multiple levels (project, program, strategy, and agency). At the project level, ratings may be conducted by project staff and/or external evaluators. Generally, final project ratings are conducted solely as self-assessments by the project staff at specific times during implementation or at project completion. However, obtaining objective ratings based on self-assessments may be problematic, especially if managers fear reprisals or funding cutbacks for poor performance ratings. Many UN agencies include some sort of a validation process to counter subjectivity in self-assessed ratings.

In other cases, UN agencies require the undertaking of self-assessed ratings as an input into subsequent independent project evaluations. The independent evaluation is typically conducted by an external evaluator. The final project ratings are produced by the external

evaluator as an outcome of the evaluation process. These ratings are based on actual program data, interaction with beneficiaries and stakeholders as well as on project performance documents (which include self-assessed ratings). Finally, some UN agencies do not require project staff to prepare self-assessed ratings. Final project ratings for these agencies are produced completely as an outcome of the independent evaluation process.

At higher organizational levels, ratings are generated by the Monitoring and Evaluation department at agency headquarters. The M&E staff aggregates individual project ratings by geographic region and/or key strategic goal areas in preparing a meta-analysis at the country or agency level. Because ratings at higher organizational levels are primarily based on project ratings, they are conducted after project-level evaluations are completed. Many donor agencies also recruit independent evaluators to conduct country program or agency level ratings.

In special cases, project ratings are conducted by headquarters M&E staff after the completion of the independent external evaluation of the project. In other words, the headquarters M&E staff rates projects ex-post based on the assessment provided in the final evaluation report. While project evaluations are conducted independently, they may not be mandated to include project ratings. In order for the agency to produce a meta-analysis, the M&E staff is required to rate based on interpretation of written evaluation reports. The M&E staff then aggregates these ratings to develop a meta-analysis at the country or agency-level.

The issues of who does the ratings and when in the reporting/aggregating process they are done are critically important to the quality of the rating system. The presumption typically is that ratings done by project staff or managers are more likely to be biased than those done by independent evaluators. However, a reasonable counter-argument may be that project staff and managers have more intimate knowledge of the project and are consequently in a better position to make valid assessments. Independent auditing, systematic comparative ratings, and other mechanisms for cross-checking ratings for quality are essential for establishing the credibility of the rating system regardless of which choices are made.

Aggregation

Ratings are aggregated in various ways. The smallest unit of analysis to receive a score is usually the rating criteria. Each criterion typically receives a distinct score. As mentioned earlier, either the scores from every criterion are averaged to come up with an overall project rating or some percentage is applied across criteria to arrive at an aggregate value. In some cases, only the "core" criteria are averaged and the evaluator is asked to make an informed judgment regarding the overall project rating based on the averaged "core" rating and individual ratings provided to each supplementary criterion. In generating an average rating, agencies do not typically use weights for different criteria.

The most common aggregation practice involves categorizing the percentage of projects with ratings above or below a particular response category (e.g., xx% of projects/programs receiving a satisfactory or higher rating). Many UN agencies use this approach to report on agency performance across the whole portfolio, within different goal/sub-goal areas, or within different geographic regions. Problems, however, still remain with the meaning and comparability of ratings across a wide diversity of projects and country settings. Some agencies also report problems with coverage -- i.e., not all projects are routinely rated. In almost all cases, only completed projects are rated. UN agencies also employ stratified sampling methods to randomly sample projects by relevant agency goal areas. For example, in a recent meta-evaluation, the ILO aggregated projects ratings by the agency strategic framework – 59 projects were randomly sampled and stratified by 19 outcomes in the agency-level strategic framework.

Only a few agencies use weights in the process of aggregation. For instance, some of the multi-lateral development banks weight individual projects by amount of disbursements (size of loans and credits). These weights are taken into account when ratings are aggregated at country, program or agency-level. Another example of a weighted approach is when organizations weight individual rating criterion in the overall aggregation process – The ILO employs this method in its Evaluability Assessment tool.

CONCLUSION

Rating systems present organizations with a fundamental challenge. While on the one hand, such systems have important advantages – mainly, they facilitate the quantification of qualitative judgments and allow for performance aggregation within and across projects and programs. On the other hand, there also exist key methodological and interpretive issues that threaten the accuracy and consistency of rating systems.

These issues lead us to examine a basic question: how can we assess and ensure the quality of rating systems in international evaluation? This paper has attempted to frame an answer to this question by first laying out the important elements of rating quality and then providing an overview of the choices international organizations make in developing and implementing rating instruments.