Knowledge Sharing Program

2012 Modularization of Korea's Development Experience:

Performance Management System of Budgetary Programs in Korea

2013

MINISTRY OF STRATEGY AND FINANCE

Korea Fixed Income Research Institute

2012 Modularization of Korea's Development Experience:

Performance Management System of Budgetary Programs in Korea

2012 Modularization of Korea's Development Experience

Performance Management System of Budgetary Programs in Korea

Title	Performance Management System of Budgetary Programs in Korea
Supervised by	Ministry of Strategy and Finance (MOSF), Republic of Korea
Prepared by	Korea Fixed Income Research Institute
Author	Chang Gyun Park, Chung-Ang University, Professor
Research Management	KDI School of Public Policy and Management
Supported by	Ministry of Strategy and Finance (MOSF), Republic of Korea

Government Publications Registration Number 11-7003625-000076-01

ISBN 979-11-5545-035-2 94320

ISBN 979-11-5545-032-1 [SET 42]

Knowledge Sharing Program

Government Publications Registration Number
11-7003625-000076-01

Knowledge Sharing Program

2012 Modularization of Korea's Development Experience

Performance Management System of Budgetary Programs in Korea

MINISTRY OF STRATEGY AND FINANCE

Korea Fixed Income Research Institute

Preface

The study of Korea's economic and social transformation offers a unique opportunity to better understand the factors that drive development. Within one generation, Korea has transformed itself from a poor agrarian society to a modern industrial nation, a feat never seen before. What makes Korea's experience so unique is that its rapid economic development was relatively broad-based, meaning that the fruits of Korea's rapid growth were shared by many. The challenge of course is unlocking the secrets behind Korea's rapid and broad-based development, which can offer invaluable insights and lessons and knowledge that can be shared with the rest of the international community.

Recognizing this, the Korean Ministry of Strategy and Finance (MOSF) and the Korea Development Institute (KDI) launched the Knowledge Sharing Program (KSP) in 2004 to share Korea's development experience and to assist its developing country partners. The body of work presented in this volume is part of a greater initiative launched in 2010 to systematically research and document Korea's development experience and to deliver standardized content as case studies. The goal of this undertaking is to offer a deeper and wider understanding of Korea's development experience with the hope that Korea's past can offer lessons for developing countries in search of sustainable and broad-based development. This is a continuation of a multi-year undertaking to study and document Korea's development experience, and it builds on the 40 case studies completed in 2011. Here, we present 41 new studies that explore various development-oriented themes such as industrialization, energy, human resource development, government administration, Information and Communication Technology (ICT), agricultural development, land development, and environment.

In presenting these new studies, I would like to take this opportunity to express my gratitude to all those involved in this great undertaking. It was through their hard work and commitment that made this possible. Foremost, I would like to thank the Ministry of Strategy and Finance for their encouragement and full support of this project. I especially would like to thank the KSP Executive Committee, composed of related ministries/departments, and the various Korean research institutes, for their involvement and the invaluable role they played in bringing this project together. I would also like to thank all the former public officials and senior practitioners for lending their time, keen insights and expertise in preparation of the case studies.

Indeed, the successful completion of the case studies was made possible by the dedication of the researchers from the public sector and academia involved in conducting the studies, which I believe will go a long way in advancing knowledge on not only Korea's own development but also development in general. Lastly, I would like to express my gratitude to Professor Joon-Kyung Kim and Professor Dong-Young Kim for his stewardship of this enterprise, and to the Development Research Team for their hard work and dedication in successfully managing and completing this project.

As always, the views and opinions expressed by the authors in the body of work presented here do not necessary represent those of the KDI School of Public Policy and Management.

May 2013

Joohoon Kim

Acting President

KDI School of Public Policy and Management

Contents | LIST OF CHAPTERS

Chapter 3

Contents | LIST OF CHAPTERS

Chapter 4

Contents | LIST OF TABLES

Chapter 2

Chapter 3

Contents | LIST OF TABLES

Chapter 4

Contents | LIST OF FIGURES

Chapter 2

Chapter 3

Contents | LIST OF FIGURES

Chapter 4

2012 Modularization of Korea's Development Experience
Performance Management System of Budgetary Programs in Korea

Chapter 1

Introduction

Introduction

Korea launched a major reform to introduce performance-based budgeting into the government sector in the 2000's. The Korean case is particularly interesting in that the government pushed the reform ahead very rapidly while other reforms in the budgetary system were also pursued concurrently known as the Four Major Fiscal Reforms, which provided an extraordinarily favorable environment for building an effective performance management system. Due to such a big push forward on a large scale, the Korean government was able to establish a comprehensive and robust performance management system in a short period of time. The Four Major Fiscal Reforms consisted of the establishment of a medium-term expenditure framework known as the National Fiscal Management Plan, introduction of top-down budgeting, establishment of the performance management system, and building of a digital budget information system. The reform process has not been completed yet and the Korean government is still committed to providing a significant amount of time and resources to making the reform successful and to building a strong and efficient budgetary system. It might be too early to cast a verdict on the reform process but many commentators offer favorable opinions on the accomplishment of the reform efforts. If the reform is completed and turns out to be successful, Korea will possess a very strong and efficient budgetary system that incorporates virtually all of the best practices.

These ambitious reforms were motivated by the deteriorating fiscal conditions and prospects in Korea. The Korean government experienced a dramatic increase in public debt after the Asian financial crisis in the late 1990s. The growing debt was mainly driven by a rapid increase in public expenditures to strengthen the social safety net which became an urgent policy agenda in response to widening income disparities resulting from the economy-wide restructuring. Looking ahead, the aging population in Korea is progressing at a pace that is unprecedented among countries, generating additional pressure on public finances.

The medium-term fiscal plan puts government spending decisions in a five-year framework. Based on prudent projections on future economic growth, the plan determines the level of annual overall expenditure over the medium term and allocates the total amount available among major sectors of government spending. Consistency between such medium-term resource allocation decisions and annual budget appropriations are enforced through the top-down budgeting system. The system assigns firm spending ceilings on the expenditure of each ministry according to the medium-term fiscal plan, but delegates lower-level budgeting decisions to ministries, provided that the latter's aggregate expenditures remain within their assigned ceilings. The greater autonomy given to the ministries in turn requires greater accountability on their part. This is ensured through the performance management system, which was introduced to monitor and analyze the performance of government spending programs andthus strengthen the link between budgeting and performance. The digital budget information system allows the budget office to monitor the ministries' spending in real time.

Performance management system was introduced to Korea in four phases. The first phase was the experimental pilot project carried out during 2000-02. The project experimented with a modified version of GPRA (Government Performance and Results Act) from the United States. Twenty-two ministries and the program agency that participated in the project were asked to develop annual performance plans. The pilot project was terminated as a new administration was inaugurated. Building on that experience, the second phase began as a core component of the Four Major Fiscal Reforms in 2003. Twenty-two ministries and agencies were selected and asked to submit their annual performance plans along with their annual budget requests. The second initiative was also inspired by the GPRA but implemented only on a limited subset of GPRA features. The third initiative, the Self-Assessment of the Budgetary Program (SABP), was introduced in 2005. This system was basically based on the "Program Assessment Rating Tool (PART) of the United States, with some modifications. Under the SABP, about a third of all government programs have been reviewed every year, a pace which would allow the Ministry of Strategy and Finance (MOSF) to review every major budgetary program over a three-year cycle. Each ministry and program agency selected for SABP was asked to fill out the checklist that includes questions on planning, management and results of a government expenditure program. The fourth phase started in 2006 with the launch of In-depth Evaluation of Budgetary Program (IEBP). IEBP is a Korean version of program evaluation that examines the performance of government expenditure programs with analytical and scientific methods typically by external experts. The results of IEBP are incorporated into the budget process to improve program performance.

The Korean performance management system is still a work-in-progress. It is very difficult to predict how the system will evolve in the future. However, many practitioners and researchers seemed to have reached an agreement that the reform effort by the Korean government to establish a robust and efficient performance management system is a sign of success. This report documents the reform process for introducing the performance management system into the Korean government and summarizes lessons learned.

2012 Modularization of Korea's Development Experience
Performance Management System of Budgetary Programs in Korea

Chapter 2

Theoretical and Practical Aspects of Performance Management

Theoretical and Practical Aspects of Performance Management

1. Performance Management

The policy cycle in the public sector consists of four phases; inputs, process or activities, outputs, and outcomes. Conceptually, performance management means the efforts to manage the policy cycle systematically for the purpose of achieving the mission or objectives assigned to the public sector. Performance management can be classified into three categories based on the developmental stage of the concept; traditional performance management, result-oriented performance management, and integrated performance management.

The traditional performance management system focuses on the early phases of the policy cycle. It is based in the presumption that an appropriate control of inputs and processes lead to achieving higher outputs and better outcomes. The advocates of the system argue that much attention should be paid to establishing sound management and a control system centering on inputs and process phases rather than outputs or outcomes phases. Unlike the private sector where it is relatively easy to identify tasks and performances of an organization, outputs and outcomes provided by the public sector are difficult to measure or delineate the boundary that the traditional performance management system put much emphasis on inputs and process. Consequently, program agencies have a tendency to make sure that inputs are used appropriately following the due process. Unfortunately, the practice creates an obstacle to achieving efficiency, as well as effectiveness of the public sector. In other words, the members of a public organization are not interested in achieving the mission of the organization in an effective and efficient way, but in strengthening discretionary power by increasing the amount of resources under control or complying with the procedures or processes. It is highly likely that program managers, as well as their staffs and employees, mechanically follow past practices and display extremely risk adverse behaviors.

As an optimistic expectation of the traditional performance management system crumbles, the societal pressure on the public sector has accumulated to a point that justifies spending on the programs managed by the public sector. A convincing way to persuade the taxpayers is to show them the benefits brought to society by the programs and result-oriented performance management system. Result-oriented performance management system pursues to manage and control activities of the public sector based on performance, free from bureaucratic control over inputs and process under the traditional system. It is not enough for the personnel of a public organization to abide by the rules and procedures. They have to show the results and performance that taxpayers ultimately appreciate. A significant degree of autonomy in personnel and budget is granted to the organization in exchange for the promise of results and performance. We can understand result-oriented performance management system as serious attempts to transplant management principles from private organizations into the public sector since public organizations are held accountable for results in exchange for extended organizational autonomy. The term performance has different meaning in integrated performance management system from the traditional or result-oriented performance management systems. Both the traditional and the result-oriented performance management systems concentrate only on one aspect of the policy cycle. On the contrary, performance in the integrated performance management system means the achievement relative to objectives in every aspect of the policy cycle such as inputs, process, outputs and outcomes.

Modern performance management system can be defined as a systematic process in which all members of an organization are engaged to accomplish mission and goals of the organization efficiently. Specifically, the fundamental goal of a performance management system is to enhance the organization ability to accomplish its mission and objectives efficiently. The system consists of several important components; to make a plan to achieve a mission from a strategic point of view, to carry out the plan by using scarce resources efficiently, to measure performance of individual members and the organization as a whole, to provide feedback of the results of measurement to improve the policy process and resource allocation by linking them with compensation. To sum up, a performance management system is a consistent and cycling management system consisting of four stages; performance planning, program design and execution, performance evaluation, and feedback.

The first stage in the performance management system is performance planning where performance goal is identified. The performance goal is the state the program agency tries to achieve through budgetary programs. The performance goal should be clearly specified in a clear and concrete manner to contribute to achieving the mission and objectives of the program. An important task in performance planning is to select performance indicators and their target levels.

The action plan of the program is designed and executed in the second stage. The program agency and delivery system for the program services are selected. In addition, a timetable for commitment of budgetary resources should be fixed at this stage. Moreover, a detailed plan concerning how the performance evaluation is actually executed and data are collected should be discussed. Executing the program as planned, the program agency should make sure that the performance indicators are measured on a regular basis and enough data necessary to carry out a high quality evaluation are accumulated.

Various forms of evaluation on the performance of the program are executed in the third stage. Records on the measurement of performance indicators from the previous stage and related information are utilized as important inputs. A variety of evaluation methodologies can be employed, including performance monitoring, program evaluation, and job evaluation. Performance monitoring, also called performance measurement is a very popular evaluation technique in which performance indicators are measured and compared to the target levels regularly. The program agency in charge of execution of the program itself leads the process and outside organizations are involved for verification of the self-evaluation. Performance monitoring is a relatively inexpensive and convenient evaluation technique and can be utilized to revise the business plan or improve the efficacy of the delivery system of the program. However, performance monitoring cannot provide information on the contribution of the program in accomplishing the performance goals established in the planning stage without questioning the appropriateness of the performance indicators or relevance of the program objectives. Performance monitoring also has a limitation in portraying performances of programs with multiple objectives since it tracks small numbers of performance indicators to ensure promptness and convenience of the evaluation. Nonetheless, those features of performance monitoring have prompted many governments around the world to introduce it into the performance management system. According to the EU (2005), the program evaluation is the judgment of government intervention based on its results, impacts, and needs it aims to satisfy. That is, the program evaluation is a process to produce information to be used in decision making on whether to continue the program or how to modify the program to improve the performance by examining the accomplishments in a scientific and objective manner. Program evaluations are in-depth examinations on the causal relationship between the program activities and program performances. In addition, comprehensive and scientific investigations on the program objectives, delivery system, and performance indicators are carried out in program evaluations. Due to their in-depth and complete nature, it takes significant amounts of time and resources to conduct full scale evaluations for all programs. Therefore, it is often the practice to select a small number of programs for special attention and to organize evaluation projects. Another popular technique in performance evaluation is job evaluation. Job evaluation determines the relative worth of jobs in carrying out tasks by the program agencies. The evaluator assigns rank to each job according to

its relative worth based on an important criteria such as roles and responsibilities of jobs, and difficulties and complexity of tasks. Job evaluation relates output and results of an organization or a program to each program personnel. It is, therefore, not unusual to use the results of job evaluations in assessing promotions or compensation for program personnel.

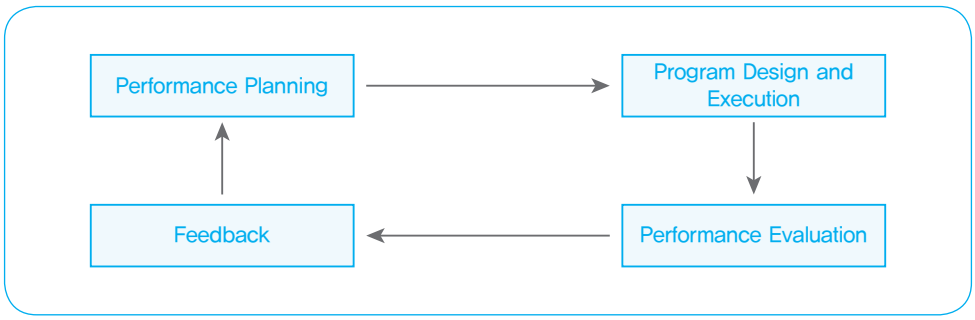
Though not included in the performance management system, performance audits have attracted the attention of practitioners recently. Performance audit examines whether activities of a public organization comply with the 3E (economy, efficiency, effectiveness) standards. Specifically, the task list of the performance auditor includes thorough investigations on the economy of activities of the public agency, efficient use of public resources, and effectiveness or accomplishment of the public organization. Under the traditional environment, performance audit used to put much emphasis on financial audit or examining the legality of decision making or actions by program agencies. Recently, performance auditors shifted their attention to effectiveness or efficiency of public agencies as result-oriented performance management systems take place of the traditional systems focusing on inputs or process. Unlike performance evaluations, such as performance monitoring and program evaluation, performance audit is not carried out or commissioned by the agency that is in charge of executing the program but initiated and conducted by independent external agencies. Performance auditors generally show a limited level of interest than program evaluators, examining long-term sustainability or adequacy of the activities of public agencies in addition to the traditional standards like 3Es.[1] While many performance evaluation techniques have been suggested and tried in various evaluation studies, performance monitoring and program evaluation takes the central place in the field and some commentators treat performance management equivalent to performance monitoring and program evaluation.[2]

In the fourth and final stage of the performance management system, the results of the performance evaluation are sent back to various stakeholders and efforts are made to improve the program performances by incorporating the feedback. The purpose of the performance evaluation is not the evaluation itself but to utilize the results to improve program performance. Program agencies may revise the program design, modify delivery system of services, and reshuffle investment plans on human and material resources. The results of performance evaluation can be used as a vehicle to secure the accountability of the program agencies. Therefore, it is important to ensure practical usefulness, feedback and utilization of the evaluation results through legislative supports.

1. For further discussion on performance audit, see Davis (1990) and GAO (2008).

2. See Perrin (2003) for recent development in researches on performance management systems.

Figure 2-1 | Performance Management System



2. Performance Planning

At the first stage of the performance management system, performance planning starts with clearly delineating the mission of an agency that manages government expenditure programs. With a well written mission statement, one can proceed to set up strategic goals and performance goals. The former indicates the medium-to-long term plan and the latter shorter-term schemes. The final core step in performance planning is to adopt performance indicators that measure the achievement of the performance goals in an objective and verifiable manner. In the following, we divide performance planning into three stages and discuss the procedures and methodologies one can utilize in each stage. The three stages are documentation of mission statement, establishment of strategic goals and performance goals, and selection of performance indicators.

2.1. Documentation of Mission Statement

The mission statement should identify the major results an organization or a program seeks. It is the starting point for identifying outcomes to be measured and the performance indicators needed. The term mission denotes both the over-reaching vision of the organization or the program and the more specific purposes that flow from the mission. Therefore, the mission normally should be stated in general, not quantitative terms and should remain relatively stable. Note that specific targets especially defined in quantitative terms are likely to change, often frequently, because of new circumstances.

The basic form of the mission statement consists of two parts, purpose (To-part) and tools (By-part). To-part of the mission statement identifies the basic objectives or results the organization or a program seeks while the By-part of the mission statement identifies the basic way the service is provided by the organization. An example from the U.S. Department of Education's distance learning program is shown below.

To improve student learning and employability, including providing access to, and improving instruction in, a wide range of subjects by the use of distance-learning technologies.

Note that the To-part includes both end outcomes (improved student learning and employability) and intermediate outcomes (improved instruction and student access to a wide range of subjects). The general approach of the program is use of distance learning technologies. Specific technologies are not mentioned to avoid limiting the options of those delivering the program services. Some programs use wording that implicitly places the By-part first. For instance, the above example can be written in a different way like;

Use distance-learning technologies that improve student learning and employability, including providing access to, and improving instruction in, a wide range of subjects.

This is not a good practice. Leading with the To-part keeps the focus more immediately and therefore more strongly on results. A By-part of the mission statement may not be necessary for programs whose approach is expected to be clear to users of the performance information. Basic municipal services, such as waste collection and recreational programs, for example, are sufficiently clear in their approaches that even a good By-part is not likely to be needed or helpful.

The following are suggestions for developing a mission/objectives statement;

- Focus on how program activities are expected to affect both the program's specific customers and the public at large.
- Identify all the major objectives that the program hopes to achieve. Most programs have multiple objectives. It is better to include too many objectives in the statement than to run the risk of excluding objectives that may later be found important to one or more customer groups.
- Call explicitly for minimizing negative effects of the program. Transportation is a good example of a program with negative effects that can be anticipated. Pollution is an inevitable by-product of transportation. Therefore, the mission statement of a transportation program might well include the words "and to minimize air, water, and soil pollution".
- Include conflicting objectives as appropriate, and recognized in the statement the need to balance them. Environmental and economic development programs, for example, have potentially conflicting effects on each other. Public land management program may need to aim a balance between promoting economic development and preserving green space, flora, and fauna.

- Consider including objectives about reducing the magnitude of unmet needs, not just about helping customers who come in for service. For example, "reduce the number of households with incomes below the poverty level."
- Include objectives that are related to the quality of services delivered – characteristics that are important to customers, such as timeliness and convenience of the help received. While these qualities are intermediate outcomes rather than end results, their importance to customers may warrant their explicit inclusion in a program's objectives, to help ensure that they receive ongoing attention.
- Include the objective of providing a service as efficiently as possible. This is an objective of virtually all programs even if only implicitly. It explicit inclusion even serves to remind program personnel of its importance.
- Include only qualitative, not quantitative, objectives to enhance the likelihood that the statement will remain stable over time. Numerical targets should be avoided because they are unlikely to be valid for longer than one measurement period. In some instances, public officials have chosen to include long-term numbers in their mission statement, such as "By 2015, the outcome indicator will double" or "by 2015, our jurisdiction will have the best outcome indicator value in the world." These are likely to be political statements aimed at securing public support. In general, they should be avoided.
- Avoid vague or obscure wording that makes later measurement a guessing game about the statement's original intent. The strategic plan for one state's transportation department included "having all transportation systems and services work smoothly together." Such statement makes it very difficult to determine how to track progress toward that objective.

The mission statement should clearly identify who the program's or the organization's customers are, unless it is already obvious to users. Almost always, organizations and programs have multiple categories of customers. Questions such as the following are helpful in identidying customer information from each source such as;

- Who benefits from the services provided by the program or the organization? Who are direct recipients? Who are indirect recipients?
- Who might br hurt by the activities program? – This question may also help identify potential negative effects of the program that should be identified in the mission/ objectives statement.
- What other people not directly targeted by the program can be significaltly affected by it?

- Which demographic or interest groups are particularly affected by the program?
- Is the public at large likely to have a major interest in what the program accomplishes rather than just what it costs? For a program that help businesses reduce hazardous waste and poluution generated by their activities, for example, the general public clearly is a major customer. But the assisted businesses are also customers, and the performance measurement process should include outcome indicators that address their concerns as well, such as higher costs.

As with defining the mission, it is often very complicated than it first appears to determine the program's customers. These complications, as well as difficulties in identifying the mission, are not caused by the performance management system. The complications are already there and should not be ignored in a comprehensive performance managment system.

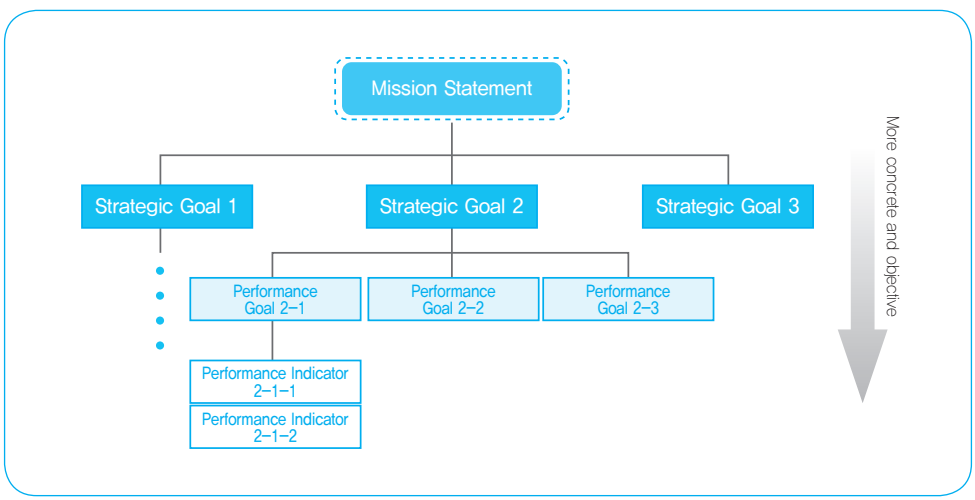
2.2. Establishment of Strategic Goals and Performance Goals

Strategic goals can be defined as the medium-to-long term objectives of an organization and may include objectives, value and function of an organization. An organization conducts various activities or programs to fulfill the mission or fundamental objectives stated in the mission statement. It is a common practice in performance management to classify into several categories all programs of an organization according to important features such as objectives and customers and then assign to each category an objective that can be shared by the programs in the same category.

Strategic goals can also be defined as the common objectives of each category of programs an organization runs. There are several important points that require our attention in setting up strategic goals. First, strategic goals should be value neutral and objective. While the mission statement can include, to some degrees, abstract or value related expressions, strategic goals should be objective and concrete as much as possible. Second, strategic goals should be written in a terse manner. One should avoid listing too many strategic goals and try to minimize the number of strategic goals directly related to core programs. The fundamental reason we explicitly establish strategic goals is to convey information on the function and objectives of an organization to both employees and customers, it is not desirable to have a long list of strategic goals written in a lengthy language. Third, multiple organizations belonging to different administrative units may share common strategic goals as long as they are involved in a set of programs with common objectives. Fourth, several important points in writing up the mission statement should also be applied to strategic goals. For example, strategic goals should focus on the final results rather than inputs and keep the balance between conflicting policy objectives, if they exist.

The performance goal is a concept subordinate to the strategic goal and spells out the concrete targets programs seek to achieve. Performance goals should be defined in more specific terms than strategic goals so that they are measurable with the performance indicators we will discuss later. Performance goals are used as baseline information in evaluating the efficacy of programs and convey the information on the expected performance levels to the interested parties and the general public. Programs can be consolidated or disaggregated in establishing performance goals. Multiple programs in the same account or fund should be aggregated if they have common performance goals. On the other hand, a single program should be disaggregated into several sub-level programs if it possesses multiple performance goals. In addition, several programs belonging to different accounts or funds should be consolidated if they share the same goals. [Figure 2-2] illustrates the hierarchical structure in the mission statement, strategic goals, performance goals, and performance indicators. The structureis based on the degree of concreteness and objectiveness of each concept.

Figure 2-2 | Hierarchical Structure of Performance Planning



2.3. Selection of Performance Indicators

The performance indicator is a tool with which the achievement of a performance goal is examined. Selection of performance indicators is the last step in performance planning that has a hierarchical structure within the performance management system as shown in [Figure 2-2].

2.3.1. Objects of Measurement

The central function of performance indicators is to provide regular and relatively accurate information on the way a program is executed and whether the effects of a program are realized. Therefore, performance indicators should not be limited to information on outcome or efficacy of a program but include information on inputs or activities even though we focus on results rather than inputs in performance management. A consistent set of definitions categorizing various types of performance information is the cornerstone of any performance management system since, all too often, much confusion arises due to unclear, inconsistent use of terms. No two people will reach an easy agreement on how to categorize performance information into several meaningful groups. A gray area exists because it is not always clear where a particular piece of information falls. In addition, for some performance information, the category may depend on the perspective of the agency that is in charge of the programs.

One of the most popular ways to classify performance information is to categorize them according to the sequence a program is executed and the program's effects are realized; inputs, process, outputs and outcomes. Information on the amount of resources expended for particular programs (inputs) differ from internal information on the amount of activity a program is undertaking (process), These types of information, in turn, differ from the products and services produced or provided by a program(outputs), which should be distinguished from result-based information (outcomes). Each of these categories is discussed briefly in turn.

Input information is the amount of resources used in a program and typically expressed as the amount of funds or the number of man-hour injected into the program. For performance management purposes, the amounts that were actually used, not the amounts budgeted, are the relevant information. Agencies occasionally call the work that comes into them an input but those kinds of information should not be regarded as input information because the amount of incoming work, i.e., workload is different from the amount of pecuniary resources or staff time expended.

Process information is also called workload information or activities information. It includes the amount of work that comes into a program or is in process but not yet completed. The amount of work is not considered proper performance information that it does not indicate how much product the program generates. Workload information provides useful data, however, when program managers want to track the flow of work into and through their programs. While the amount of work by itself cannot be performance information since it does not offer information on outputs or outcomes, workload data can be used to produce outcome information. In some programs, the amount of work not complete at the end of a reporting period can be considered a proxy for delays of services to customers.

Output information refers to the amount of products and services delivered during a reporting period. It is a common practice among agencies to report output of a program on a regular basis and keeping track of the amount of output achieved is a good practice. Examples of outputs include kilometers of roads paved, reports issued, training sessions held, and low-income single mother served by the program. Outputs do not by themselves tell anything about the results that are the ultimate target a program pursues to achieve. That is, outputs are what a program's personnel have done, not changes to people outside the agency in charge of the program or changes that outside organization made.

Outcomes are events, occurrences, or conditions that indicate progress toward a program's mission and objectives. Therefore, outcomes have direct links to the program's overall mission and are of utmost importance to customers and the general public. Outcomes are not what the program itself did but the consequences of the program. An example may help understand the difference between outputs and outcomes. The number of patients treated or discharged from a state mental hospital is different from the percentage of discharged patients who are capable of living independently. The former is output information, the latter outcome information. Here is another example. Suppose that sewer cleaning crews are rated based on how many miles of streets they clean. Crews may have the tendency to focus their operations on cleaning sewers they have already cleaned because cleaning dirty sewers slows them down. However, it is not the practice that the designer of the sewer cleaning program originally intends since the purpose of cleaning sewers is to keep sewage away from backing up into people's homes and businesses, not to rack up cleaning mileage.

Outcomes should be something the programs intend either to maximize, such as average income or better nourishment, or to minimize such as crime rate or poverty rate. Some outcomes are financial in nature. For example, the outcome of a public assistance program is the reduction of the dollar amount of incorrect payment whether overpayment or underpayment and the amount of owed child support payments recovered from absent parents is an appropriate outcome for child support offices.

In these cases, outcomes can be expressed in monetary terms. Outcomes include side effects, whether intended or not and whether beneficial or detrimental to customers or the general public. If one can anticipate the possibility of serious side effects intended or accidental and beneficial or detrimental, the performance management system should have built-in procedures measuring them on a regular basis. In addition, as long as outcomes are important and can be tracked down with reasonable easiness, they should be included in a performance management system, even if they are not explicitly identified in the program's mission and objective statement. We should bear in mind the fact that no formal program mission and objective statement include all important outcomes an agency needs to track. It is not the function of such statements to list all the outcomes the program should seek, just the most important and central ones.

In measuring outcomes, it is particularly important to distinguish intermediate outcomes from end outcomes. The distinction will help us differentiate between the ends ultimately desired from the program and interim accomplishments, which are expected to lead to the final ends. However, in many cases, it is not clear-cut to distinguish between intermediate outcomes and end outcomes but we can offer two useful criteria to differentiate the two key dimensions of outcomes; their importance and when they occur. The primary criterion should be the importance while when outcomes occur should be utilized as a complimentary device. Though it is usual that intermediate outcomes occur before the end outcomes, an end outcome can occur very early.

Intermediate outcomes are those which are expected to lead to the ends desired but are not themselves ends. Examples of intermediate outcomes include the following;

- People completing employment training programs where program participation is voluntary. This reveals how successful the program has been in convincing customers not only to participate in, but also to complete, the sponsored training sessions. However, completion is only one step toward the ultimate end of improving the condition of people in the program.
- Citizens exercising more or switching to a better diet, as recommended in an agency-sponsored health program (perhaps as measured by surveying clients 12 months after completing the agency's program). Such changed behavior is expected to lead the participants to better health, but since this connection is uncertain, the behavior is an intermediate outcome.
- A state or local agency developing a comprehensive plan of action encouraged and supported by a federal program (where acceptance of the assistance is voluntary). For the federal government, states or local governments actually completing a reasonable plan can be considered an initial step toward improving services, although completing the plan says little about the end outcome of the service improvement.

End outcomes are those which are the desired results of the program. They are conditions of the fundamental importance to the general public as well as program customers. End outcomes might, for example, be aspects of health, safety, educational achievements and earnings, or decent housing and neighborhoods, such as reduced incidence of specific diseases, improved student test scores, lower crime rates, less violence in schools, reduced number of households living in substandard housing, increased real household earnings, and reduced household dependency on welfare. For some programs, customer satisfaction with the results of a service can be considered an end outcome even though those programs have aims that go beyond satisfaction. For example, customers' satisfaction ratings of libraries can be considered end outcomes. Note that the program's mission to enhance access to

the library would be to increase public awareness of literature rather than to increase the number of visitors or the number of books borrowed. Many programs produce both short-term and long term end outcomes. Education is a classic example. Many programs in education not only produce early improvements in student learning and self-esteem, but they help students enjoy a better chance of employment and higher salaries later on. In most cases, long-term end outcomes cannot be utilized to guide program personnel on the success of most of their current activities since they are not available early enough. Therefore, it is very important to identify short-term end outcomes to assess and encourage ongoing program improvements. Short-term outcomes are tracked. Short-term outcome themselves also have value. Intellectual development and dropouts rates, for example, are outcomes of key concern to education managers, staff, and parents, and can be considered end outcomes for this reason.

There are several issues on-going discussions regarding the relationship between intermediate and end outcomes. First, intermediate outcomes, by definition, occur before end outcomes and are expected to lead to them. Thus, intermediate outcomes usually provide more timely information than end outcomes. For example, customers complete job training programs (intermediate outcome) before they obtain employment (end outcome), which is expected to occur after the completion of the program. When a program has long-term end outcomes for which data may not be available for many years (such as reduction in adverse health effects due to smoking and achieving rewarding employment careers), the program can usefully focus on short-term ends (such as reduced smoking and improved learning skills). It, however, should be noted that program designers should choose intermediate outcomes that are proven to have a close relationship with end outcomes. Second, early occurrence of an outcome does not necessarily mean it is not an end outcome. For example, family counseling programs aim to produce more stable and happier families in the short run as well as in the long term. Some interventions quickly produce end outcomes, while others require many years before end outcomes start to appear. Thus, it is very important to pay attention to the proximity to the mission, as well as the time they occur in tracking outcomes. Third, intermediate outcomes usually are related to the particular way the program delivers the service, whereas end outcomes typically do not vary with the delivery approach. For example, a government attempting to improve the quality of rivers and lakes can achieve this goal in many ways, such as by providing funding for wastewater treatment, providing technical assistance to certain classes of businesses and encouraging lower levels of government to pass stricter laws and ordinances. Each of these approaches would have its own intermediate outcomes. But regardless of the approach, the same end outcomes, such as the quality of rivers and lakes, apply.

In addition to the indicators to quantify the four important aspects of a program, program managers and stakeholders may also be interested in information on quality of services provided by the program, degree of customer satisfaction, and productivity or efficiency of the program.

The most common dimension of the quality of public services are timeliness, turnaround time, accuracy, thoroughness, accessibility, convenience, courtesy, and safety. Thus, the percentage of customers who wait in line more than fifteen minutes before being able to renew their driver's license, the number of calls to a local child support enforcement office that are returned within twenty-four hours, the percentage of claims for disability benefits that are not adjudicated within seventy working days are typical quality indicators. Quality indicators are often process indicators measuring compliance with established standards, such as the percentage of highway maintenance jobs that are performed according to prescribed operating procedures. Others focus on the quality of the outputs themselves and the need for rework, such as the number of completed highway crack sealing projects that have to be repeated within a year.

Measures of customer satisfaction are often closely related to service quality and program effectiveness, but it may be more helpful to consider them as constituting a separate category of performance measures. For example, measures of customer satisfaction with a vocational rehabilitation program might be based on data from client evaluation forms asking how satisfied they were with various aspects of training, counseling, and placement assistance they received. They might also incorporate survey-based measures of former clients' satisfaction with their jobs after they have been employed for six months, relating more to outcomes. Such customer satisfaction ratings may or may not square with more tangible measures of service quality and program effectiveness, but they do provide a complementary perspective.

Efficiency is typically measured as the ratio of amount of input to the amount of output or outcomes. The inverse of efficiency, the ratio of the amount of output or outcome to the amount of input is called productivity. Therefore, efficacy and productivity convey exactly the same information. Efficiency and productivity have traditionally related costs to outputs. However, to the extent that the performance system provides data on outcomes rather than outputs, it provides a much more complete picture of efficiency and productivity. Focusing on output-to-input ratios carries with it the temptation for managers to increase output at the expense of the quality of services or the end outcomes that are directly connected to the mission or objectives. Here are some examples of outcome based productivity measures; number of people getting jobs within six months after completing a training program per one unit of program cost, number of customers who reported that the service received had significantly helped them per dollar cost of the service, number of clients who 12 months

after completing the service had stopped the risky behaviors targeted by the program per dollar cost of the service, etc. The inverse of the productivity measure becomes the efficiency measure, such as the program cost expanded to train a person who got a job within six months after completing the program, the cost of serving a customer who reported that the service received had significantly helped him, etc.

One cautionary note in calculating productivity or efficiency ratios with output measures is that the outcomes need to be expressed as something to be maximized. Otherwise, it usually does not make much sense in presenting a productivity or efficiency ratio for a program. For example, consider an efficiency ratio for a crime prevention program defined as "cost per reported crime". Though the indicator is very easy to calculate, it does not make any sense as an efficiency indicator for the program. Note that reported crime is the object that we try to minimize not maximize through the program. The outcome to be maximized is the crime prevented not reported so that the efficiency indicator should be defined as "cost per crime prevented". Since reliable data on crimes prevented are never available, we have to resort to an estimate of the number of crime prevented by the program, which requires ad hoc and costly studies.

Efficiency or productivity ratios can be calculated with both outputs and outcomes. However, efficiency or productivity ratios using outputs are common while efficiency or productivity ratios using outcomes are rare. This is partly because few agencies in charge of the programs have developed outcome data due to high cost and technical difficulties in measurement. With the growing interest in outcome-based performance management system at all levels of the government, increased use of outcome-based efficiency or productivity ratios has become more prevalent.

2.3.2. Identification of Performance Indicators

A performance management system is only as good as the performance information it tracks. However, selecting the performance information that should be tracked is essentially a judgment call. Public service agencies almost always have multiple objectives and multiple categories of customers. Thus, program staff selecting performance information should attempt to include all these perspectives, at least to the extent practical.

Various methodologies are available to identify important performance information that should be measured and monitored. We will consider four of them; focus group interviews, meetings with other partners, role-playing by program staff acting as customers, and the logic model.

Interviews with a focus group are an excellent way to identify a program's performance information to track the performance management system. Members of a focus group can be

chosen from lists of customers without regard to the statistical representation of the selection. The information obtained from focus group participants does not provide statistical data, so statistical sampling, though an optional selection method, is not necessary. The main selection criteria are that the participants have experience with the program and be at least somewhat varied in their characteristics. The size of a focus group can vary according to the size and purposes of a program under consideration but is typically limited to 8 to 12 participants to maintain the intensity of the interview. The facilitator chosen by the program agency plays a crucial role in every focus group interview by establishing an open, non-threatening environment to obtain input from each participant and stimulating discussions among participants. The facilitator, however, should be cautious not to directly participate in the discussion or to lead discussion to the direction of his or her own preference. Two kinds of focus groups are particularly useful; focus groups of customers and those of program personnel. Customers are the main target of the program and therefore the natural candidates for an intense interview to identify the performance information, especially outcomes of the program. Focus groups of programs or project personnel, especially those who frequently work in the field with customers, are another useful way to obtain customers' perspectives on performance information.

Focus group interviews are a relatively inexpensive way to identify performance information to measure. Participants do not usually need to be paid, and the meetings, which could be held in various locations within the program's service area, do not need to be in luxurious surroundings. Considerable staff preparation and administrative effort are needed, however, to ensure that the process goes smoothly and produces the information sought.

Many budgetary programs require corporative efforts with, and active participation of other organizations or agencies, public or private. In identifying performance information to track, it is strongly recommended to seek the inputs from those partner agencies and organizations and information that can be gathered through various communication channels such as meetings, telephone and conference calls, mail, faxes, and the internet.

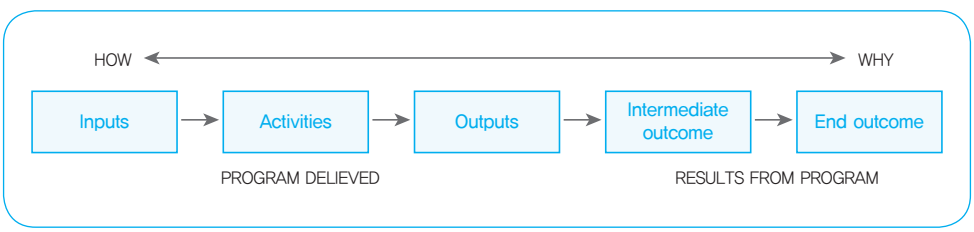
Partnerships are warranted when the program believes that desired outcomes would be best achieved if organizations agree voluntarily on the performance information and performance indicators to be tracked, how the data should be collected, the short and long-term targets for each outcome indicator, and the roles and responsibilities of each organization in providing the particular service. The intricate involvement of all partners can facilitate data collection efforts and possibly reduce the costs of data collection. Such a process is not easy, however, and takes considerably more effort than the traditional go-it-alone approach. Such agreements are called performance partnerships. They are a relatively new concept and require significant time and effort to work out with other organizations.

Role playing in identifying performance information of a program to track is to have program staff take the position of program customers in various hypothetical situations. For example, a child welfare agency may ask its staff members to play roles of program administer, parents and other childcare givers, and children themselves. Staffs who work directly with customers including field workers are more likely to be particularly valuable participants. This procedure is especially useful for programs for which it is not possible to hold customer focus group sessions.

Each participant, in her customer role, should be asked the same questions on which aspects of the programs he likes or not and then draws on her own knowledge of the program and what her experiences have indicated are the likely reactions of customers. In every role playing session, someone should be explicitly appointed as the recorder to take down the findings of the session, especially the potential outcome characteristics identified during the session. The recorder should then draft a report listing all outcomes explicitly or implicitly identified by the role players as either intermediate or end outcomes.

A logic model is a plausible and sensible model of how the program will work under certain environments to achieve the outcomes. Every program has implicit hypotheses about what actions will produce what results. Logic models attempt to identify these hypotheses by showing the flow of intermediate and end outcomes expected to result from program activities and the outputs produced by those activities using inputs. Logical models can be the bases for a convincing story of the program's expected performance, telling stakeholders and others the objectives the program focuses on and how it is effectively qualified to address the task. The elements of the logic model are inputs, activities (or processes), outputs, and outcomes, intermediate and end.

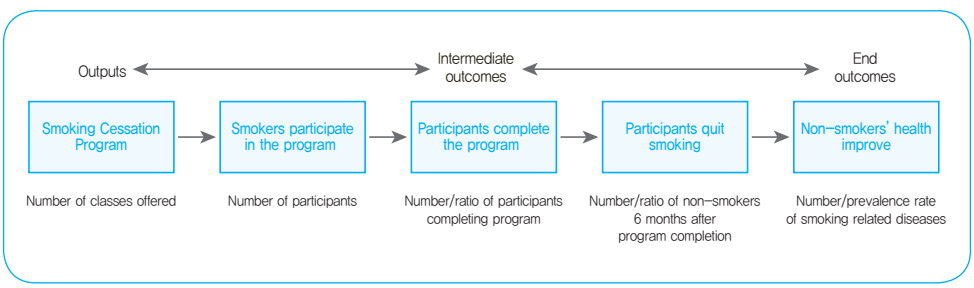
Figure 2-3 | Basic Logic Model



As the emphasis on accountability and managing for results spreads out, there is a growing interest among program managers in logic models. There are several benefits of logic models when they are integrated into the performance management system. First, the logic model is extremely useful conceptual and visual device in identifying core outcomes of programs. Second, it helps with program design or improvement by identifying programs that are crucial to attaining objectives, are redundant, or have inconsistent or implausible linkages to program objectives. Third, one of the uses of the logic model that should not be overlooked is communication. The process of developing a logic model brings people together to build a shared understanding of the program and program performance. The model also helps communicate the program to those outside the program in a concise and compelling way and program staff to gain a common understanding of how the program works and their responsibilities to make it work.

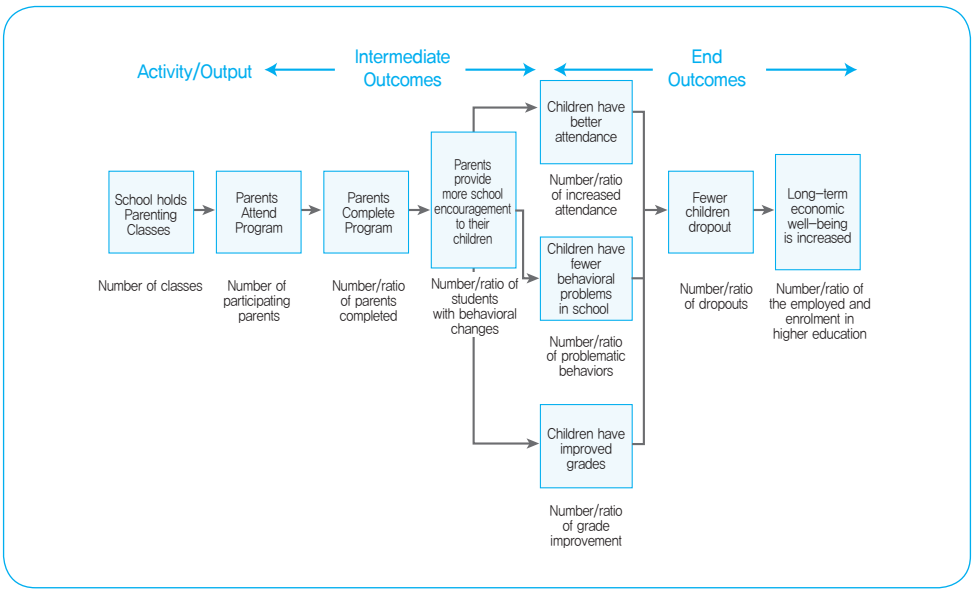
An example may help us understand the use of logic models in identifying performance information to track. Consider a smoking cessation education program to reduce the smoking rate and eventually improve general levels of health. In [Figure 2-4], inputs and activities of the smoking cessation program are omitted for simplicity. In the program, output is the smoking cessation class and measured by the number of classes offered. Smokers will participate in the program and some of them will complete the program, which constitute intermediate outcomes. Intermediate outcome can be easily measured in terms of the number of participants who sign up for the classes and the number or ratio of participants who complete the required classes among total participants. Some of the participants who complete the program succeed in quitting smoking and others do not. In the end, those who succeed in quitting smoking are expected to experience improved health. These end outcomes can be measured by the number or ratio of non-smokers who still manage not to smoke six months after the program completion and the number or prevalence rate of smoking related diseases such as heart disease and lung cancer.

Figure 2-4 | Smoking Cessation Program



The next example clearly shows the usefulness of logic models in tracking performance information in the program with multiple outcomes. [Figure 2-5] illustrates a sequence of expected events for a dropout prevention program focused on parental involvement. The program provides classes to help parents better support their children's learning efforts. Holding classes is the program activity and is an output of that work. The number of classes held is an associated output indicator. The program then hopes that parents will enroll in, and attend, these classes, and then complete the program. The numbers of parents enrolling in and completing the program are intermediate outcomes. These outcomes indicate, respectively, the program's success in attracting parents into the program and retaining them through the end of the program. It is hoped that as a result of the program, parents encourage their children to learn. This outcome indicates that the program actually affected those parents, which is a more advanced outcome, but still an intermediate outcome. While it is expected to lead to improved student learning, this encouragement does not guarantee that improved learning occurred. Increased attendance, fewer behavioral problems, and improved grades of students whose parents completed the program activities are likely to be the specific desired short-term end outcomes. Fewer students dropping out of high school is also hoped for, but this outcome cannot be completely determined until all current students are through their final year of high school. It is, therefore, a longer-term end outcome. The logic goes even further. School learning and completion are hoped to lead to better employment and earning histories of the students-very long-term end outcomes. Each outcome is important and should be included in the program's performance management process, along with the outputs. This includes the regular measurement of outcomes expected to occur far into the future, such as post-high school work histories. The example in [Figure 2-5] considers improved attendance, fewer behavioral problems, and improved grades end outcomes. Others might consider them intermediate outcomes. People can legitimately disagree over outcome categories. The example also considers parents entering and completing the program short-term intermediate outcomes. Others may prefer labeling these actions outputs. However, because these programs are voluntary and involve customers themselves taking steps, we consider these items outcomes. Whether something is classified as an output, intermediate outcome, or end outcome rarely affects the measurement process itself. The label, however, can affect the importance the organization attaches to the outcome. Therefore, program managers should consider carefully the categories assigned to its measurements: whether particular measurements are outputs or outcomes and whether outcomes are intermediate or end outcomes.

Figure 2-5 | Dropout Prevention Program



2.3.3. Selection of Performance Indicators

The performance information obtained from all sources described in the previous section should be assembled into one list and should be categorized as information on inputs, processes, outputs, and outcomes, intermediate or end. The resulting list should be compared with the mission statement to make sure that the list is comprehensive enough to include all aspects of the statement. Duplication or redundancy should be corrected from the list of performance information and any performance information should be deleted from the list when it is not possible or prohibitively costly to measure the information. The next step is to select performance indicators out of various candidates identified through the methodologies discussed above. Performance indicators are not the same as performance information. Each performance information tracked needs to be translated into one or more performance indicators. A performance indicator identifies a specific numerical measurement to represent the progress toward the desired state of performance information. Performance indicators usually begin with the words number of, percent of, ratio of, incidence of, proportion of, or similar phrases.

It is difficult to provide an agreed-upon set of criteria in choosing performance indicators and different agencies offers a different list of criteria. We discuss three of them that have received attention from various agencies, public or private, interested in introducing the

performance management system. The Office of Management and Budget of the United States suggested the following five features for good performance indicators.[3]

- Quality over quantity; performance indicators should be relevant to the core mission of the program and the results the program intends to achieve. However, program designers should not feel compelled to collapse complex activities to a single measure, particularly if that measure is a proxy for the true objective.
- Importance to budget decisions; performance indicators should provide information that helps make budget decisions. Agencies can maintain additional performance indicators to improve the management of the program.
- Public clarity; Performance indicators should be understandable to the users of what is being measured. Publicize (internally and externally) what you are measuring. This also helps program partners understand what is expected from the program.
- Feasibility; performance indicators should be feasible, but not the path of least resistance. Choose performance indicators based on the relevancy of the outcomes and not for other reasons, especially not because you have good data on a less relevant measure, for example. If necessary, terminate less useful data collections to help fund more useful ones.
- Collaboration; agencies and their partners need to work together and not worry about "turf". The outcome is what is important.

Harty (2006) offered a set of characteristics we should focus on in choosing performance indicators.

- Relevance; performance indicators should be relevant to the mission or objectives of the program and to the outcome the indicator is intended to help measure.
- Importance; performance indicators should measure important aspects of performance information.
- Understandability; performance indicators should be easily understood by everybody who uses them.
- Program influence or control over the outcome; do not use this criterion as a way to avoid measuring important outcomes. A program will almost always have less than full influence over most outcomes, especially end outcomes. As long as the program is expected to have some tangible, measurable effect on a specific outcome, and indicator of that outcome should be a candidate for inclusion-whether the effects are direct or indirect.

3. OMB (2003).

- Feasibility; it should be feasible to collect reasonably valid data on the indicator.
- Uniqueness; if an indicator is duplicated by, or overlaps with, other indicators. It becomes less important.
- Manipulability; do not select indicators that program personnel can easily manipulate to their advantage.
- Comprehensiveness; the set of indicators should include outcomes that identify possible negative or detrimental effects.

HM Treasury of the United Kingdom published a list of characteristics that good performance indicators should possess.[4] Performance indicators should be relevant, incentive compatible, attributable, well-defined, timely, reliable, comparable, and verifiable. Relevancy requires that performance indicators should be relevant to what the program is aiming to achieve. It is easy to fall into the trap of targeting easily measured processes. But they often do not address the core objectives of the program. The idea is to find measures that fully capture and represent the objective in question. It is often better to try and measure important objectives imperfectly, than ignore them altogether. An imperfect measure can still help in setting priorities, planning, and providing at least a starting point for further search for a better performance index. Incentive compatibility implies that performance indicators should be able to avoid perverse incentives that encourage unwanted wasteful behavior. A wrong choice of performance indicators may result in the behaviors that exist to meet the indicators, but not to achieve the objectives of the program. An example of a performance indicator with a perverse incentive is "the average time of answering letters", which may result in efforts to respond to letters as fast as possible at the cost of quality of responses. The way data are used may also cause perverse incentives. For example if payment to medical staff is linked to the number of medical treatments provided for the month, then staff may been encouraged to deal with lots of easy treatments rather than a small number of difficult treatments. It should be noted that whether a perverse incentive actually causes changes in behavior can depend on the culture of an organization. Performance measures which focus directly on the objective to be attained avoid creating some of the perverse incentives that can arise when specifying measure around intermediate stages or processes. Measuring outputs or processes in lieu of outcomes, for example, can encourage management to be unresponsive to changed circumstances, or to maintain old processes when new ones could yield better outcomes. Being able to be attributable requires that the activity measured must be capable of being influenced by actions which can be attributed to the program. It should be clear where accountability lies. Performance indicators should measure something that the program can reasonably be expected to

4. HM Treasury (2001).

influence. It should be clear who is responsible for the program's performance against that indicator. Ideally the indicator should also give a clue to the fact that how much of any change can be attributed to the program. Some measures may reflect more than one aspect of a service, and so it may be difficult to attribute the program's role in any change. For example, the number of complaints about a service will reflect both the standard of the service and how willing service users are to complain-which may in turn depend on their confidence that complaints will be taken seriously. An increased number of complaints could reflect an increase in the confidence of customers that the agency in charge of the program will take their complaints seriously. Contextual information may help to shed light on what is driving changes in these sorts of indicators. In some cases the degree to which an agency's activities create the desired outcomes will not be clear. In these cases it is still appropriate to set outcome indicators key to the stakeholders. These indicators help focus an agency on priorities and its overall goal. General indicators may be supplemented by more specific outcome indicators at other levels in an agency. Next, performance indicators should have a clear, unambiguous definition so that data will be collected consistently, and indicators are easy to understand and use. The definition should be easy to understand to the users of the indicator, whether they are internal managers, user groups, or individual users of the program service. The definition of the indicator should be unambiguous. This is important to ensure that data is collected consistently, and to ensure that people have a common understanding of the indicator. There are two elements to a performance indicator providing timely information. The indicator should provide data frequently enough to track changes that are taking place in order to take action. The indicator should provide up to date information, with a short time-lag between the period the data covers and when the data becomes available. The importance of having frequent data, soon after the event, will depend on the speed at which policies can be changed to affect the outcomes and the length of time before the target is due. There is often a trade-off between accuracy and timeliness. Producing statistics quickly means that some data sources cannot be used, and less time can be spent on checking the data. There is also a trade-off between the cost of collecting data, and the frequency of collection. Efforts should be taken to take the balance between those conflicting attributes. Performance indicators should also be reliable, accurate enough for its intended use and responsive to change. All stakeholders must have confidence that any performance information faithfully represents what it purports to represent. A performance indicator should be statistically valid. It should be borne in mind that an indicator based on a very small sample of cases may show large fluctuations. It should also be responsive to change, that is it should pick up significant changes in performance. For example, a measure which relies on a yes/no question for customer satisfaction will fail to register the difference between someone being just satisfied and very satisfied. In order to assess progress we need to be able to make comparisons between current and past performances

or the program's performances and performances of similar programs elsewhere. In order to help ensure the indicator remains comparable, the changes in definition over time should be minimized. Where changes are necessary one should try to estimate their effect on existing measures so that comparisons can be made between periods before and after the change. It is also important to help ensure the comparability of the indicator to use standard definitions where these exist. For example, other agencies in charge of similar programs may be measuring similar concepts, or international definitions may exist which allow comparison between countries. Verifiability requires that the performance indicator should have clear documentation behind it so that the processes which produce the indicator can be validated.

2.3.4. Difficult-to-Measure Performance

Not everything that counts can be counted. Some performance information may require indicators that are extremely difficult or expensive to track directly, in which case surrogate indicators that are easier and less expensive to measure are used. Here are some examples of activities that are particularly difficult to measure.

First of all, the performance of programs related to deterrence or prevention of unwanted situations or behavior is very difficult to measure or quantify. These include programs such as crime and fire prevention, child abuse prevention, and disease prevention. Regulatory programs such as environmental protections and food safety programs also face the same difficulties in terms of measurement of performance information. These programs are ultimately intended to prevent a variety of unwanted and hazardous behaviors or situations. Most importantly, prevention or deterrence measurement requires consideration of what would happen in the absence of the deterrence program. Also, it is often difficult to isolate the impact of the individual program on a situation or behavior that may be affected by multiple other factors. Accurate measurement of counterfactual incidents requires very sophisticated and expansive schemes that provide some way to measure what would have happened in the absence of the program. Such high cost is a great disadvantage for the performance management system in which one should measure performance information in a regular manner.

Less sophisticated and less expensive alternatives are needed. An easy and quick solution is to use the number of incidents that were not prevented as a surrogate for cases prevented. The indicator, however, is valid only if we are willing to accept the presumption that the total number of incidents is very stable over time. In addition, surrogates can be found that track reduction in major factors known to lead to undesirable incidents. Those factors can be named as risk factors that, if reduced, are expected to help prevent the unwanted incidents. Even though the indicators tracking risk factors are important, they are intermediate not end outcome indicators.

If performance measures reflect a continuum from lower-level outputs to higher-level outcome measures related to the overall strategic goal, it is important for deterrence or prevention programs to choose measures that are far enough along the continuum that they tie to the ultimate strategic goal, as well as to the program's activity. This will help ensure that the measures are both meaningful and genuinely affected by the program. Care should be taken, as some measures may create perverse incentives if they do not reach the correct balance between output and outcome. For some programs, deterring a majority of the negative outcome is appropriate. For other programs, most, if not all, of the negative outcome must be avoided. In principle, the target for the program should reflect consideration of the maximization of net benefits. In any event, understanding the costs and benefits of compliance at the margins will help the program to determine the correct target level for compliance. For programs where failure to prevent a negative outcome would be catastrophic including programs to prevent terrorism or nuclear accidents, traditional outcome measurements might lead to an "all-or-nothing" goal. As long as the negative outcome is prevented, the program might be considered successful, regardless of the costs incurred in prevention or any close calls experienced that could have led to a catastrophic failure.

The next category of programs whose performance is difficult to measure is the ones with long range activities. Some programs take many years to produce outcomes. As the time span until major outcomes are realized is widened, the usefulness of performance indicators declines and more factors are likely to have intervened, which makes the task of relating outcomes to program activities very difficult. Program agencies can track several intermediate outcomes that are realized in an earlier stage than the end outcomes. Examples include such intermediate indicators as the percent of time that reports and plans were provided on schedule, the results of peer reviews, and the number of citations in the technical or academic literature. These indicators, however, tells us little about the results of the program. In-depth analysis like program evaluation is a more adequate tool to assess the end outcome of the program than program monitoring that requires frequent and regular measurements of program performances.

It may also be useful to track process-oriented measures, such as the extent to which programs make decisions based on competitive review. For example, research programs can have many uncertainties, including their expected outcomes. So, while research programs are encouraged to define measures that can track progress, not all of them will be able to. Such programs may rely, in part, on process measures, such as the extent to which the program uses merit-based competitive review in making awards. To qualitatively address the research itself, some programs develop measures to reflect meaningful external validation of the quality and value of the program's research. To address the uncertainty

of research outcomes, programs may also be able to demonstrate performance in terms of the broad portfolio of the efforts within the program. Expert independent evaluators might also help determine if the process of choosing appropriate long-term investments is fair, open and promises higher expected payoffs in exchange for higher levels of risk. Rotating evaluators periodically may help ensure independence and objectivity.

It is also difficult to measure the performance of the programs whose major outcomes apply to a very small number of events. For some programs, the results of a small number of particularly important events may have significance far beyond statistical incidence. For example, an emergency response program's performance cannot be adequately assessed with traditional quantitative indicators such as the number of lives saved if only a few major emergencies occur during a reporting period. In these instances, quantitative data are not sufficient and a pragmatic alternative is to provide qualitative information separately for these few but very important cases.

Often programs from various levels of government, federal, state, and local, private-sector or non-profit activities, or even foreign countries all contribute to achieving the same goal. The contribution of any one program may be relatively small or large. In those cases, it is difficult to separately measure performance of each individual program agency. Examples of programs with these characteristics include international peacekeeping programs, special education pre-school grants programs, highway maintenance programs, vocational education programs, and many education, labor, and housing formula grant programs. One approach to this situation is to develop broad, yet measurable, outcome goals for the collection of programs, while also having program-specific performance goals. For a collection of programs housed primarily in one federal agency, a broad outcome measure may be one of the goals in an agency strategic plan. The broad outcome goal can often be tracked using national data that is already being collected, while the program-specific goals may require more targeted data collection. It is important to "right size" the measure to suit the program. Sometimes a program is such a significant contributor, or leverages so much funding, that an appropriate goal is a societal outcome. Other times it is more appropriate to write measures specific to program beneficiaries. There is no rule of thumb on where that threshold is. It is only suggested that programs of similar size, or with a similar percentage contribution to the desired outcome, approach this issue similarly. Sometimes programs are designed to work together toward a common goal, but each provides a different piece of the service or activity. In other cases, programs are designed to merge funds and support the same activities as well as goals. When programs fund different activities and do not co-mingle funds, programs should be able to develop activity-specific performance goals that support the broader outcome. It is likely, however, that these will be output goals and the challenge will be agreeing on how each of the separate activities contributes to the outcome.

When programs co-mingle funds in support of a goal, it is extremely difficult to assess the marginal impact of the program dollar since all funding supports similar activities. Programs may seek to claim responsibility for the entire outcome and output, despite having a shared, and sometimes small, role in the overall activity. However, we should seek to evaluate whether such claims are realistic.

Some programs are designed to address multiple objectives or support a broad range of activities or both. Block grant programs often have these characteristics, with the added feature of allowing grantees the flexibility to set priorities and make spending choices. Increased flexibility at the local level can limit efforts to set national goals and standards or create obstacles for ensuring accountability. In other cases, the program may focus on a limited set of activities which in turn are used for multiple purposes by many distinct stakeholders. Establishing performance indicators for these types of programs can be challenging. Moreover, some block grant programs provide resources to non-federal levels of government to focus on specific program areas, such as education, job training, or violence prevention. While the funds can often be used for a variety of activities, they are for a specific purpose. In these cases, national goals can be articulated that focus on outcomes to highlight for grantees the ultimate purpose of program funds. Targets for these measures may be set by surveying grantees to gauge the expected scale of their work or by looking at historical trend data. A system could be developed that uses performance measures and national standards to promote joint accountability for results. With this approach, after agreeing on an appropriate set of performance measures, program targets can be set at the local level and aggregated up to national targets.

Many programs in the government are administrative or process-oriented in nature, which presents a number of problems when it comes to measuring performance. One issue is the appropriate balance between outputs and outcomes. Realistically, output measures may be useful for evaluating the efficiency of process oriented activities. However, for larger administrative efforts, consideration should still be given to ultimate outcomes. In some cases, it may make most sense to evaluate the administrative costs as part of the overall program, rather than as a separate activity. For example, a grant program may contain separate accounts for the grants themselves and for administrative salaries and expenses, yet both accounts might be viewed as providing inputs into a single program. Benchmarking with other agencies or the private sector, competitive sourcing, and the use of intermediate outcomes such as returns on investment are all approaches that can assist where data availability is an issue. As many administrative functions run across agencies, the development of common measures is also encouraged.

3. Performance Evaluation

Equipped with performance indicators and related information, one can proceed to the next step in the performance management system; that is performance evaluation. There are several available evaluation techniques. We discuss two most popular evaluation procedures; performance monitoring and program evaluation. As briefly discussed, performance monitoring and program evaluation are two complementary not competing tools and a good performance management system should have both of the procedures firmly built in it.

3.1. Performance Monitoring

Performance monitoring, also called performance measurement is a descriptive evaluation procedure that complements, informs, and supports more methodologically rigorous evaluation studies such as program evaluation. Although it is obviously more superficial than an intensive investigation such as the program evaluation, it can often provide a timely and inexpensive view of program or agency performance on an ongoing basis. Thus, governments that are interested in managing for results and improving their performance are increasingly using performance monitoring systems.

Performance monitoring systems are designed to track selected measurements of programs called performance indicators at regular time intervals and report them to managers and other specified audiences on an ongoing basis. Their purpose is to provide objective information to program managers and policy makers in an effort to improve decision making and thereby strengthen performance, as well as to provide accountability to a range of stakeholders, such as higher-level management, central executive agencies, governing bodies, funding agencies, accrediting organizations, clients and customers, advocacy groups, and the public at large. Thus, performance monitoring systems are critical elements in a variety of approaches to performance-oriented managements.

The core tasks in performance monitoring are to select performance indicators and to compare measured levels of performance indicators with target or expected levels. We have already discussed the selection of performance indicators in the previous section. The next step is to measure program performance in terms of the selected performance indicators.

3.1.1. Classification of Data

Data can be classified in a variety of ways. First, data can be classified into quantitative and qualitative data. Quantitative data are measured in terms of numerical values such a crime rate, gross domestic products. Qualitative data are measured in terms of categories such as satisfaction level and ethnic background. Next, data can be classified into subjective and objective data. Subjective data describe attitude, feeling, perception at individual or

group levels. Objective data are based on facts, excluding subjective judgment or evaluation. Both subjective and objective data could be either quantitative or qualitative. Third, data also can be classified into cross-sectional and time series data according to the dimension of measurement. Time series data record measurements on the same object along a time line while cross sectional data measure performances of many individuals or groups at one point of time. Finally, data can be classified into primary and secondary data according to sources. The primary data are collected by the researchers and the secondary ones by others.

3.1.2. Data Sources

a. Agency Records

Most agencies in charge of programs routinely record data on customers, activities, and/or transactions for administrative purposes. Acquiring data that already exist and are maintained by agencies is the most widely used for producing performance data. Records may come from the program itself, from other programs within the agency, or from other agencies. The relevant information for performance monitoring should be extracted from those records in order to yield the desired performance indicators. Agency records are particularly a useful source on input and output information as well as outcome information. Records also, in most cases, contain demographic and socioeconomic variables of program participants and other characteristics of the workload for breaking out indicators for further in-depth analysis of performance indicators.

Agency records have advantages. First, the data are readily available at low cost. Second, most program personnel are familiar with the procedures for transforming the agency data into performance indicators since they are responsible for collecting and maintaining the data. Agency records also have disadvantages. First, agency records seldom contain enough information on quality and outcome data to create an adequate set of performance indicators. Second, existing record collection practices often need modification to generate performance indicators. Third, obtaining data from other programs or agencies can be administratively difficult and can raise the important issue of confidentiality. For example, indicators of the success rate of health and social service treatment programs may take the rate of recidivism as the core performance indicator. Such data might require access to data archives of police, hospitals, courts, and other health and social service agencies where failed customers end up, which raises very difficult and delicate issues on privacy and jurisdiction.

b. Statistics from Outside Sources

Statistics from an outside source can be utilized to construct performance indicators. Outside statistics are particularly important as a data source since they may offer an

opportunity to assess the effectiveness of a program by comparing data for participants (treatment group) and non-participants (control group). Types of available outside statistics include government statistics and privately published statistics. The government collects data and publishes various data for administrative or policy purposes. Official statistics from the government agencies generally provide data that are relatively reliable and consistently collected. The government agencies tend to collect certain statistics on a more regular basis than many private sources. Private organizations, such as trade associations and advocacy groups, collect data that may be valuable to an organization's performance management or simply to promote their activities. Special care should be taken when data from a private source are used since these kinds of data are prone to interrupted collection, irregular methods, non-uniformity, and uncontrollable bias. The careful performance monitoring professional will only use data that conforms to the researcher's needs and will specify data limitations or seek to apply multiple lines of evaluation methods when any data is in doubt.

The biggest advantage of outside statistics as a data source for the performance management system is that many useful statistics can be collected with little or no assistance from program participants. Therefore, it is relatively an inexpensive and quick way of collecting data. However, outside statistics also have some disadvantages. Available data may not exactly measure the desired characteristics, and the proxy measures force evaluators to either create an additional measure or simply accept the measure. The bias that may exists in privately collected data could influence the performance monitoring procedure in unintended ways. Moreover, since statistics themselves, in most cases, offer little evidence for program performance, especially outcomes since most of the measures were originally conceived for other purposes that they may not measure the phenomenon to the desired level of detail and accuracy.

c. Customer Surveys

Customers are an important source of information for the performance management system and a major way to acquire reliable information from customers is through professionally designed surveys. Information obtainable from a customer survey includes the customer's condition and attitude after participating in the program, as well as the results of the programs, customer action or behavior after participating in the program, overall satisfaction with the programs, ratings of specific service quality characteristics the programs provide, extent of service use and awareness of services the programs offer, suggestions for improving program performance, and demographic and socioeconomic information on the respondents.

Surveys are often the most feasible, if not the only, way to acquire data for some performance indicators and provide direct input from the program's customers, adding not

only valuable information but also credibility. In addition, surveys offer the opportunity to involve stakeholders in both the program assessment and the program improvement processes. A well planned and conducted survey presents an efficient method of collecting personal information and perceptions of individuals impacted by a program. Surveys may provide a leading indicator of what changes may become necessary in the near future. On the other hand, surveys require special expertise, especially for the development of the questionnaire and sampling plan and for training of interviewers. Surveys also require more time and are more costly than other forms of data sources. Moreover, evidence based on respondents' perceptions and memory may be less convincing than other data sources that are based on objective assessments. Some customers may not respond or will not be honest in their responses. This problem can be alleviated by well-worded questions and good interviewing skills but it is not impossible to eliminate potential problems due to non- and dishonest responses.

How a survey is designed and conducted can be critical to the success of the survey and to the collection of a body of data that is of adequate size and quality for the intended purpose. Good surveys can be developed by professionals with abundant experiences in the field and extensive knowledge of the program but the following procedure may help optimize most survey designs;

- Define the areas for evaluation and develop applicable questions.
- Identify the target population and designate a comparison group, if applicable.
- Develop a sampling protocol that includes a well thought out method of data collection, sampling techniques and method of analysis.
- Develop the questionnaire.
- Field-test the questionnaire, the individual questions, and the time it takes to administer the test.
- Distribute the questionnaire to respondents with a stated return date.
- Provide a follow-up contact with non-respondents, if the sample size is small enough to be able to track non-respondents individually.
- Analyze data and share the results with stakeholders.
- Report the results.

Designing the survey questionnaire is the most important element to extract high quality information from the customer survey. A well-designed survey questionnaire may include questions that can offer information on the important topics such as outcomes of the program, the type and amount of the services used, reasons why respondents gave particular answers

or ratings, especially unfavorable ones, suggestions on improving the services program provides, and demographic and socioeconomic variables for further in-depth analysis.

There are several issues that should be resolved in designing customer surveys; the coverage of survey subjects, survey administration methods, and the size of survey, to name just a few.

First, the program agency needs to decide whether to survey the entire population in its jurisdiction, called household surveys, or to survey its own customers or program participants, called user surveys, or both. Household surveys include representative samples of all potential customers in the jurisdiction regardless of whether they have used the service provided by the program under consideration. Household surveys can produce information on several important services simultaneously and hence costs of the survey can be shared among agencies, reducing the costs to each agency. Also, household survey can produce information from non-participants, enabling the program agency to estimate the determinants of program participation and to construct the comparison group that is very important in evaluating "true" program effectiveness. Since household surveys are administered centrally, they are likely to offer opportunities for better quality control and are less of a burden on individual agencies. On the other hand, user surveys are administered to customers who have actually participated in the program and used the services provided. User surveys usually provide more in-depth information on particular services because survey subjects are familiar with them and do not need to be asked about other services. Sample selection is easier for user surveys because the program agency usually keeps records on program participants' contact information while samples for household surveys need to be drawn from some census data. For user surveys, higher response rates are expected because users have personal interest in and knowledge of the service and the information collected is likely to be more useful to program personnel due to the extensive and detailed nature of the information.

The next issue we need to address is the methods through which the survey is administered. Customer surveys, household or user, can be administered through mail, telephone, in-person, the internet, or a combination of those aforementioned. Mail survey is a low-cost method, including second and third mailings possibly combined with telephone follow-ups for non-respondents to secure response rates high enough to yield reliable information. To obtain satisfactory completion rates, mail questionnaires should be short and simple, preferably not exceeding four or five pages. Telephone surveys can achieve good response rates at lower cost than the in-person alternative. However, they are more expansive than mail surveys mainly due to higher costs required to compensate for interviewer time and training. Telephone surveys have faced a serious problem of ever increasing rates of non-responses due to people's greater resistance to phone surveys. The increased use of phones

in sales promotion by private companies and opinion polls by political entities has induced people to resist to revealing their preferences or opinions to interviewers whose identity they cannot confirm. In-person surveys can be administered either at the respondent's home or business or at a service facility depending on the nature of services provided by the program. The former tend to yield high response rates and can provide detailed information since they can be longer and more complicated than mail questionnaires. The fundamental difficulty with the survey methods is that they are the most costly to administer and too expensive for repeatedly collecting the data needed for the performance management system. The latter has the advantage of obtaining high response rates without the high cost of finding respondents at their homes or businesses. This option is a good one if the primary purpose of the survey is to capture the quality of the participants' immediate assessments of the services provided by the program. It is, however, impossible to use for assessing the program outcomes that occur after the customers leave the facility. Internet survey is an emerging option gaining growing popularity in performance management for low cost and widespread access to computers. Internet surveys should be complemented with advance and follow-up mails to encourage the completion of the questionnaire. A good strategy in choosing the administrative method is to use a combination of various methods, if possible. For example, mailings can be supplemented with telephone calls to people who failed to respond to follow-up mailings, mail or telephone surveys might be used to supplement internet surveys, and so on. Combining different survey methods may raise questions of accuracy or consistency of responses from more than one method. But, it is not clear whether this significantly distorts the survey findings. In choosing a method of survey administration, the program agency needs to weigh the trade-offs between response rates, larger sample sizes, and cost of administration. Higher response rates provide greater confidence that the findings represent the views of the population surveyed and larger sample sizes generally yield estimates with higher statistical reliability.

For some programs, it may be feasible to survey all customers, such as by routinely mailing questionnaires to all customers. This is generally possible only when the number of customers participating in the program is small enough for the program agency to be able to administer the survey with manageable cost and in a reasonably short amount of time. When the size of the population of interest is very large, survey administrators are forced to take samples from the population. The crucial issue is how large the sample size should be. Larger samples are needed if the program agency wants to measure performance indicators with high precision in spite of higher administrative costs. However, it is not the usual practice to pursue measurements with very high precision in customer surveys.

One of the biggest obstacles of the customer survey is costs. Surveys costs depend on a variety of factors; the number of surveyed people, the frequency of the survey, the mode

of administration, and efforts to increase response rates. Several suggestions are offered to save survey costs. First, if the population represented is large, it is best to construct a sample rather than surveying the entire population. Second, one should try to reduce the required sample size by not pursuing a measurement with excessive precision. Third, it is a good idea to use agency personnel or volunteers when possible and appropriate. Fourth, it is possible to cut costs significantly by using or adopting questionnaires that are already available especially from other agencies and shorten the questionnaires as much as possible. Fifth, adding questions to already scheduled surveys, such as those of other agencies or universities, might be another useful way to save survey cost.

d. Trained Observer Ratings

Trained observer ratings are used to extract information related to performance indicators through direct observation by observers with enough expertise in the area to which the program of interest belongs. The key element for performance management is that the rating scales and procedures should be robust enough to provide values and reliability. The challenge is to ensure that different observers at different times assign the same or similar evaluation results to similar conditions. In order to secure the reliability of the method, we need systematic rating scales, adequate training and supervision of the observers and the process, and periodic examination of the quality of the ratings. Trained observer ratings have been used for assessing quality of services such as cleanliness of streets, condition of facilities, condition of traffic signals and signs, and quality of food served in shelters for the poor. Trained observer ratings are used for human services as well. Agencies have rated the ability of customers with physical and mental problems through observations and examinations and teachers have rated children's readiness-to-learn by observing the behavior of children. Trained observer ratings can provide reliable and reasonably accurate assessment of conditions that are otherwise very difficult to measure. Assessment of performance by trained observers can provide a timely and inexpensive source of performance information since it can be conducted with little preparation as long as the program agency maintains a list of experts or assessors with adequate training. In particular, periodic ratings by trained observers can provide information that can be used not only for periodic reporting of performance indicators but also to encourage early responses to problematic conditions. However, ratings by trained observers are very labor-intensive procedures that take a significant amount of personnel time and resources. Moreover, the ratings should not be done by the people in charge of administrating the program or delivering the services since the results from such ratings, in most cases, will lack credibility especially from external stakeholders.

Implementing a trained observer rating session typically requires the following steps;

- Decide what should be rated.
- Develop a rating scale for each assessment.
- Determine which facilities re areas should be rated, when, and how frequently.
- Select observers with adequate qualification and train them for rating sessions.
- Test the scale and observers on a small number of sites in the facility or area to make sure reasonably well trained observers give consistent ratings.
- Establish procedures for supervising the observers and for recording, transcribing, and processing the data collected.
- Conduct the ratings regularly.
- Establish procedures for systematically checking the ratings of trained observers to evaluate the quality of assessment.
- Develop and disseminate reports on the findings.

e. Expert Panel Evaluation

Expert panel evaluation, also called peer review, involves the reviewing of one's work by those with expertise in the field. Expert panel evaluation is premised upon the assumption that a judgment about certain aspects of science, for example its quality, is an expert decision capable of being made only by those who are sufficiently knowledgeable about the cognitive development of the field, its research agenda, and the practitioners within it. The method has been used in many areas in which a significant degree of expertise is required to assess the achievement of the program and research and development programs is one of the major fields in which the expert panel evaluation technique have been intensely utilized.

Three approaches to the designation of peers have been developed. It was posited that peers should, whenever possible, include members of the applicable "invisible college" who study the program or area to be studied. These potential investigators may be those within an organization conducting the review or those knowledgeable professionals working in the field. A different approach would be to use those evaluated to evaluate their own work, although this method remains open to criticisms of its objectivity. A more recent approach is to use stakeholders to evaluate the program or work.

Expert panel evaluations may offer a way of evaluating very complex matters or to provide an insightful ranking of technical alternatives. They also can be used to evaluate projects that are not near a mature stage or projects that may produce immeasurable outputs and may quickly accumulate expert opinion or advice for use in developing an evaluation framework. However, we should point out a serious disadvantage or limitation to the method.

Experts commissioned to evaluate the program may possess biases that may preclude its use in many cases or at least erode the credibility of the evaluation results. It is crucial to find a group of experts who possess an objective attitude and are free from collective interests but still have enough expertise.

f. Use of Special Technical Equipment

Special equipment is needed for some programs that require scientific measurement to collect data such a noise levels, air or water pollution levels, or road conditions. Appropriate technical equipment usually provides accurate, reliable data and in some cases is the only reasonable way to acquire credible information on performance indicators. The equipment can be expensive to procure, operate and maintain, which may pose a serious obstacle in developing countries. The information obtained from the equipment should be interpreted to be useful to program personnel and outside stakeholders, some of whom may experience significant trouble in the process.

g. Focus Groups

Focus Groups are small, group-facilitated sessions, designed to quickly gather in-depth information while offering stakeholders a forum for direct participation. They are usually facilitated by an outside third party and can yield invaluable information. Focus groups can be used to gather a wide variety of information; to obtain general background information about a topic of interest, to generate research hypotheses that can be submitted to further research and testing using more quantitative approaches to stimulate new ideas and creative concepts, to diagnose the potential for problems with a new program, service, or product, to generate impressions of products, programs, services, institutions to learn how respondents talk about their interests to facilitate the design of questionnaires, survey instruments and other research tools, to interpret previously obtained quantitative results, to provide a forum for stakeholders to present their views and participate in the process. Focus groups enable performance monitoring to collect data from a group of people quickly and at a lower cost than individual interviews and to interact directly with respondents allowing follow-up questions, clarification of responses, contingent answers, and observation of nonverbal responses. Moreover, focus groups allow respondents to build on responses of other group members and give flexibility to examine a wide range of topics and subjects.

To conduct a successful focus group, that is, to optimize the amount of information that can be gathered from a focus group, the following procedure is typically followed;

- Define the problem and formulate the research question. An internal focus group composed of members of the same organization promotes interaction between different levels of power and measures the group opinion on a topic. An external

focus group is composed of participants from inside and outside an organization and can bridge the gap between professionals and the target audience.

- Identify the sample population for the study. When using a single focus group, the size of the group should range from eight to 12 participants and should represent a cross section of opinions. In a larger study, divide the participants into groups of similar opinion or organizational level, using group data to form a composite picture.
- Carefully choose a facilitator. A good facilitator is well trained in group dynamics, familiar with a variety of interview techniques, genuinely interested in hearing other's thoughts, and is expressive of his or her own feelings and biases.
- Generate and pre-test the interview guide.
- Recruit the sample. Although meal and transportation costs are often sufficient, financial incentives may help induce active participation.
- Conduct the focus group session. Facilities geared toward focus groups often including the use of one-way mirrors, sound equipment, and video recording devices can dramatically enhance the data recorded.
- Analyze and interpret the data. Use transcripts and recordings to recreate the discussion and use content analysis and other standard data analysis techniques to glean the greatest amount of information from the material provided by the participants.
- Write the report.

h. Other Issues in Data Collection

There are several criteria in selecting how to collect data. Five of them are particularly important that each of them merits detailed discussion; cost, feasibility, accuracy, understandability, and credibility. Cost is always the primary concern in selecting data collection method. There exists a fundamental tension between cost and accuracy or adequacy of data collected. Feasibility criterion covers identification of non-financial obstacles that are likely to make data collection very difficult or impossible. The level of accuracy and reliability that can be achieved with each procedure should also be one important criterion in selecting the data collection procedure. Data collected by a procedure should be understandable to program managers and outside stakeholders as well as the general public. Credibility criterion includes the potential for data manipulation especially by people with a vested interest in making the data look good. For cost reasons, many agencies tend to resort to internal resources to obtain data, which may raise the issue of credibility in a serious manner. Such a problem can be alleviated if an agency uses agency personnel who are not involved in delivering the services and has a reasonably robust data control process.

It is generally a good idea that all new or substantially modified data collection procedures should be put under a pilot-test to identify and eliminate possible problems before full implementation. The pilot test should approximate the conditions that are expected to appear after full implementation, but it will usually test the procedures only on some segments of the program.

Program agencies should pay enough attention to the frequency of data collection and reporting. Frequent collection and reporting is important for making the data useful to operating managers. However, higher frequency typically leads to higher cost, which may cause problems for program agencies working under tight budget constraints.

3.1.3. Data Analysis

Producing data does not mean that they will be useful and additional efforts should be taken in the performance management system to help transform collected data into useful information such as performance indicators. There are several alternatives among which a performance analyst can choose but two of them, breakouts of performance information for each indicator and comparison of a program's performance information to benchmark data, have drawn much attention.

a. Breakouts of the Sample

Breakouts of the sample according to distinguishing attributes can reveal highly useful information on performance hidden by aggregation. Two functions of breakouts are particularly important. First, breakouts can clearly distinguish differences in performance among relevant subgroups. Identifying such differences is important since it is the first step to examine why different levels of performance are observed across subgroups. Second, breakouts can help identify why some groups have significantly better performance than others, if they exists. There are several criteria suggested to break the sample into different subgroups; organizational unit or project, workload or customer characteristics, geographical location, difficulty of workload, type and amount of service provided, and reasons for outcome or rating.

When a program is administered by multiple organizations or a program sponsors different projects perhaps using different grantees or contractors, providing performance information for each organization or project will greatly increase the usefulness of the information. Examples of organizational units for which breakouts are likely to be relevant include individual facilities, particular offices that are each the responsibility of a specific supervisor or manager, and groupings of offices or facilities that are an upper-level manager's responsibility.

Providing performance data on every organizational unit or project level can be expensive. For such cases the program needs to make a choice between being satisfied with less than complete coverage of all individual units or projects and use smaller sample sizes that will result in less precise estimates.

Breakouts by categories of customers or other forms of a program workload are likely to be useful in providing information about whether particular categories are achieving the expected performance. Such breakouts also provide important information on the distribution of benefits across customers. Programs with individuals as the basic unit of treatment may consider breaking the sample into several categories according to age, gender, income, household composition, and amount of formal education, etc. Programs whose basic treatment unit is business may categorize based on the criteria such as business size, product type, ownership characteristics, and location of the main office. For some programs, categorization of the sample can be done based on the types of workload rather than customers. For instance, for education programs, breakouts can be achieved based on school characteristics such as the number of student enrollment, the number of teachers, location, and percent of student body under the program.

Data can be broken out according to geographical location and this is one of the most popular ways of breakouts. Information on the geographical distribution of quality or extent of services provided by the program can be used to identify problematic areas and come up with remedies. Useful geographical breakouts may include neighborhood, political boundaries, zip codes, and regions.

All programs receive workloads with different difficulties, which could be another criterion for data breakouts. The proportion and degree of difficulty are likely to vary depending on reporting time and organizational unit. The more difficult the workload, the more time-consuming, and the more expensive, it is to achieve the desired performance. For example, some students are more difficult to help than others, perhaps because of limited mental or physical ability. Providing performance information at aggregate level can be very misleading and unfair if the difficulty of the workload is not considered. A program unit may exhibit a poor performance not because it did a poor job but because it happened to deal with more difficult cases than other units.

Program performance can also be grouped by the amount or type of service if multiple procedures are used to deliver the service. This kind of breakouts provides program agencies with information to help them assess which service delivery approaches are accomplishing desired performance and which are not. If the agency classifies each item of workload by particular type, or amount of service applied to the item and then links that information to the performance data on the item, the program agency can subsequently obtain and compare performance information for each type or amount of service.

If information is available on reasons for outcomes, particularly reasons for poor outcomes, breakouts by reason can be used to infer actions needed to correct the problems and achieve the desired outcomes. For example, a disease control program may have shown unsatisfactory performance because patients participating in the program did not follow the regimen prescribed, medical staffs made errors, or the disease was essentially very difficult to cure. To be most useful, reasons should be grouped into those reasons the program can do something about and those it cannot. For instance, if a significant proportion of respondents identify the lack of time for nonuse of the services provided by the program, the program may be able to alleviate the problem and subsequently increase usage by modifying the way or the time services are delivered.

b. Comparison with Benchmarks

Once a program has collected performance information, the program agency should determine whether the level of performance reflected in the data is satisfactory or not. Comparing the outcomes to benchmarks is a fundamental and essential element of the performance management system as well as the performance monitoring process. Many types of benchmarks to which performance information are compared have been suggested. Examples include performance in the previous period, performance of similar organizational units or geographical areas, performance for different workload or customer groups, different service delivery practice, a recognized general standard, performance of other jurisdictions, performance of the private sector, and targets established at the beginning of the performance period.

Comparison with performance in the previous period is the most common type of comparison and is applicable to almost all programs. It conveys the information on whether the program performed better in a given service environment over time. Data on past performance should be readily available, except for the first-time performance indicators. Reporting periods compared should be the same length such as month or year. The frequency of comparison and the duration of comparison period are two important issues that should be resolved before the comparison is actually performed. As for frequency of comparisons, the fundamental principle is that the more timely the feedback from the analysis of performance information, the more useful it is for agencies and staffs in charge of the programs. However, the resource constraints program agencies are facing should also be considered in determining the frequency of reports. Time length for comparison could be year, quarter, or even month depending on the nature of the performance information. Once time length shorter than a year is chosen, programs whose outcomes are believed to be significantly affected by seasonal factors should compare data for a particular quarter or month with data for the same quarter or month in previous years. To track the changes in performance in the initial stage of a new program practice, data for several periods after the

introduction of the new program practice should be compared to data for several periods before their introduction.

Performance data can be compared with those of similar organizational units or geographical areas. An important use of this type of comparison is to enable comparisons among different organizational units or geographical areas that provide essentially the same service to essentially the same type of customers. Through the comparison, we can identify which units or areas are performing well and which are not relative to one another. In addition, the information from the comparison can be used to motivate program personnel in each unit and be a source of best practice information. The crucial requirement for the valid comparison is that the type of customers as well as services delivered should be reasonably similar across units or areas compared.

When workload and customer breakouts of performance data are available, categories can be compared so that managers can focus on the groups that need special actions. That is, comparisons indicate whether the program is more or less successful on particular performance indicators with certain categories of customers or workload than with others.

Programs from time to time adopt new or different practices to deliver the services in terms of operating procedures, staffing arrangements, amounts or levels of service provided to individual customers, and providers. The performance information from different delivery practices help program agencies assess the performance of different ways of delivering services and examine the performance of the current practice for improvement. Using performance information to compare alternative delivery practices with the current ones, program agencies may choose one out of two alternative approaches. The first option is to introduce new practices across the board to replace the current practice and the other one is to adopt new practices into part of an operation and running the old and new practices simultaneously for a period of time. For the former, performance information for a period before the change should be compared with that for a period after the change while performance information for the old practice should be compared with that for the new practice over the same period for the latter.

When an external entity such as another level of government or a professional association has developed a standard for performance indicators, the standard can be used to assess the performance of the program. Aside from those formal standards, another type of standard useful in the performance monitoring process is the “rule of thumb”. For example, it is not clear what level of customer satisfaction should be considered acceptable in a customer survey; is it bad if 10% of the respondents rate the service as fair or poor rather than good or excellent? There is no widely accepted standard to judge those kinds of qualitative assessments. However, well-experienced experts may offer a rule-of-thumb standard to infer

comparable information from data on quantitative assessment. In the previous example, a rule-of-thumb standard may suggest that if more than 20% of respondents give ratings lower than good, the issue warrants attention. One cautionary note is that rule-of-thumb standards have to be considered in light of the particular service being rated. For example, if past records indicate that customers have been unhappy with a service, the agency might want to raise the bar for the level of satisfactory performance.

In some cases, comparable performance information may be available from other jurisdictions for comparison with the performance data collected. This type of comparison is useful as long as the activities of the other jurisdictions are similar to those of the program under evaluation and compatible data on the performance information are available. Comparison with performance of similar programs has the advantages of indicating the realistic target level of performance indicators and identifying good practices of high-performing programs that can be adapted. The crucial condition for the usefulness of comparison is that the program agencies should be able to find the programs in other jurisdictions that showed good performances under similar circumstances to their own. Some private businesses provide services comparable to publicly provided services and therefore their performances can be a good benchmark for programs operated by the public sector agencies. Examples include private bus companies, solid waste collection firms, and vehicle maintenance shops.

A usual practice in the performance management system is to establish a target level at the beginning of the performance period and compare it with the actual level of achievement. Most countries conduct performance monitoring by comparing the actual performance with the target level. Target levels should be challenging as well as realistic. Experience with an indicator is important in setting up targets. If a performance indicator is new so that the program agency has not collected enough data to be confident of setting plausible values, it is wise to defer the target setting. In most cases, annual targets are likely to be required for an agency's budget preparation process. However, programs will typically benefit from setting up targets for the reporting period shorter than a year. On the other hand, longer term targets can encourage program personnel to think in the longer term perspectives and reduce the temptation to overemphasize current results at the expense of future progress. Different targets should be established for each performance indicator in each breakout categories when the sample is broken into several categories. Different targets will make comparisons much more useful and fairer if the targets taken into consideration are difficult. This will also reduce the temptation for program personnel to concentrate on easier cases in order to show high performances. Targets can be set in various ways but the following points should always be taken into consideration in setting up targets. First, consideration of previous performance should always be important factors. Second, targets should be

established by benchmarking against the best practice available. If the program has more than one unit that provides the same service for the same types of customers, it is better to consider using the performance level achieved by the most successful unit as the target for all units. Third, if benchmarking against the best seems too ambitious, the average performance of all units can be considered an alternative. If the program wants to be more conservative, it could use the worst value as the target to emphasize the need to achieve the minimum acceptable level of achievement. Fourth, the program agency can choose as the target level the performance achieved in the past for different customers or workload categories. Fifth, the program agency should make sure the targets chosen are feasible given the program's budget and staffing plan for the fiscal year. Sixth, targets should be adjusted to reflect changes in internal and external conditions that may affect the program's potential performance. Seventh, a target does not have to be a single value. A range is an acceptable alternative if substantial uncertainty exists.

3.1.4. Reporting and Uses of Performance Information

a. Reporting Performance Information

How the findings from analyses of performance data are reported to audiences, internal and external, is just as important as what is reported. More often than not, the importance of good presentation for performance information is not fully recognized especially in the public arena. Technological advances enable program agencies to make much clearer and more attractive presentations with the use of diverse forms of much easier and cheaper to use presentation tools. Of course, it should be a given that the contents and information the presentations convey are much more important than the format or skill used. There are two types of performance reporting, external and internal.

External reporting of performance information is a major device used in ensuring accountability of program agencies. It enables elected officials, interest groups, and the general public to verify whether they are getting what they paid for. External reporting also has the potential benefit of motivating the program agencies to perform better. In addition, external reporting offers more benefits by making it possible for outside stakeholders to compare with performances of other public agencies or private organizations that provide similar services, if any. The development of information technology fundamentally transforms the way to report the program agency's performance information to outsiders. Web-based reporting is gradually replacing the traditional form of reporting, paper reports. Many public agencies have their own websites and started to post the performance data on their websites. The trend is expected to continue, making electronic reporting the major channel citizens and interest groups obtain performance information. Notwithstanding web-based reporting is a fast and inexpensive way of performance reporting, program agencies

should make a lot of efforts to make sure easy and fair access to performance information through the internet and to keep the information on the websites up to date in a timely manner. One important decision making that should be made before the performance information is provided to external stakeholders is how much detail should be given to those outside the program. The principle is that program agencies should provide information as much as possible as long as the readers are not overwhelmed by too much information.

Internal performance reporting is the presentation of performance information to program staffs and personnel. It is vital to stimulating service improvement since feedback is a key ingredient of good management. There are several key issues for internal performance reporting. First, all reports should be clear and substantial. Second, the reports should be timely. Performance information should be provided with sufficient frequency and reasonably up-to-date. Some performance information may need to be reported more frequently than others. For example, it is enough to report performance information constructed from household surveys on an annual basis while performance reports containing incidence of crimes need to be reported considerably more frequently than annual frequency to enable program managers to take timely corrective actions. Third, the reports should adequately summarize or highlight the important portion of the performance information to allow program managers to digest in a reasonable amount of time. Fourth, the internal reports should be disseminated to all those who need and can use the information.

b. Uses of Performance Information

Performance information should be used to help improve program performance through learning and not to shoot the managers of the program. We can list the major uses of performance information as follows;

- Respond to elected officials' and the public's demand for accountability.
- Help formulate and justify budget requests.
- Help allocate resources throughout the year.
- Initiate in-depth examination of why performance problems exist.
- Formulate and monitor the performance targets of contractors.
- Provide data for special, in-depth program evaluation.
- Support strategic and other long-term planning efforts.
- Analyze options and establish priorities.
- Communicate better with the public to build trust and support for public services.
- Help motivate program personnel to continue improving the program.

Though accountability has been directed at legal and appropriate use of taxpayers' money such as avoiding waste, fraud, and abuses, accountability for producing expected results also becomes important. However, it is not possible to require full accountability of program agencies since public agencies possess the limited influence over program outcomes. It is more realistic to recognize that accountability for program outcomes is usually shared with other agencies, with other levels of government, with private organizations involved in the program, and with the customers. Therefore, performance information should be used as an element, though instrumental, in checking whether the program agency meets the demand for accountability by elected officials and the public.

A major use of performance information is to help officials determine what resources and activities are likely to produce the best outcomes. That is, past performance information can be used to help agencies develop their budget requests, rather than first formulating their requests and then including available performance information as part of the submitted budget to justify their demand. Tracking outcomes of the program and linking them with budget allocation can give potential funders such as the congress greater confidence that the money they provide will be used beneficially. If a program does not provide substantial evidence that it is producing benefits to the customers and the society, the program's budget is likely to become more vulnerable to scrutiny of budget watchers.

Performance information, especially outcome information should be actively utilized throughout the year in identifying problematic areas, which is an important step in determining the need to reallocate resources. That is, performance data can be used to identify elements with poor performance so that additional resources can be focused on them to improve the performance or program managers may decide to reduce or sever the resources allocated to the problematic areas. Performance information also can be used to help prioritizing among competing options. The practice has been done for many years for various infrastructure projects, such as selecting among street repair projects. For each road to be repaired, existing condition and severity are obtained to be combined with estimates of the cost to bring deficient conditions to a satisfactory level, and data on the number of people affected. Using the information the program agency can order the projects and select based on the funds available.

One of the key uses of performance information is to raise questions on why the performance of a program is good or bad. Breakouts of performance information according to certain criteria can provide at least partial clues to why problems exist. In some cases, formal in-depth studies are called for to examine the causal relationship between the program services and performance.

If the program agency contracts out or provides grants to other public or private organizations for service provision, performance targets, especially outcome based ones,

can be included in the agreement. This is called performance contracting. The targets included in the contract should be carefully developed and compatible with indicators in the program's performance management system. Both rewards and penalties can be included in the agreement to ensure that contractors put their best efforts forward to accomplish the targeted level of performance. A lot of service contracts include termination options for nonperformance, but these options generally apply to extreme circumstances and do not appear to provide much incentive for improving performance. An additional motivator for good performance is to make past performance an explicit criterion for future rewards.

When monetary compensation for good performance is included in the contract, the program agency needs to work out incentive schemes that are fair both to the public and to the contractor. Developing such schemes requires considerable skill and cooperation between the program staffs and contractors. To make performance contracting more effective, an agency also needs a strong contract oversight function that either collects the performance data itself or regularly checks the quality of performance data provided by the contractors. In general, state verification of the performance indicators in the contract is costly and disagreement on the state of the indicators may result in lengthy controversies between the agency and the contractor. The most feasible approach when contractors are obliged to regularly provide relevant information is to help them maintain their own performance management system but require them to allow the program agency to periodically examine the system.

A program agency's performance monitoring system can often provide an excellent starting point for more specialized and in-depth program evaluation on the program performance. Data from the system are particularly useful for program evaluators that would otherwise have to collect them. Even when the system does not provide enough data that program evaluators need to collect additional data by themselves, the performance monitoring system can shed light on the issue addressed by the program evaluation and even lead to raise new issues.

Strategic planning is about the future and therefore involves many forecasting, projection, and extrapolation beyond the scope of the performance monitoring system. Strategic planning can be helped considerably, however, by information from the performance monitoring system. Strategic planning requires identifying long-term objectives of the program and an examination of alternative means of meeting the objectives. The planning that goes into development of the strategic plan should examine various tradeoffs, including the performance results that are expected from each alternative means of meeting the objectives. The estimates for performances of those alternatives are most likely to be based on the past performance information available. Roughly speaking, information from the performance monitoring system can be used for three purposes in strategic planning; to

provide baseline values for each of the plan's performance indicators, to provide historical data on each performance indicator so that performance can be projected for each option examined, and to provide data on key performance indicators that can be used in regular reports that track progress toward meeting strategic objectives.

One of the most difficult tasks a program agency is consistently confronting is to decide among options and establish priorities among competing uses of scarce resources. Information from the performance monitoring system can usually provide important information to help make those choices.

Information from the performance monitoring system can also be used to enhance the ways to communicate better with the public and, over the longer run, to increase the public's trust, confidence, and support for the organization as well as the program. Here are some points to remember in involving the public in the performance monitoring system. First, program agencies can utilize the customer focus groups in identifying performance information they should track. In this case, the agency should note in performance reports that customers' inputs had been included in the determination of what performance information to track. Second, program agencies should actively pursue feedback on the performance and quality of services from the customers and the general public. Third, if performance reports present information important to the public and the information is presented in a clear, fair, and balanced way, the public may be more likely to support the agency and its service. This will make it more likely that the public's concern about performance helps motivate agency personnel to focus on performance.

Performance information can motivate program managers and their staff members to identify and implement ways to continually improve service. Many public employees are surely motivated by the desire to serve the taxpayers better by achieving good performance. Regular performance reports may provide a strong incentive for the employees of the program agency to improve. Incentives for high program performance derived from performance information can be categorized into two groups, non-monetary and monetary. Monetary incentives especially in the public organizations are quite controversial. Monetary incentives have typically relied on a superior's subjective opinions or judgment, which may result in a contention if the supervisor's decision is not fair enough to the supervisees. The performance monitoring system can provide more objective data and the use of the objective data, if the data are believed to be appropriate, is likely to increase the acceptance of rewards as motivators by employees, elected officials, and the public. There are two types of incentives program managers can use performance information to motivate program personnel, monetary and non-monetary incentives. We discuss them in the following section.

3.1.5. Use of Performance Information for Motivation

a. Non-monetary Incentives

Non-monetary incentives have the advantage of being inexpensive. They are probably the most common form of reward in the public sector, but are not generally believed to offer strong motivations.

Providing the latest information on the state of performance indicators relative to the targeted levels immediately after each reporting period can encourage program managers and their staff members to pay attention to the reported performance and shortfalls. One popular example is to post regularly updated progress toward the target in the form of a thermometer registering rising temperatures. Calling attention to performance information through regular performance reports provides motivating feedback to all program personnel contributing to the program performance. Regular reports comparing performance indicators broken out by organizational units delivering similar services to similar customers can be particularly powerful in motivating poorly performing units. Such information may encourage program personnel of poorly performing units to try to identify why their performance is unsatisfactory and help find ways to improve. However, an excessive drive by program managers motivated by a poor performance report may result in destructive feedback effects, if mishandled.

Program reports can also be utilized to have program personnel set targets for their performance indicators and relate performance data to the targets. The use of the program report will be particularly effective if reporting periods are frequent, targets are set for each reporting period, and the results are available soon after the end of each reporting period.

Program managers can be given more flexibility in exchange for a higher level of accountability in the program performance. Increased flexibility in exchange for more performance accountability is granted in the form of more budgeted funds, authority to make purchases without going through extensive red tape, or authority to hire, remove, compensate, and move program personnel to other tasks and positions. The flexibility should be allowed only after it is confirmed that the targets for performance indicators are achieved. However, increased flexibility may be used in illegal, unethical, or dishonest ways without a proper internal control system.

Each program personnel or agency sets up targets for performance indicators that should be achieved in a reporting period and the level of achievement can be explicitly incorporated into the performance appraisal process. There are two basic approaches to the task of incorporating performance information into individual program personnel's or agency's performance appraisal process. The first approach is to compare actuals to targets for each indicator over which a subset of program personnel had some control. Achievement

of target levels can be considered in the appraisal process. All people in the group receive the same ratings on that particular part of the appraisal. Second approach is to identify how program managers have implemented and used performance information. This approach is in much less demand than the first approach and therefore offers weaker motivation. Agencies should experiment with the second approach until all program personnel have had sufficient experience with the performance information to be responsibly rated on them.

Performance contracts between high ranking government officials and the heads of program agencies have been used as an important tool to stimulate the motivation of program personnel. In New Zealand, individual agency heads agree to produce specified amounts of products in return for specified budget levels and more management flexibility. As a part of this process, agency heads can receive extra or reduced compensation based on an agency's performance relative to promised targets.

b. Monetary Incentives

It is a very popular practice to link financial incentives to performance as an important motivator. However, the practice is fraught with pitfalls and has often produced counter-productive side effects especially in public environments. The problem is that it is very difficult to draw agreement from all stakeholders, particularly program personnel, on the fact that the performance criteria and measurement are fair and valid. Almost always, compensation systems linked to performance have ended up relying heavily on the judgments of supervisors. Employees resist such process particularly when they do not accept the fairness and validity of criteria and measurement of performance indicators. On the other hand, if a sound performance monitoring system is used as a major part of the reward criteria and is perceived by all stakeholders as reasonably objective, fair, and valid, then the use of data from performance reports in determining payment to employees should reduce potential negative effects. Another major dilemma is that external factors can play significant roles in affecting the program's performance that a rigid linkage between pay and performance may result in unexpected side effects. It is impossible to exclude the possibility that an excellent program performance has little to do with the employees' efforts. If it is the usual case, utilizing the performance report to establish a linkage between pay and performance cannot achieve the purpose. An important condition for monetary incentives to work properly is for all stakeholders including program personnel, high ranking government officials, and the public to accept the principle that program personnel will be rewarded as announced if performance matches the pre-specified target level, regardless of the extent to which the program personnel actually contributed to the improvement. Conversely, all those involved should clearly agree that employees of the program agency will not be rewarded if performance levels do not meet the expectations, regardless of what may have been an excellent effort and contributions.

A less controversial monetary incentive than directly linking pay to performance is to reward an agency or individual for good performance with a fund that can be used only for organizational purposes, such as employee training or improvement of workplace conditions. These incentives are most likely to be better received by government officials and the public because little additional funding is required and that funding is used not for individual but for organizational purposes. The most frequently used application is to use a part of cost saved to reward the group that contributed to cost saving. These are sometimes called shared savings or gain-sharing programs. When a public agency links monetary incentives to performance indicators, the agency should also include indicators that track negative outcomes to be avoided. It is tempting for program personnel to focus on performance indicators linked to rewards at the expense of other indicators that are not related to rewards. In the United States, a federal law enacted in 1997 promised to pay states $4,000 to $6,000 for every child adopted over a baseline number to promote adoption. This created a substantial financial incentive for the states. Several newspapers reported that caseworkers had been pressured to seek adoptions before the children or families were ready or place children with inappropriate families. To reduce this kind of danger, agencies may as well make the performance monitoring system comprehensive in that the system should be designed to cover potentially important negative effects as well as positive effects. Monetary incentives for many programs should depend on whether performance lasted a reasonable amount of time. For the financial incentives for the states for the number of adoptions, the monetary incentives should also be tied to the extent to which the placements were trouble free after a reasonable period of time had elapsed.

The performance report can also be utilized to provide justification for sanctioning low performers. The most drastic and threatening sanctions for an individual are salary reduction, demotion, and discharge. For organizations, a substantial cut in funding would be the most serious concern. As more performance data are available, they are likely to be used as the basis for such sanctions. Special care is needed when using performance data for disincentive. Most of all, it should noted that the performance data should be based on relatively objective findings. In addition, since outcomes are seldom fully controlled by the program agency or program personnel, continued low performance rather than temporary poor results, should be an important element in such decisions.

Monetary sanctions reduce or withdraw the funding of agencies or programs that have not met expected performance levels or have not provided adequate information to assess performance. Such sanctions should be carefully designed and implemented not to penalize the customers of poorly performing programs or programs struggling with insufficient funding.

3.1.6. Other Issues in Performance Monitoring

a. Common Problems with Monitoring System

Performance monitoring systems are not panaceas for improving the performance of budgetary programs and managing them effectively. Sometimes the data are not good enough to have any reasonable credibility. Non-comparability of data from different sources can create a serious reliability problem, especially with respect to benchmarking efforts. Even with a high degree of validity and reliability, the data generated by the performance monitoring system are basically descriptive and limited in terms of evaluating causal relationships regarding program effectiveness. Sometimes real outcomes are too elusive to measure and do not lend themselves to regular, systematic, quantitative measurement on a real-time basis as is done in most performance monitoring systems.

Interpreting performance data out of context can be misleading and produce erroneous impressions regarding performance. The data also can be over-interpreted in terms of causal effects and lead to unwarranted conclusions that a program is in fact producing desired outcomes. Worse, the data generated by the performance monitoring system can be misused or abused in ways that are unfair to program managers and employees and counterproductive in terms of program or agency performance. The performance monitoring system can also set up unbalanced or suboptimal incentive structures that may result in goal displacement and opportunistic behaviors at the expense of overall effectiveness. Another common problem concerns unrealistic expectations regarding the costs and benefits of performance monitoring systems. Most monitoring systems of any substantial size and complexity require a significant investment of time, effort, and financial resources, and if this is not clearly understood at the outset, enthusiasm for a system may wane as the costs are piled up. In addition, there is often internal resistance to monitoring systems on the part of managers and employees who feel threatened by the measures or concerned that they might face adverse system impacts. Finally, performance monitoring systems often fail to make a difference because they are not responsive to other stakeholder concerns or the decision makers do not deem them useful by decision makers.

b. Strategies for Developing Good Performance Monitoring Systems

Strategies are available for overcoming the shortfalls of performance monitoring systems cited above and developing effective ones. First, with respect to building credibility for the system and increasing likelihood that it will in fact be used, those who are developing performance monitoring systems should;

- Secure a commitment from the top decision makers to support and use the system.
- Communicate realistic expectations regarding the benefits of the system, as well as the time and effort required to develop and maintain it.

- Be candid about the limitation of the system with all stakeholders.
- Involve stakeholders in identifying criteria, measures, targets, and data.
- Measure and design features to users' needs and preferences.
- Communicate how and why measures are being developed, and prompt management to demonstrate commitment to suing the measures.
- Provide training to program managers on using performance data to improve their programs.

At the core of the process, developing a performance monitoring system is both an art and a science, and it often involves weighing trade-offs among competing criteria. Thus, program evaluators who are designing systems should;

- Tie measures directly to mission, objectives, goals, service standards, and targets.
- Use program logic models or other relevant framework to ensure a systematic and comprehensive approach.
- Try to be result driven rather than data driven in defining measures, but use available data when appropriate.
- Be pragmatic in evaluating measures in order to build a workable system that produces worthwhile information.
- Try to anticipate likely goal displacement or opportunistic strategies, and balance measures in order to produce such reactions.
- Install procedures to ensure data integrity.

Finally, it is important to build features into a performance monitoring system that will help generate acceptance and encourage its use. Thus, system designers should;

- Keep measures and presentations simple and straightforward and not employ more measures that are not absolutely necessary.
- Emphasize useful comparisons in reporting systems, and break out data by important characteristics that can draw attention from all stakeholders.
- Provide adequate explanatory information along with the performance data, and provide fields in reporting formats for explanatory comments.
- Give program managers and others a chance to see the performance data first and make corrections and comments before reports go to higher level management.
- Avoid over-interpretation of the data and drawing unwarranted conclusions regarding a causal relationship between programs and outcomes.

Although effective program monitoring is not easy, it is not rocket science either. Rather, it is a commonsense approach to the performance based management system that can be useful in providing though descriptive but still very useful information on program performance on an ongoing basis.

3.2. Program Evaluation

3.2.1. Program Evaluation: Definitions and Important Features

A program is a set of organized but often varied activities directed towards the achievement of specific objectives. It may encompass several different projects, measures and processes and also tend to have a definite time schedule and budget. In the context of the performance based budgetary management system, a program is thought to indicate a group of individual budgetary expenditure projects that are designed and implemented to achieve common specific policy objectives. The program evaluation can be defined in many different ways partly because it is closely related to various academic fields such as economics, policy studies, statistics, public administration, and psychology. It is also because many program evaluations are conducted by a wide range of people with diverse purposes.

Given that it is probably impossible to arrive at a single definition which can obtain universal acceptance, the OECD's definition merits our attention. OECD (1999) defined program evaluation as "a systematic and analytical assessment addressing important aspects of a program and seeking reliability and usability of findings". We can identify several important characteristics required for all program evaluations. First, program evaluations should be analytical discourses based on recognized research techniques. Second, program evaluations should be systematic, requiring careful planning and consistent use of the chosen techniques. Third, the findings of a program evaluation should be reproducible by a different evaluator with access to the same data and the same methodology. Fourth, issue-orientedness requires that program evaluation should seek to address important issues related to the program including its relevance, efficiency, and effectiveness. These are three important evaluation criteria adopted by many evaluation processes. Fifth, program evaluations should be user driven in that successful program evaluations should be designed and implemented to provide useful information to decision-makers, taking into consideration the political circumstances, program constraints, and available resources.

The comparison of program evaluations with other similar analyses such as scientific studies, audits or performance monitoring would help us grasp the meaning of program evaluation with ease. Program evaluations differ from scientific studies, though both identify the causal chain in analytical, systematic and reliable manners. However, whereas scientists may undertake academic research in order to expand the realm of human knowledge and

frequently confine themselves to one highly specialized discipline, program evaluations are undertaken for more practical reasons. They are intended to be of practical use by informing decisions, clarifying options, reducing uncertainties and providing information about program performance. Next, program evaluation is also different from audit. While audit is primarily concerned with verifying the legality and regularity of the implementation of resources in a program, program evaluation, on the other hand, is necessarily more analytical and focuses on the results of a program. Program evaluation looks at the validity of the strategy adopted and whether objectives are appropriate given the problems to be solved and the benefits to be achieved. Auditors tend to have coercive powers, sometimes defined in legal texts, whereas program evaluators must often rely on the power of their arguments. Audit has traditionally covered activities such as the verification of financial records. A more recent innovation is known as performance audit, which is conceptually closer to program evaluation. Performance audit is strongly concerned with questions of efficiency of a program's outputs and good management. Performance audit and program evaluation share the same aim of improving the quality of services provided by programs, but program evaluation goes much further. It also looks at issues such as sustainability, relevance and the long-term consequences of a program. Finally, program evaluation must be distinguished from performance monitoring. Performance monitoring examines the delivery of the services produced by the program to target customers. It is an on-going process, carried out during the execution of the program, with the intention of immediately correcting any deviation from operational objectives. Program evaluation, on the other hand, is specifically conducted at a certain point in the life cycle of a program, typically after a certain amount of time has elapsed since the program started and consists of an in-depth study on the outcomes of the program. Both performance monitoring and performance audit are important in improving program performance and contribute to successful fulfillment of program evaluation, particularly because data collected from a performance audit or performance monitoring can be utilized in program evaluation.

Program evaluations focus on a group of projects sharing common goals or objectives rather than a single project. The core task in program evaluation is to examine the validity of intervention logic forwarded by its designers. Programs are always conceived with a given set of needs, which are problems programs seek to address from the perspectives of the program's customers who are the beneficiaries of the program. For example, in a program to reduce the youth smoking rate, the program customers are young smokers and the problem to be addressed by the program is premature death due to smoking related diseases.

Before discussing intervention logic in the program evaluation, let's go back to a logic model shown in [Figure 2-3]. A logic model typically consists of several steps, inputs, activities or processes, outputs and outcomes. The model offers a simple conceptual

framework that illustrates the sequential chain of actions and events from inputs to outcomes. In other words, a logic model is a plausible and sensible model of how the program will work under certain environments to achieve the outcomes. Intervention logic of a program refers to the conceptual link from inputs of the program to its outputs and outcomes in a logic model. The examination of validity of the intervention logic of a program step by step is the central task in a program evaluation. The evaluator asks a series of questions on how the inputs of a program lead to outputs, and ultimately the outcomes, immediate and end. Typically intervention logic of a program contains hidden assumptions about the causal chain between the program and its supposed effects and about how the program influences, and is influenced by, other factors. An important task is to identify these hidden assumptions so that they can be critically assessed by the evaluator.

The evaluator examines the validity of intervention logic focusing on the following five aspects of the program; relevance, effectiveness, efficiency, utility, and sustainability. Relevance means that to what extent the program's objectives are congruent to the evolving needs and priorities of society. Discussion on relevance of a program may lead to a more serious discussion on whether a program should be allowed to continue in its current state, altered significantly, or merely allowed to elapse without renewal. Examining relevance of a program, the evaluator asks the question whether broad changes in society have altered the rationale for a program, or may do so in the future. The discussion of future relevance leads the evaluator to examine alternative options. The next issue the evaluator should address is efficiency of a program. Efficiency refers to the notion of how economically various inputs were converted into outputs and outcomes. Efficiency compares resources expensed by the program with the services and impacts a program provides. At a practical level, checking efficiency of a program involves asking the question, "Could the same benefits have been produced using fewer inputs?" Alternatively, "Could the same inputs have produced greater benefits?" Focusing on efficiency necessarily leads to comparisons between the program under evaluation and alternatives. The main difficulty is the choice of appropriate benchmarks. The evaluator needs to specify against which benchmarks the efficiency of a program is being measured. Difficulties may arise when no comparable programs have been implemented or the evaluator has not been exposed to the evaluation projects of programs with similar characteristics.

Examining effectiveness of a program, the evaluator should ask the following question, "How far have the outcomes of a program contributed to accomplishing program objectives and satisfying the needs of program customers?" Even if a program is efficient, it can still be ineffective in achieving the program objectives. In other words, efficiency and effectiveness should be treated as two different things in evaluation projects. It is most likely that for a program with poor designs, objectives may not have been stated clearly or may even be

missing altogether. The evaluator may be required to modify vague or abstract program objectives into clear and concrete ones before he continues the evaluation project. In addition, it must be remembered that effectiveness is concerned with only one aspect of a program's impact: the positive, expected aspect. A program may also have negative and unforeseen effects as well as positive and expected ones. A balanced evaluator may want to take into consideration all impacts of a program in examining effectiveness. In order to assess the overall effectiveness of a program, the evaluator is required to identify causality between the program and its impacts, positive or negative. Attributing causality is the key problem that should be solved in program evaluations. Other possible explanations for the effects which are to be attributed to the program must be identified and eliminated so that the evaluator can show that the positive effects would not have arisen anyway.

The next issue is the utility of a program, which involves the degree of a program's impacts or outcomes compared with the needs of the target population. Programs are useful only if they manage to bring about socially beneficial changes in response to the needs of the target population. A particular problem with utility criterion is that it is difficult to arrive at a universally acceptable definition of needs. However, evaluators somehow should be able to identify the contents or boundary of the needs that draw consents from the majority of the population. The final issue is sustainability. Sustainability requires that the positive changes brought about by the program should last for a reasonable amount of time after the program has been terminated. Even if a program generates benefits in tune with the needs of its target population, it may be of little value unless these benefits are still being enjoyed at some stage in the future. Sustainability is therefore concerned with what happens after a program has been completed. For example, there is little value in training unemployed workers in skills which are likely to become obsolete after a few years.

It should be noted that a wide range of individuals and private or public groups have legitimate interests in program evaluation. We have used the term stakeholders to indicate them. The Stakeholders are various individuals and organizations who are directly or indirectly affected by the implementation and outcomes of a program, and who are likely to have an interest in its evaluation. A list of stakeholders of program evaluation, though not comprehensive, consists of policy makers and high level decision makers, the person in charge of the evaluation of the program, the target customers of the program, program managers and staff members, and other individuals and organizations with a legitimate interest in the program and its evaluation. The evaluator in most cases is tendered and selected by the sponsor of the program evaluation and is responsible directly to the sponsor and indirectly to all stakeholders. The evaluator should bear in mind that he should be able to demonstrate the understanding of the diverse information needs of various stakeholders and the importance of different stakeholders at various stages of the evaluation process.

Program evaluations are typically categorized in three ways; formative versus normative, intermediate versus ex-post, and internal versus external. First, distinction between formative and summative evaluations is based on the intended users of the evaluation. Formative evaluations are concerned with examining ways of improving and enhancing the management and implementation of programs. Formative evaluations tend to be conducted for the benefit of those managing the program with the intention of improving their work. Summative evaluations are concerned with determining the essential effectiveness of programs. Summative evaluations tend to be conducted for the benefit of external stakeholders who are not directly involved in the management of a program, for reasons of accountability or to assist in the allocation of budgetary resources. Although the distinction between formative and summative evaluations seems to be clear-cut, in practice it is often blurred. A general concern with improving public programs usually requires a combination of both approaches. Second, the difference between intermediate and ex-post evaluations mainly lies in the timing of the evaluation. Intermediate evaluations are conducted during the implementation of a program and ex-post evaluations either on or after the termination of a program. In many cases, intermediate evaluation often focuses on a program's outputs and do not attempt a systematic analysis of the program's outcomes and effectiveness. They tend to rely on information provided by the performance monitoring system and show formative concerns such as improving the program's delivery mechanisms. In some cases, intermediate evaluations do look at outcomes at the full sense, but only in a limited way. Ex-post evaluations are more likely to be summative in nature, and are often conducted with the express intention of analyzing a program's effectiveness. However, since the information needed to assess a program's impact may often not be fully available until several years after the end of the program, even ex-post evaluations can be limited in the extent to which they can provide a complete assessment of impact. Since many public programs are replaced by successor programs with similar objectives, ex-post evaluations can be justified even if they are conducted even after the termination of the programs. Third, another important way of classifying program evaluations is to make a distinction between internal and external evaluations. Internal evaluations are performed by members of the agency in charge of conducting the program under evaluation while external evaluations are conducted by a person outside the agency managing the program. A lot of program evaluations tend to be performed by external evaluators in order to secure independence and expertise of the evaluation. In order to ensure that external evaluations are conducted properly, evaluation sponsors must pay particular attention to drafting the terms of reference. Furthermore, unless there is proper supervision of the external evaluator during the conduct of the evaluation by the evaluation sponsors, a number of problems can arise. For example, evaluation reports prepared by external consultants may produce misguided recommendations due to a lack of sufficient knowledge on internal politics and organizational culture. In addition, external

evaluators may be too far removed from the chain of management for their findings to be taken into account so that the findings and recommendations of the evaluations may end up sitting in the cabinets. It is very important to make sure that the supervision of external evaluators by the evaluation sponsor does not compromise the evaluator's independence.

Internal evaluations also have their own benefits. Internal evaluations provide excellent opportunities to educate the internal members of the agency by promoting learning by doing since good management of public programs is closely linked to questioning the 'how' and the 'why' of their activities. However, internal evaluations cannot go well with summative evaluations since it is generally quite difficult to convince other stakeholders, especially outside the program agency, that an internal evaluation has been conducted objectively.

3.2.2. Preparation of Program Evaluation

a. Establishing Management Structure

Preparation of a program evaluation starts with the establishment of the management structure of the evaluation task. An efficient management structure should ensure that the evaluation project is of high quality, available in good time and produced at a justifiable cost. The chief task of the management structure is to delineate object and scope of the evaluation and draft the terms of reference for the evaluation, in particular, if it is entrusted to an external expert. An efficient way of establishing the management structure is to set up a steering group. The steering group should include representatives of the agency in charge of the evaluation, evaluation sponsor if it is different from the program agency, independent evaluation experts, and relevant stakeholders from the civil sector. It is also possible to include representatives of those who are entrusted with evaluation as long as doing so does not compromise the objectivity and independence of the evaluator. It is always recommended to establish a steering group especially when programs are of major budgetary significance, or of a controversial nature, or when the evaluation's focus is not simply confined to the implementation of the program but also looks at the program's effectiveness and future relevance. By establishing a steering group, we expect it to encourage the various stakeholders to get actively involved in the evaluation, to reduce the chance that program managers become too closely associated with the evaluator, posing the danger of compromising independence of the evaluator, to allow for quality control of the evaluation by experts. Creating a steering group helps convince people that the evaluation is regarded as an inclusive process. Stakeholders are then more likely to have confidence in the evaluation's conclusions and recommendations, especially if they have had the opportunity to influence the management of evaluation. A steering group should not be too large since it may lose the roles as management body and become a mere negotiation platform threatening the impartiality of the evaluation process.

b. Making Evaluation Plan

Planning an evaluation project is involved in the following steps;

- Identification of the goals of the evaluation.
- Delineating the scope of the evaluation.
- Drawing up the analytical agenda.
- Setting benchmarks.
- Taking stock of available information.
- Mapping out the work plan.
- Selecting the evaluator.

Every evaluation planning should start by questioning the goals of the evaluation; that is, why do we want to launch an evaluation project? There are three specific reasons why program evaluations are conducted. First, a program evaluation is conducted to improve the program management. Examining implementation and delivery systems of the program should be the major concern in the evaluation if the primary purpose is the improvement of the program management. The final evaluation report could be quite technical since the primary audiences are people who are already familiar with the program such as program managers, program supervisors, and direct beneficiaries. Second, a program evaluation is conducted with a view to enhancing accountability of the public agencies in charge of the program. If it is the primary purpose of the evaluation, it should focus on the effectiveness of the program. Possible side effects and specific issues associated with equity and transparency are also important in securing accountability of the program agency. The final report should be written in a plain style with non-technical terminology that every stakeholder can easily understand. Third, a program evaluation is also very useful in providing data and analytic results to assist in the allocation of budgetary resources. If the evaluation sponsor puts emphasis on obtaining data to decide a program's renewal and related budgetary needs, the goal of the evaluation should be to shed light on the efficiency or cost effectiveness of the program, justifiability of continued government expenditure, and searching for possible alternatives. The evaluation report should be written in a style that is easily comprehended by decision makers and opinion leaders who may not have enough knowledge on the program.

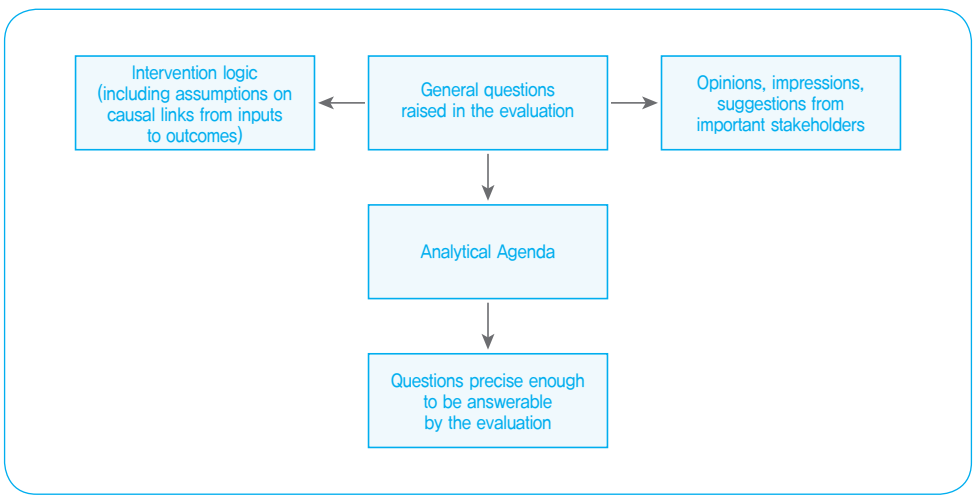
Delineation of the scope of the evaluation is to decide what to be evaluated. Without proper delineation of the scope at the planning stage, the program evaluator is most likely to be pressed to be as comprehensible as possible. Then, it would be an endless task to look into every imaginable aspects of a program, or into all its actual or potential ramifications

with other programs. Another major topic of discussion related to the scope is which key evaluation issues to focus on among relevance, efficiency, effectiveness, utility and sustainability. The choice will be influenced by various factors such as availability of data, time constraints and limitations on financial resources allowed to the evaluator. Regardless of the delineated scope of the evaluation, there is a core question that should be asked in all program evaluations; has the causal link from inputs to outputs, and subsequently to the achievement of outcomes occurred as initially envisaged? If the answer to the question is negative, the evaluator should do his/her best to provide the reason why the program performance was not satisfactory.

The next step in the evaluation planning is to formulate the analytical agenda. Analytical agenda is a logical structure imposed on the different questions asked in the evaluation. The aim to clearly draw up an analytical agenda is to transform the general, often vague, questions raised by those who called for the evaluation into questions which are precise enough to be manageable by evaluators relying on scientific methodology in examining the issues. The procedure of analytical agenda formulation is illustrated in [Figure 2-6]. The analytical agenda is a way of transforming the general questions into more precise questions that program evaluators can answer the questions on a scientific basis. The two main sources for the general questions are intervention logic that was advanced when the program was designed or prepared for the evaluation and opinions from important stakeholders. As for intervention logic, special attention should be paid to the hidden causal assumptions. Retrieving the original intervention logic of a program is usually easier said than done. Official documents and records in most cases do not contain any systematic explanation on the causal chain assumed and even the program's objectives are not stated clearly. More often than not, a substantial amount of study on official documents and records is needed to retrieve the correct interpretation of truly meaningful questions for evaluation purposes. It is nonsense to judge or evaluate success or failure of a program when we do not know its objectives or purpose. Whenever the objectives of a program have to be reconstructed from scratch, this should be done transparently by the management structure, preferably under the responsibility of the steering group. Another useful source of information in drawing up the analytical agenda is important stakeholders' opinions on the program. The opinions can be expressed in various forms such as impression, suggestion, even complaint. These opinions should be examined carefully as working hypotheses in the evaluation. One cautionary note is that the opinions from important stakeholders should not compromise the impartiality and independence of the evaluation process. When the general questions have been identified, the analytical agenda should be drawn up. This means that the evaluator should come up with a set of precisely written questions to which the evaluator should provide answers. Once the analytical agenda for the evaluation has been drawn up, both the evaluation sponsor and steering group need to ask whether the program is indeed evaluable.

The questions that were identified when the analytical agenda was drawn up should be answerable by an evaluator using appropriate research methodologies. If it turns out that a program is not evaluable, this can lead to a decision to postpone the evaluation or to draw up a new, more realistic analytical agenda. Nevertheless, it should always be remembered that it is better to have imprecise answers to important questions than to have precise answers to unimportant questions.

Figure 2-6 | Procedures of Analytical Agenda Formulation



Program evaluation is basically about revealing the value of a program. This requires the evaluator to make value judgments on the program performance. To ensure that the value judgments do not become arbitrary or biased, it is a good idea to announce in advance transparent and understandable benchmarks against which the program performance would be compared. For example, the evaluation steering group should be able to provide the criteria to rate the effects or the standards to evaluate the functions or success of a program. Setting benchmarks could be a tricky job. Objectives of a program are sometimes expressed in very vague terms and a single program may have multiple objectives, some of which may carry relatively more weight. Moreover, objectives of a program may change over time as the program environment evolves.

There should be more to benchmarking than simple reconstruction, clarification, and prioritization. Ideally, benchmarks should allow us to compare the program's performance with that of other programs with similar objectives. Even if the performance of a program falls short of the target level, it may not mean that the program performance is unsatisfactory. It may compare favorably with results achieved by similar programs executed in the past,

or by other agencies. It is always possible that the performance of a program fails to reach the target level not because of poor performance but because of the unrealistically high ambition of the agency in charge of the program.

When judging the program performance by means of benchmarks, the fundamental caveat to be kept in mind is that the benchmarks may have been reached by virtue of factors not attributable to the program. An effort should be made to separate out these factors in order to identify the net effect of a program on the achievement of its objectives.

Having done the benchmarking, the next step is to tap into the existing stock of information and knowledge on the program under consideration itself and the evaluation strategy. For most programs, the performance monitoring system should be the first source of information. The quality of information from the performance monitoring system is crucial in the quality of the program evaluation. Unfortunately, in a lot of cases, the data collected from the performance monitoring system is insufficient or inadequate to be used in the program evaluation. Other sources of information readily assessable include academic literature on program evaluation, media reports, administrative data, and published statistics. Ex-ante evaluation reports, if they exist and are available, can also be a very useful source of information to design an in-depth program evaluation since they must have taken into account the existing knowledge at the time of the program's inception. However, the limitation of the ex-ante evaluation reports should also be clearly recognized. It is not possible to conduct systematic ex-ante evaluation for all programs and, even for the programs that had undergone ex-ante evaluation, it is highly likely that several years have already elapsed since the program was launched, calling for significant updates and revisions of reports. The primary benefit of tapping into the existing stock of information comes from comparing it with the needs that were raised by the analytical agenda. The comparison clearly reveals the information gaps which, in turn, set collection and interpretation of the data to be undertaken by the evaluation. The analytical agenda may be the result of a maximalist approach without realistic consideration on feasibility or desirability, raising questions on the data that can only be of doubtful quality or obtained at large cost. Some of these questions may be fairly remote from the key objectives of the program. Evaluations face a time and budget constraint, so that before launching data collection activities, it should be decided whether the data to be generated are liable to shed any significant new light on the subject under scrutiny.

After finishing formulating the analytical agenda and checking the available information, management hierarchy of the evaluation project should be able to draft a work plan that maps out a detailed roadmap that the program evaluation should follow once it is launched. The work plans should include detailed descriptions on data collection, analytic agenda, and research methods to be employed. For management purposes, it is always a good idea

to divide the work plan into several distinguishable stages setting the timetable for the completion of each stage. The work plan should also include the assessment of the evaluation cost. When the evaluation is done internally, the cost analysis should take into consideration the opportunity cost of time to be spent in evaluation as well as explicit expenditure. If external evaluators are employed, the total cost should be estimated before the call for tender is announced. This is done in order to verify that the budget for an evaluation by external experts is compatible with the analytical agenda contained in the work plan. Cost estimates should always be realistic. All too often, evaluations do not achieve what they set out to do because initial expectations were too high. For example, collecting data that cannot be garnered from the performance monitoring system would be quite expensive.

The final step in the planning of a program evaluation is to select the evaluator. Selection of evaluator should reflect the diverse features of evaluation works. Some evaluation activities are complicated and technical so that they require dedicated research by experienced specialists. On the other hand, many evaluation works can be carried out by evaluators with reasonable experience and sector specific knowledge on the program. In any case, the selected evaluator should maintain some professional distance from the program agency in order to ensure the objectivity and independence of the evaluation process. The most important criterion in selecting an evaluator is the technical ability to carry out the evaluation, but it is not sufficient by itself. Other important issues involved in the choice of evaluators are the ability to obtain access to various sources of information, knowledge and previous experience of the field related to the program, and independence from main stakeholders.

There are a number of different types of organizations or groups that can be appointed as external evaluators. Two of the most often used are management consultancies and academic institutions. Management consultancies are private companies that possess expertise and resources for program evaluations. Large companies with considerable experience in carrying out a range of different evaluations tend to be preferred to smaller ones that possess highly subject-specific expertise in a narrower area. Management consultancies are often perceived by stakeholders to embody a businesslike approach. They tend to have advantages in terms of time and expertise over other types of external evaluators. They also have disadvantages. Their services are expensive, which may cause a serious concern in public agencies under tight budget constraints. Also, they may try to save cost by applying one-size-fits-all style solutions rather than devising a tailored approach to the evaluation. Academic institutions are likely to offer a high degree of methodological expertise and technicality that may help enhance the quality of the evaluation. They may possess an excellent knowledge on specific fields and be regarded as independent by stakeholders. Academic institutions are most likely to be a better option than management consultancies

when it comes to cost to the evaluation sponsor. They may offer better value for money than management consultancies.

For large size programs covering various layers of customers or regions, it is often strongly recommended that a consortium of evaluators is commissioned to conduct the evaluation. The consortium may consist of different types of evaluators. The normal practice is to appoint a lead evaluator to supervise the evaluation processes. The lead evaluator should possess ample knowledge and experience in the field related to the program under evaluation, expertise in evaluation methodology, independence and external legitimacy, and ability to meet the required timeline of work.

c. Drafting Terms of Reference

The terms of reference are official documents that outline the work to be carried out, the questions to be addressed, and time schedule of the evaluation. Therefore, the primary use of the terms of reference is to offer the evaluation sponsors the opportunity to define their requirement for the evaluation works and to clearly inform the evaluators of what is expected from the evaluation to be undertaken. Though the terms of reference must include clauses concerning specific circumstances of the program under evaluation, they cover some of the core elements such as;

- Legal base and motivation for the evaluation - It is helpful for both the evaluator and the sponsors if the terms of reference specify the legal and contractual requirement upon which the evaluation is based.
- Future utilization of the evaluation results - Evaluator will benefit from the information on how the findings will be used and who will be the primary users, in identifying the underlying intention of the sponsors such as relative emphasis on implementation and outcomes of the program and the level of detail in research.
- Description of the program - The terms of reference should normally include a succinct but comprehensive description of the program to be evaluated including, for example, intended target population, general and specific objectives, inputs and outputs, and delivery systems.
- Scope of the evaluation - The terms of reference should also include the clear demarcation of the scope of the evaluation. They should provide the answer to the following questions; will the evaluation cover the entire program or a part of the program? Will the program evaluated as an independent stand-alone unit or should the evaluator consider the relationship to other program? Is it the duty of evaluator to examine just whether the expected outputs and outcomes of the program are actually achieved or not, or to include unforeseen results and outcomes, positive or negative, as well?

- Main evaluation questions - It is important to specify the questions from the analytical agenda in the terms of reference to provide the evaluator with precise guidelines on what are expected from the evaluation by sponsors and stakeholders. The main question one should not omit from the terms of reference is whether or not the intervention logic of the program is valid.
- Data collection and analysis methodologies - Evaluators will benefit if the evaluation sponsors provide clear guidance on the data collection and analysis methods. However, it should also be noted that there is no single, universally applicable methodology applicable to all evaluations. The methodology for data collection and analysis must reflect the specific circumstances of the program and the particular questions whose answers the sponsors want to know. In the case of external evaluations, broad guidelines can sometimes be preferable, at least at the "call for tenders" stage. This allows the chosen contractor to use any knowledge and experience to refine the suggested approach through a process of negotiation and discussion with the sponsors.
- Work plan - The work plan for the evaluation should include factors such as the length of the contract and the deadline for reporting. By specifying the organizational structure of the evaluation project, the roles of different players in the evaluation process should be clarified. This is especially important when the evaluation is conducted by multiple evaluators. Work plan also specifies the responsible agent for final reporting and dissemination of evaluation of the results. Budget for the evaluation should also be stated in detail.
- Structure of final report - There is no universally acceptable structure for evaluation reports, although all reports should include an executive summary as well as a copy of the terms of reference. A typical evaluation report includes title, table of contents, executive summary, introduction, explanation on research methodology, evaluation results and findings, and appendix, if necessary.

3.2.3. Establishing Evaluation Design

a. Evaluation Designs

An evaluation design means a logic model is used to gather evidence on results or changes attributed to a program. In evaluation literature, a program is regarded as an experiment and the purpose of the evaluation is to identify the effects of the experimentation on the subjects. The basic principle of identifying the program effect through an experimental setting is illustrated in <Table 2-1>. Evaluation design is involved in comparing two groups, one of which, called the treatment group, exposed to the program, and the other, called the control

group, is not, and attributes the differences between the two groups to the program. This type of design is referred to as the ideal evaluation design.

Table 2-1 | Evaluation Design

	Before	Exposure to treatment	After
Treatment Group	T_0	O	T_1
Control Group	C_0	X	C_1

In <Table 3-1>, the subscript 0 indicates the measurement or observation before the experiment and the subscript 1 after the experiment. Here, all measurements such as T_0 and C_1 should be understood as the average of the measured values for all members of the corresponding group. Therefore, $(T_1 - C_1)$ is the difference between the averages of the control group and the treatment group rather than the difference in individual observations.

Under the ideal evaluation design, the outcome attributable to the program can be measured by the difference between treatment and control groups after the program and expressed as $(T_1 - C_1)$. Control group has been constructed to possess similar characteristics with the treatment group except for the fact that no member of the group participates in the program. It is natural to conclude that the difference between the two groups that are found only after the program should be attributed to the program. The crucial requirement for the ideal evaluation condition is that treatment and control groups should be identical, especially in terms of the attribute the program intends to affect. The significance of the ideal design is that it serves as the underlying proof of program attribution for all evaluation designs described in the following discussion.

Causal inferences are made by comparing identical groups before and after a program for all evaluation designs. What distinguishes the various designs is the extents to which the comparison is made between groups that are identical in every respect save for exposure to the program. Three types of evaluation designs are frequently utilized in program evaluations depending on data availability and budget; experimental design, quasi-experimental design, and implicit design. Experimental design, also called randomized design, is the most rigorous design setting, ensuring the initial equivalence between the treatment group and control group by creating them through the random assignment of participants to each group. Random assignments of group members ensures the expected values and other distributional characteristics of T_0 and C_0 are equal that they provide the ideal evaluation design to the evaluators. Quasi-experimental designs come close to experimental designs in that they use experimental groups to make causal inferences, but they do not use randomization to construct treatment and control groups. In quasi-

experimental settings, the treatment group is already given to the evaluator. The evaluator is now obliged to construct one or more comparison groups to match the treatment group as closely as possible before getting program treatment. Without randomization, one cannot assume that the condition for the ideal evaluation design, comparability between the two groups before the program treatment, is satisfied. Therefore, the evaluator should be ready to deal with potential incompatibility of the two groups. Nonetheless, quasi-experimental designs are the best that can be hoped for when randomization is not possible. With implicit designs, measurements are tried for the treatment group after the program treatment without any reference to the control group. Therefore, these evaluation settings impose strong assumptions on conditions of the treatment group before the treatment. The inferences based on the implicit evaluation designs also impose assumptions that an unspecified comparison group would experience no changes whatsoever due to the program. Under these strong assumptions, the program effects are measured as the difference between the average of the treatment groups after the program exposure and the assumed state of the treatment group before the program treatment. Implicit evaluation designs possess logically fragile structures but we are often forced to resort to the implicit evaluation designs in evaluation in practice. Many budgetary programs are designed and executed under the presumption that they would bring beneficiary effects to their customers. The majority of them provide the services to the people who can satisfy pre-specified conditions or criteria, which does not fit into the settings of the ideal evaluation design. In practice, most of the budgetary programs are introduced without due regard to the ideal evaluation design that requires randomized decision on program participation that program evaluators are forced to resort to implicit designs.

In the real world where there hardly exists such a thing as the ideal evaluation design posing various potential threats to the validity of causal inference, the evaluator needs to choose an evaluation design among three alternatives. In selecting an evaluation design, the main criteria the evaluator needs to considerate internal and external validity. Internal validity refers to the confidence the evaluator can put on the conclusions about what the program actually accomplished. A threat to internal validity occurs when the causal link between the program and the observed effects is uncertain due to some weakness in the evaluation design. External validity refers to the confidence one can have on the possibility of generalization of the evaluation results into other cases in terms of policy environment, time, and the customers. The external validity is threatened when the evaluation design does not allow causal inference about the program to be generalized to different times, places or subjects to those examined in the evaluation. Evaluators must ask themselves what sorts of decisions are likely to be made as a result of an evaluation, and be aware of the challenges to internal or external validity.

b. Randomized Experimental Designs

Experimental designs are the most rigorous approach available for establishing causal links between programs and their results. When successfully applied, they furnish the most conclusive evidence of program impacts. Unfortunately, experimental designs are impossible to implement for many government programs after the program has been running for some time. Nevertheless, they are important for two reasons. First, they represent the closest approximation to the ideal evaluation designs. As such, even when it is not feasible to implement an experimental design, less rigorous designs are often judged by the extent to which they come close to an experimental design. Second, in spite of the difficulties involved in conducting them, experimental designs can be and have been used to evaluate many programs.

Experimental or randomized designs are characterized by a random assignment of potential participants to the treatment and comparison groups to mimic the ideal evaluation designs as closely as possible by ensuring the equivalence before the treatment. They are experiments in the sense that program participants are chosen at random from potential candidates. There are a large number of experimental designs, three of which are discussed here; classical randomized comparison group design, post-program-only randomized comparison group design, and randomized block design.

The structure of classical randomized comparison group design is illustrated in <Table 2-2>, where "R" represents randomized assignments of the target population. The potential program participants from the target population are randomly assigned either to the treatment group that will receive program services or to the comparison group that will not. Measurements are taken before and after the program, and the net program effect is calculated as $[(T_1 - T_0) - (C_1 - C_0)]$. The change in the comparison group is subtracted from the change in the treatment group to measure the program effects, which is the origin of the name "difference-in-difference estimator". Randomization ensures that every member of the target population has a known probability of being selected for either the experimental or the comparison group. Then, the treatment and the control groups are mathematically equivalent. The expected values of T_0 and C_0 are equal. However, the actual pre-program observations may differ due to random chances that the evaluator cannot control, such as measurement error. As such, pre-program measurement allows for a better estimate of the net outcome by accounting for any accidental differences between the groups that exist despite the randomization process. In this design, the program treatment is the only difference, other than chance, between the treatment and the control groups.

Table 2-2 | Classical Randomized Comparison Group Design

	Before	Exposure to treatment	After
Treatment Group (R)	T_0	O	T_1
Control Group (R)	C_0	X	C_1

The main drawback of the classical randomized design is that it is subject to a testing bias. Testing bias indicates the fact that the pre-program measurement may affect the behavior of the treatment group, the control group, or both. This can potentially affect the internal validity of any causal inferences the evaluator may wish to make. To avoid the mishap, the evaluator may wish to drop the pre-program measurement as illustrated in <Table 2-3>. The design is almost equivalent to the ideal evaluation designs as long as the randomization is carried out in a proper way. And, it causes no problem to disregard the pre-program observations in measuring the program effects. However, one should keep in mind that, despite the randomization, it is possible that the two groups constructed will differ significantly in terms of the measures of interest and one cannot be completely certain of avoiding initial group differences that could affect the evaluation results.

Table 2-3 | Post-Program-Only Randomized Comparison Group Design

	Exposure to treatment	After
Treatment Group (R)	O	T_1
Control Group (R)	X	C_1

In order to minimize the sampling errors in the process of random assignment, the evaluator may want to construct as large a sample as possible. Unfortunately, this can be extremely costly. When the evaluator is not allowed to extract a sample large enough to ignore the problems related to sampling error, he may combine the randomization with blocking to address the problem. Blocking is involved in dividing the target population into several blocks and randomized selection of the treatment and the control groups could then be separately performed in each block. For example, if a social program affects urban and rural dwellers in a different way and the evaluator cannot afford to construct a sample large enough to disregard problems related to the sampling error, two blocks, an urban block and a rural block, can be formed. Randomized selection of the treatment and control groups could then be performed within each group. This process would help ensure a reasonably equal participation of urban and rural inhabitants. In fact, blocking should always be carried out if the variables of importance are known. Blocks can, of course, be matched on more than one variable. However, the number of blocks and ultimately the required sample size increase

very rapidly as the evaluator increases the number of variables characterizing blocks. That negates the initial motivation for blocking the population before randomization.

Randomized experimental designs provide the most rigorous way of conducting causal inferences on the impacts of the programs. Threats to internal validity are taken care of by making use of various tools such as control groups, randomization, and blocking. The primary difficulty with the randomized experimental designs is that they are often difficult to implement. There are cases where the evaluator cannot adopt a randomization device. When the entire target population participates in the program, there will be no basis for constructing a control group. In addition, it would be illegal or unethical to grant the benefit of the program to some people (treatment group) and withhold the same benefits from others (comparison group). Moreover, if the program has been under way for a significant amount of time, distinguishable differences may have already been firmly established between those who have benefited from the program, the treatment group, and those who have not, the comparison group. Randomized experimental designs are exposed to some threats to external validity. The difficulty of generalizing conclusions from a program evaluation is not automatically ruled out even in randomized experimental designs. Randomization for generalization purpose is a different issue from the random selection of the treatment and the comparison groups. In addition, several threats to internal validity still remain despite the random assignment of the samples. For instance, differential attrition rates of the treatment and the comparison groups would make bias in the initial randomization and diffusion of treatment between the two groups could contaminate the results.

c. Quasi-Experimental Designs

When randomization is not feasible, an alternative approach is to construct a comparison group that is similar enough in all important aspects to make valid inferences on the program effects. Quasi-experimental designs choose to use a non-randomized comparison group to make an inference with reasonable accuracy on program results under the environment, disallowing the evaluator to construct a comparison group through randomization necessary for exact inference. The comparison group in quasi-experimental designs could be either a constructed group, which was not exposed to the program, or a reflexive group, namely the treatment group itself before it was exposed to the program.

Three types of quasi-experimental designs are discussed here. They are pre-program/post-program designs, time series designs, and post-program-only designs.

There are two basic designs in pre-program/post-program designs. They are the pre-program/post-program non-equivalent comparison group designs and one group pre-program/post-program designs. The former uses a constructed comparison group and the latter does a reflexive comparison group.

The pre-program/post-program non-equivalent comparison group designs are structurally similar to the classical randomized comparison group designs in that they make use of both pre-program and post-program observations from both the treatment and the comparison groups. The comparison group is selected so that important characteristics of the group resemble those of the treatment group as closely as possible. The degree of similarity between the two groups is determined through the pre-program comparison. The accuracy of the inference based on this evaluation designs crucially depend on how similar the two groups before the program are in terms of those characteristics that are thought to affect the program results. A variety of methods to select a comparison group are available from simple comparison to highly sophisticated statistical mechanism. One statistical device drawing much attention in the evaluation literature is propensity score matching[5] which select the comparison group by comparing simulated probability of program participation. Unfortunately, it is extremely difficult to construct a perfectly matched comparison group showing similarity with the treatment group in all important variables. This means that at least one rival explanation for observed net program impacts will remain, namely that the two groups are unequal in some important aspects, to begin with.

One group pre-program/post-program designs take the difference between pre-program and post-program measurements on the treatment group as the program effects as shown in <Table 2-4>. The treatment group before being exposed to the program is taken as the comparison group. The lack of an explicit comparison group means that the internal validity of the evaluation design is seriously threatened. History may cause serious problems since the design does not control for events outside the program that may affect program outcomes. Moreover, the treatment group is fundamentally different from the entire population so that the measured program effects from the design tells us the program impacts on the people participated in the program rather than overall program impacts on the entire population. For example, even if we observe a significant decrease in unemployment among the participants of a job training program, we cannot reach a conclusion that the program has a positive effect on employment in the entire population. Those who voluntarily participated in the job training program could have found the job with higher probability than non-participants even without the training program. It is highly likely that participants in job training programs are more eager and willing to spend more resources to find a new job than non-participants. In addition, normal maturation of the program population itself may also explain any change. As well, the change may be an artifact when T_0 is unusually low by chance, so that (T_1 - T_2) is measuring the tendency to regress to the normal level rather than program effects. The design is also vulnerable to the problems of mortality of the sample, testing bias and instrumentation. The sole advantage of the design is its simplicity. If the

5. For detailed discussion on the issue, see Guo and Fraiser (2009).

evaluator can achieve enough control over external factors affecting the outcome variable, the design furnishes reasonably valid and conclusive evidence. Unfortunately, controlling the external factors is a very hard task to achieve in social sciences.

Table 2-4 | One Group Pre-Program/Post-Program Design

	Before	Exposure to treatment	After
Treatment Group (QR)	T_0	O	T_1

Time series designs are characterized by a series of measurements over time, both before and after exposure to the program. Any of the pre-program/post-program designs discussed above could be extended to the time series design. This means that time series designs possessing only a few before-and-after measurements are subject to all of the threats to internal validity the corresponding single measurement design faces. A more complete set of measures, on the other hand, allows the evaluator to eliminate many of these threats by analyzing pre- and post-program trends. Two time series designs, the basic time series designs and the time series designs with a non-equivalent comparison group, are discussed here.

In the basic time series designs, multiple numbers of before-and-after measurements on the treatment group are made as shown in [Figure 2-5] and the evaluator tries to identify the program effects by the change in the pattern of the time series covering both before- and after-the-program periods. In <Table 2-5>, $T_0(1)$ indicates the first measurement on the treatment group before being exposed to the program and $T_1(2)$ the second measurement on the treatment group after being exposed to the program. The evaluator examines the time series data, $T_0(1)$, $T_0(2)$, $T_0(3)$, $T_1(1)$, $T_1(2)$, $T_1(3)$ and tries to find any differences between before- and after-the-program observations.

With adequate time series data, this design can be fairly rigorous, ruling out many threats to internal validity, particularly maturation and testing effects. Still, the fundamental problem stemming from making inference based only on the treatment group remains. That is, the basic time series designs cannot eliminate the possibility that some factors other than the program treatment may have caused the change in the outcome variable.

Table 2-5 | Basic Time Series Design

	Before	Exposure to treatment	After
Treatment Group (QR)	$T_0(1)$, $T_0(2)$, $T_0(3)$	O	$T_1(1)$, $T_1(2)$, $T_1(3)$

Time series designs can be at least partially improved upon by adding a comparison group, typically non-equivalent one. Since both the treatment and comparison groups should experience the same external factors, it is unlikely that an observed change will be caused by anything but the program. As with any design using a non-equivalent comparison group, however, the groups must be similar enough in terms of the characteristics of interest. When this condition is met, historical designs could be quite useful contrary to what they seem. They could be robust to threats to the internal validity. This is true because, when properly carried out, a time series design allows for an assessment of the maturation trend before the program intervention. Time series designs can be used to analyze various time dependent program effects. The longitudinal aspect of the design can be used to address several questions whether the observed program effects are persistent or transitory and whether they are immediate or delayed. One serious problem with the design is that data requirement for the design is somewhat considerable and numerous data problems may exist. In particular, the time series available are often much shorter than those usually recommended for statistical analysis and different data collection methods may have been used over the period being considered.

Table 2-6 | Classical Randomized Comparison Group Design

	Before	Exposure to treatment	After
Treatment Group (QR)	$T_0(1)$, $T_0(2)$, $T_0(3)$	O	$T_1(1)$, $T_1(2)$, $T_1(3)$
Control Group (QR)	$C_0(1)$, $C_0(2)$, $C_0(3)$	X	$C_1(1)$, $C_1(2)$, $C_1(3)$

In post-program-only designs, measurements are taken after being exposed to the program. Two types of post-program-only designs are considered here, post-program-only with non-equivalent comparison group designs and post-program-only differential treatment s designs. The elements of a post-program-only with non-equivalent comparison group design are illustrated in <Table 2-7>. The difference between the treatment and the comparison groups, $(T_1 - C_1)$, is regarded as an estimate of the program effect on the outcome variable. Measuring only after being exposed to the program, the designs are free from the fear of testing and instrumentations biases. However, problems due to selection and mortality cause serious threats to the internal validity of the inferences. There is no way of knowing whether the two groups were statistically equivalent before exposure to the program. The quantity we take as the program effects, $(T_1 - C_1)$, may reflect the initial heterogeneity between the two groups rather than changes in the outcome due to the program. In addition, the estimate may exaggerate the true program effects if the attrition of the sample in the treatment group has occurred asymmetrically among the program participants who they think experienced no benefit from the program. This is called the

survivorship bias in the literature and may result in a biased estimate of the program effect even when the pre-program statistical equivalence between the treatment and the control groups are satisfied.

Table 2-7 | Post-Program-Only with Non-Equivalent Comparison Group Design

	Exposure to treatment	After
Treatment Group (QR)	O	T_1
Control Group (QR)	X	C_1

Post-program-only differential treatment designs measure the outcome variable for multiple treatment groups after the treatment. Each treatment group receives different level of treatment. This may be accomplished by providing differentiated amount of services to different groups classified according to a specific criterion such as region, income, and marital status. Even though the evaluator cannot make inferences on the program effects with the evaluation design, if sample sizes are large enough, sophisticated statistical analysis can uncover the relative performance of the program for each treatment group. For example, an estimate of (T_{B1} - T_{A1}) delivers the relative performance of the two treatment groups with different levels of treatments. Like other evaluation designs without the comparison group, the designs are exposed to the risk of selection and mortality threatening the internal validity of the inferences.

Table 2-8 | Post-Program-Only Differential Treatments Design

	Exposure to treatment	After
Treatment Group A (QR)	O	
Treatment Group B (QR)	O	T_{B1}
Treatment Group C (QR)	O	T_{C1}
Treatment Group D (QR)	O	T_{D1}

Having randomized experimental designs is an unusual luxury in social science. This often makes the quasi-experimental designs the best alternative the evaluator can use in practice. When the statistical equivalence of the treatment and the control groups cannot be established through randomization, the best approach is to use the stock of prior knowledge available to choose the quasi-experimental designs that are free from confounding effects as much as possible. The fundamental strength of the quasi-experimental designs is that they are cheaper and more practical than the randomized experimental designs since quasi-experimental designs do not require randomized construction of the treatment and the

comparison groups that demand high cost in terms of time, money, and efforts.

Quasi-experimental designs are likely to be vulnerable to threats to internal validity but the evaluator can overcome the difficulties by carefully constructing the comparison group. If the key variables of interest are identified and matched adequately between the treatment and the comparison groups, threats to internal validity are minimized. Surely, it is needless to say that more often than not it is impossible to match all variables of interest, especially when the number of variables of interest is large. Confronted with practical difficulties with randomized experimental designs, evaluators should look at the various quasi-experimental designs as alternatives and assess the risk factors in each type of design. The appropriate design will minimize the major risk factors, or at least allow the evaluator to account for their impact.

d. Implicit Designs

Implicit designs are probably the most frequently used evaluation designs, but are also least rigorous. Often, no reliable conclusions can be drawn from the design. However, an implicit design would be enough when the purpose of the evaluation is to logically examine whether the program has caused the outcomes. They are basically post-program designs with no control group as illustrated in <Table 2-9> where "I" indicates an implicit design. Neither can the magnitude of the program effect be estimated nor can anything definitive be concluded about the causal effects of the program. In its worst form, the design can be used to justify the program by conveying the information on the impression or feelings of the customers on the service provided by the program. Few will express a negative opinion for the service offered by a public agency as long as they do not have to pay for it explicitly. In spite of such obvious limitations, the evaluation designs enjoy popularity among a wide range of evaluators mainly due to the easiness in implementation. The design enables the evaluator to conduct the evaluation even when no pre-program observations are available or no control group exists. In such cases, implicit quasi-experimental designs could be possible alternatives. Three variations are discussed here; theoretical control group designs, retrospective pre-program measure design and direct estimate of difference design.

Table 2-9 | Implicit Design

	Exposure to treatment	After
Treatment Group (I)	O	T_1

Post-program-only with theoretical comparison group designs look like post-program-only non-equivalent control group designs as shown in <Table 2-10>. The only difference is that the values of the outcome variable for comparison group in the design, C_1*, is assumed

rather than measured or observed. The assumption on the state of the control group can be made with reference to theoretical arguments. For example, for a program to enhance the public's awareness of the harmful effects of caffeine, the evaluator can safely assume that the average level of awareness is negligible in the absence of the information provided by the program. As for public investment programs, the evaluator can assume that the social average rate of return on the equivalent private investments is, say 10%, against which the measured average rate of social return on the public investment projects can be compared.

Table 2-10 | Post-Program-Only with Theoretical Comparison Group Design

	Exposure to treatment	After
Treatment Group (I)	O	T_1
Control Group (I)	X	C_1^*

Post-program-only with retrospective pre-program measurement design looks similar to one group pre-program/post-program design in that it records the values of outcome variable both before and after being exposed to the program for each member of the treatment group. However, the records for both periods are taken after the program in the design and therefore the values for the outcome variable before the program are recorded retrospectively depending on the memories or assessments of the program customers.

In post-program-only with difference estimate designs, the evaluator directly asks the changes brought about by the program of the program customers, that is, members of the treatment group. The assessments are purely subjective and vulnerable to all kinds of risk factors, internal and external. It shares with the post-program-only with retrospective pre-program measurement designs a common feature that the pre-program states of the outcome variable are measured only after the provision of the service by the program.

Table 2-11 | Post-Program-Only with Difference Estimate Design

	Exposure to treatment	After
Treatment Group (I)	O	$(T_1 - T_0)$

In spite their structural weakness, implicit designs possess some strength. First, they are flexible, versatile and easy to implement. Because of the light requirements on data, implicit designs are always feasible. It is always possible to ask program participants, managers or experts about the performance of the program. This may also be a drawback in that easy implicit designs are often employed where, with more effort and ingenuity, more rigorous implicit or even quasi-experimental designs may have been possible. Second, implicit

designs can address virtually any issue and can be used in an exploratory manner. Program participants or managers can be asked any question about the program. While obviously weak in dealing with more objective estimates of program outcomes and attribution, an implicit design may well be able to answer questions about program delivery.

However, implicit designs offer little objective evidence on the program performance. Conclusions on the program effect drawn from implicit designs require major assumptions on what would have happened without the program, which makes the designs not robust to the threats to internal validity. Particularly, where attribution or incremental change is a significant evaluation issue, implicit designs should not be used alone. Rather, they should be used with multiple lines of evidence.

e. Causal Models in Evaluation Designs

Discussion up to now stressed the conceptual nature of the ideal evaluation design and its variants. The possible causal link between the program and the outcomes is isolated by utilizing two groups constructed by random assignments. An alternative way of addressing the causal problem in program evaluations is to make use of the causal model. The causal model can be represented as an equation that describes the marginal impacts of independent variables on the dependent variable. The simplest causal model is a linear model given as;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \gamma D + \varepsilon$$

where y is the outcome variable whose state we want to change through the program, (x_1, $x_2 \cdots x_k$)is the vector of explanatory variables that are thought to be affecting the outcome variable, D is the dummy variable taking the value 1 if participated in the program, 0 otherwise, and ε is the error term. All Greeks other than the error term are parameters to be estimated. Here, the key variable is the dummy for the program participation, and hence the parameter of interest is γ. If the estimate of γ is different from 0 with statistical significance, then the evaluator can conclude that the program is effective in changing the outcome variable[6]. In the experimental or quasi-experimental evaluation designs, the joint distribution of the explanatory variables is adjusted to be equivalent for the treatment and the control groups that the evaluator focuses on the program variable D, ignoring other explanatory variables that may affect the outcome variable. However, in the causal models, the program effect is only one of several independent variables that are expected to affect the program outcome. For instance, to examine the effect of an export support program, the evaluator may compare the export sales of firms receiving the program support and those of forms that are not. To account for the effects of various factors on the outcome variable, a

6. γ is the marginal effect of the program participation on the outcome variable. Note that $\gamma = \frac{\partial y}{\partial D}$.

causal model then may include variables such as the number of employees, capital intensity, industrial dummy, and dummy for program participation. Using the regression analysis, the evaluator could then determine the marginal impact of each of these variables on a firm's export sales.

In practice, most evaluators will want to use both causal and comparative approaches to examine program results. At the theoretical level, the two methodologies should yield the same answer on the program effects as long as the treatment and the comparison groups in the comparative approaches are constructed to guarantee that the joint distributions of explanatory variables in the causal models are equivalent in the two groups. Causal models are best suited to situations where sufficient empirical evidence has been accumulated, before the evaluation, on the existence of a relationship between the variables of interest.

One crucial problem with the simple causal model is that the dummy variable indicating the program participation is possibly correlated with the error term, resulting in an inconsistent estimator when the simple ordinary least squares estimation is employed. Efforts have been taken to overcome the problem in the evaluation literature and some of them will be discussed later.

3.2.4. Data Collection and Analysis

a. Data Collection Methods

We have already discussed several data collection methods when we discussed performance monitoring in the previous section. There are a few more methods of data collection particularly useful in program evaluation. Four of them, literature search, file review, natural observations, and case studies are explained here.

Literature search is the examination of previous program evaluations or academic research. It involves the examination of two types of documents. The first type consists of official documents, general research reports, published papers and books in the program area. Reviewing these documents, the evaluator examines the theoretical and conceptual issues related to the program under evaluation and considers the possible generalization of those issues to the questions related to the evaluation. Through literature search the evaluator may identify new evaluation questions and methodologies, thus leading to a potentially more effective evaluation. For example, if past researches on the industrial assistance programs reveal that the program's effect may be significantly different across firm sizes, the evaluator should elaborate a sampling scheme that ensures proper representation of all sizes of firms. The second type includes specific studies in the area of interest, including past evaluations on similar programs. The evaluator may want to compile and summarize previous research findings and use them as inputs in various stages of the evaluation study.

For instance, examining previous evaluation studies on industrial assistance programs, the evaluator finds data on employment in several areas benefitted differently from industrial assistance programs. Then, the evaluation design should incorporate the finding into the sampling scheme in which regions receiving a large amount of aid would serve as the treatment group, and regions receiving a small amount of aid would become the comparison group.

Literature search is a relatively economical and efficient way of collecting relevant data and has a high potential for payoff. It is also useful as a source of new hypotheses, to identify potential methodological difficulties, to draw or solidify conclusions, and as input to other data collection techniques. On the other hand, since data and information gathered from literature searches may not be relevant or compatible enough with the current evaluation issues, it is always important to note the compatibility or comparability between previous researches and current evaluation.

Like literature search, file review is relying on the examination of existing information to collect data useful for the evaluation work. It does, however, seek insight into the program the evaluator is currently investigating through past experiences. The documents reviewed include official and unofficial documents such as cabinet documents, departmental business plans or performance reports, reports of the audit office and records of departmental executive committee meetings. Administrative records are also an important area of file review. Examples include the size of the program or project, types of participants, experience of participants, post-project experience, costs of the program or project, and before-and-after measures of participants' characteristics. Project and program records such as critical events, project personnel, and events and alterations in project implementation should also be scrutinized.

File review is useful at least in three ways. First, review of general program files can provide invaluable background data and information on the environment surrounding the program, as well as the program. Second, review of individual or project files can indicate program results. For example, from the review of files of the international aid programs, the evaluator may find information on important quantities that may serve as inputs into the evaluation. Those include product-capital ratio, value added per unit of capital, productivity of capital, capital intensity, employment per unit of capital, value added per unit of total input, and various production functions. Although these measures do not directly assess program effectiveness, they are indicators that could serve as inputs into the evaluation. Third, file review may produce a useful framework and basis for further data gathering. File review also has limitations and problems. Program files are often incomplete or unusable. More often than not, a central filing system is relegated to a secondary position, containing brief memos from committees, agendas of final decisions and so forth. In retrospect, these

files tell an incomplete story. When researching the material that has given shape to a program, the evaluator may find that important information is held by separate individuals, instead of the central repository for program files. This can create serious problems. Experience suggests that once the project life-cycle moves beyond a working group's terms of reference, participating individuals will dispense with their files instead of keeping them active. More serious potential problem of file review is that it rarely yields information on control groups. To assess impact effectively, evaluators must have access to a control group of some sort. For a file review, this implies a requirement for file information about program participants before they entered the program, or information about non-participants. It is rare for such information to exist, except where an evaluation framework was approved and implemented beforehand. The lack of such data may make it necessary to collect new data, but these data may not be comparable with the original file data.

Observation involves selecting, watching and recording objects, events or activities that play a significant part in the administration of the program. The observed conditions can then be compared with some pre-established standards. In most cases, direct observation generally provides more powerful evidence than what can be obtained from secondary sources like a literature search or file review. In some cases, direct observation is an essential tool to understanding how the program functions. For example, a team evaluating customs clearance at airports might observe long lines of incoming passengers whenever two large airplanes arrive at the same time. Such peak-load problems would hinder the effectiveness of inspection, as well as the quality of service. Another example might be the case where dangerous chemicals were stored improperly, indicating hazardous working conditions for staff and a violation of health and safety regulations. Neither of these findings would have become apparent from examining written records only. Observational data describe the setting of a program, the activities taking place in the setting, the individuals participating in the activities and the meaning of these activities to the individuals. The method has been extensively used by many behavioral scientists such as anthropologists and social psychologists. It enables an evaluator to obtain data about a program and its impact holistically.

Observation provides only anecdotal evidence unless it is combined with other systematic scheme of data collection. Some first-hand observation can be justified in almost every evaluation, but it can be expensive since the collection of data by direct observations requires visits to program sites. Conducting direct observation, the evaluator may have the chance to find things that are missed or taken for granted by staff members of the program and find issues they are reluctant to disclose to outsiders. Most organizations involve routines which participants take for granted. Subtleties may be apparent only to those not fully immersed in these routines. This often makes it possible for an outsider, in this case the evaluator, to

provide a "fresh" view. Similarly, outsiders may observe things that participants and staff are unwilling to discuss in an interview. Thus, direct experience with and observations of the program will allow evaluators to gain information that might otherwise be unavailable. One important thing to note is that the reliability and validity of data collection through observations depend on the skills of the observer and on the awareness of any bias he or she brings to the task. Direct observation cannot be repeated. Different people carrying out a similar set of on-site observations may interpret the same phenomena differently, which is a fundamental limit to both the internal and external validity of the direct observation data.

When a program under evaluation consists of a series of projects or cases, the evaluator can extract very useful information on program performance through intensive examination of a small number of carefully selected projects or cases. Case studies are the only practical alternative an evaluator can choose when it is impossible, for budgetary or practical reasons, to take a sample large enough to guarantee reliability of statistical inferences. Case studies examine a number of specific cases or projects, through which the evaluator tries to reveal information about the program as a whole. Thus, selecting appropriate cases becomes a crucial step. The cases should be chosen so that the conclusions from selected cases can be generalized to the target population. Alternatively, some cases are chosen because it is considered critical examples, perhaps the purported best cases. If program outcomes in several critical cases turn out unsatisfactory, the effectiveness of the whole program might be seriously questioned, regardless of the performance of the program in other cases. The primary difficulty with the case studies is the lack of scientific ground for generalization. Case studies usually require significant resources and time, limiting the evaluator to small number of cases for in-depth analysis. Consequently, they are not normally expected to offer the results that can be generalized to other cases in the statistical sense. Rather, the main function of case studies is to provide a broader overview and insights into the process through which the program effects are realized.

b. Data Analysis Methods

The analytical tools for a program evaluation should be clearly decided at the evaluation planning stage. It is meaningless and not recommended to collect data before knowing how those data are to be used in the subsequent analysis. A coherent evaluation design should consider three things: the issues to be analyzed, the analysis methods and the data required to shed light on the issues. All of the pieces must fit together before the evaluation proceeds.

We discuss two ways of measuring the program effects by analyzing data, direct and indirect. The former tries to measure the program effects through direct measurement of various performance indicators and related variables, and the latter to infer indirect effects of a program based on the direct measurement.

a) Statistical Analysis

Statistics are used in a variety of ways in program evaluations. The manner in which program and pertinent contextual factors are measured affects the sorts of analytical techniques and statistical tools that are available for the evaluators. A key distinction affecting choices of statistics is the level of measurement used for coding the events of interest. For convenience, we can identify four levels of measurement; nominal, ordinal, interval, and ratio. Nominal and ordinal levels of measurement are inherently categorical, while interval and ratio variables reflect an underlying numeric continuum. Numeric distinctions are made with interval and ratio level variables that permit the values to be mathematically manipulated. Ratio measures differ from interval only in the assumption of a meaningful zero point. Nominal level measurement entails simply attaching numbers to data for purposes of assessing them to groups. Ordinal level variables differ from nominal level variables in that the categories of ordinal variables bear some ordered relationship to one another. For example, participants in a job training program might be identified at the end of the program as successful (completed the training and employed within four weeks of the program completion), partially successful (completed the training but unable to find a job after four weeks of the program completion), and unsuccessful (fail to complete the training program) with the distinction that ordinal variables are characterized by order, while nominal level categories serve only to differentiate the categories. Ordinal variables may play a key role in evaluation since ordinal attitudinal scales are typically used to measure perceptions of the program participants.

It is frequently the case that the selection of the appropriate analytical technique is virtually simple formality once the levels of measurement of the key variables in the analysis have been established. In practical application of statistical methods, other considerations, such as the audience's comfort level, also merit attention. Matching analytical techniques to the levels of measurement, audience, and evaluation questions is a big challenge for all evaluators.

When any phenomena are measured, the data can be tabulated according to a variety of procedures. If the resulting statistics, such as averages and medians, are used to describe a group of items, the figures presented are called descriptive statistics. In many situations, the population of program recipients is so large that to survey the entire population would be too costly. Instead, a sample is drawn from the population with the expectation that the quantitative results from the sample can be generalized into the population. To ensure that the statistics can be generalized with confidence, the manner in which the sample is drawn is of critical importance. If a group of units is selected in a systematic fashion such that the probability for each unit to be selected from the larger population is known, the group can be referred to as a probability sample. When statistics are computed from the sample with

the intention of generalizing from the sample to the population, the statistics are referred to as inferential statistics.

The accuracy of inferences is critically affected by the sampling procedures used. Four principles should guide evaluators when they select samples. First, the population of interest must be reasonably known and identified. This requirement presents a challenge for evaluators when records are not comprehensive. Therefore, evaluators should make efforts to ascertain whether the reason that records are not inclusive may be indicative of any bias. Second, a sampling technique should be used in which the probability for selecting any unit in the population can be calculated. Random sampling is the most recommended sampling technique. When there are specific subgroups within the population, the evaluators may divide the population into such subgroups and apply probability sampling techniques within each of the subgroups, an approach called stratified sampling. Third, an appropriate size of sample should be drawn to ensure the applicability of generalization. Fourth, even though probability sampling is applied, evaluators should examine a sample to ensure that it is truly representative of the population to which the evaluators hope to generalize on variables of critical interest. Probability sampling can help rule out chance variation that may conceal the true relationship or impede accurate identification of program effects, but it cannot be guaranteed that the sample contains certain units or phenomena in the same proportion as they exist in the population of interest.

To apply inferential statistics, a systematic procedure called statistical hypothesis testing should be adopted. First, a statistical hypothesis identifying the relationship among variables must be specified. Most of all, a null hypothesis is stated. The null hypothesis in program evaluation is that the program has no effect in achieving the desired outcome. For example, "access to home health aides does not affect medical costs for emergency care" might be a null hypothesis for an evaluation of a home health aid program. When the null hypothesis is not rejected, the sample data do not permit a conclusion that the program has had the measured outcome. When data are drawn to test the null hypothesis of no effect, if the program truly has no effect and the data support this, there is no problem. Similarly, if the program has the intended effect and the test data demonstrate this, again there is no problem. Problems arise when there is a discrepancy between the true situation and the test results. If the test result erroneously indicates that the desired outcome is achieved, an error called a false positive or type I error, is committed. On the contrary, if the test result erroneously concludes that the program fails to achieve the intended outcome, a false negative or type II error is committed. It is difficult to protect equally against both types of errors, so the cost of committing each type of error should be considered and attention paid to avoiding the more costly one. Any measurement precaution that helps protect the evaluator from committing the type II error increases the statistical power of the test. Once

the relative costs of committing two types of error are considered, evaluators can develop a decision rule that reflects the level of confidence they wish to have to generalize the existence of relationship found in the sample to the population. Since the probabilities of committing the two types of error are inversely related, the more evaluators protect against one type of error, the more vulnerable the test will be to the other type of error.

A quantified decision rule for specifying how much evidence is needed to generalize results also indicates how confident the evaluator wishes to be that the type I error occurs. This decision rule provides the confidence level for the test. The confidence level reflects the amount of evidence evaluators want to have to ensure that they are correct in concluding that the program does produce the observed effect. In social sciences, a 95% confidence level is conventionally used as a decision rule for testing statistical hypothesis, though alternative confidence levels such as 90% or 99% are also used. The null hypothesis to be tested is that the treatment does not have the intended effect. If the findings are sufficiently deviant from what the probability tables predict if the null is true, the null hypothesis is rejected. This decision allows one to generalize the program effects found in the sample to the population with the confidence that, over the long run, a test of this type should result in a type I error only five times out of one hundred in case of a 95% confidence level. When the null hypothesis is rejected, it is appropriate to state that the relationship in the sample is statistically significant at a confidence level of, say, 95%. This conclusion tells the audience that the relationship found in the sample reflects a real relationship in the population from which the sample is drawn.

When an estimate for the magnitude of a program effect is presented, it should be reported as a confidence interval, that is, the sample statistic should be stated with a margin of error for a given confidence level. Reporting an estimate of effect without a margin of error is not appropriate since it incorrectly implies too much precision in the estimate. The magnitude of the program effect should not be given too much confidence and reported to be falling within a range.

When evaluators wish to estimate or predict program effects by measuring the relationship between the alleged cause and effects, the manner in which the variables were measured limits the number of statistics appropriate for use. The most fundamental constraint is whether the variables were measured at the nominal, ordinal, or interval level of measurement. With nominal measures, contingency tables that array frequency counts are the most often used technique for analyzing data to assess the impact of one variable on another. In fact, if any of the variables of interest are nominal, contingency tables are the best option. With ordinal measures, contingency tables and frequency distributions are still the most popular choice for analysis. Some researchers prefer to treat ordinal measures as if they are equivalent to interval measures, and they choose analytical techniques typically reserved for interval

measures such as regression analysis, though this is not recommended for most of the cases. With interval measures, evaluators have the widest range of alternatives. When evaluators wish to explain an effect by other variables, regression is the most popular choice.

When multiple indicators have been used to measure a program effect, there are two basic approaches to reducing the data to a smaller number of factors; aggregating measures that are pre-specified to capture the program effect or using analytical techniques to identify patterns in the measure that indicate that there are observable patterns in the measure. When criteria for measuring a program effect are set for evaluators, the measures used can simply be aggregated. A summary index can be used that weighs different measures and then sums the total. When evaluators are unsure of what basic factor best expresses the program effect, they can use analytical techniques that sort through the indicators to identify co-movement that might permit the creation of indices. Factor analysis is the technique most frequently used for such data reduction purposes.[7] Sometimes evaluators may wish to sort units such as delivery sites into groups to identify characteristics of high or low performers. If the indicator on which the units are evaluated as low and high is known beforehand, discriminant function analysis can be used to identify the other characteristics of the units that will best predict which units will score high on the indicator. Discriminant function analysis is similar to regression analysis in that it identifies linear combinations of other (explanatory) variables that best predict the groupings of high and low performers. When the indicator on which units are to be disaggregated is not known beforehand, cluster analysis can be used to identify similar groupings. Cluster analysis differs from factor analysis in that the objective is to group objects rather than to identify groupings among variables. Characteristics of programs such as the level of administrative workload and other contextual characteristics might be used to identify clusters. An evaluator of an inter-jurisdictional program such as legal service to the poor might be interested in identifying clusters of offices that appear to operate under many of the same constraints. In this case, cluster analysis might be applied to identify characteristics that seem to differentiate most consistently across the offices.[8]

In addition to considering how statistics will be used in an evaluation, evaluators must consider other criteria when selecting a statistical technique. Sample size, for instance, may have a significant implication on the analysis. A small sample size may fail to produce a reliable conclusion on the program effect and preclude any further analysis of subgroup differences. In addition to the sample size, the number of observations recorded for the units of interest is pertinent to decision making regarding what statistical techniques are used

7. For a detail discussion on the statistical factor analysis, see a standard textbook such as Harman (1976).

8. See Hair, Black, Babin, and Anderson (2009) for more on factor analysis, discriminant function analysis, and cluster analysis.

in the evaluation. Before employing any statistical technique, evaluators should examine the distribution of the units along each of the variables or measures. Such basic frequency analysis will indicate how much the units vary on each of the variables. For example, if age is important in an analysis of the effects of a management training course on managers but a sample contains only twenty-eight and twenty-nine years old, the low variation on age rules out many analytical techniques. When a variable is measured at the interval level but the sample range is very narrow, the techniques the evaluator can choose are limited to those appropriate for ordinal variables. Similarly, if measurement was intended to be expressed in intervals but responses indicate that respondents could not make such fine differentiations, then techniques requiring interval measures are again ruled out. For example, survey questions asking researchers to report the percentage of their time devoted to research, administration, and teaching are intended to yield interval measures given in percentages. However, if almost all respondents respond "about half" or "about one-third" to these questions, this level of precision suggest that these variables should be analyzed as ordinal, not interval, measures.

b) Analysis of Qualitative Information

Most of qualitative data such as detailed descriptions of program administration, answers to the open-ended questions in surveys, transcripts of group discussions are the typical subjects of non-statistical analysis. Brief discussions on non-statistical analysis were already presented in the previous sections on data collection methods. The analysis of qualitative data can provide a holistic view on the phenomena of interest in the evaluation. The process of gathering and analyzing qualitative information is often inductive and naturalistic in that, at the beginning of data collection or analysis, the evaluator has no particular guiding theory on the phenomena under investigation.

Non-statistical analysis generally relies more on the evaluator's professional judgment than statistical analysis. Hence, evaluators should not only be knowledgeable about the evaluation issues, but should be aware of the various potential biases that could affect the results of the analysis.

Several types of non-statistical analysis exist, including content analysis, case analysis, inductive analysis, and logical analysis. All methods are intended to produce patterns, themes, tendencies, trends and motifs as well as interpretations and explanations on them. These analysis should also assess the reliability and validity of findings, possibly through the examination of several competing hypotheses. In addition, they may well analyze deviant or outlying cases

Several important decisions should be made in non-statistical data analysis. They include analytical methods, level of analysis, timing of the analysis, and the way to integrate non-statistical with related statistical analysis.

Like statistical analysis, most non-statistical analysis are typically executed after data collection is completed. But at least some part of non-statistical analysis can be carried out even during data collection, which may allow the evaluator to develop new hypotheses that can be tested later. It also permits the evaluator to identify and correct problems in the data collection process and to find information missing from early data collection efforts. On the other hand, conclusions based on early analysis may bias later data collection or may induce a premature change in program design or delivery, making interpretation of findings based on the full range of data problematic.

Conclusions based solely on non-statistical analysis may not be as accurate as conclusions based on multiple lines of evidence and analysis. Therefore, non-statistical data analysis should best be done in conjunction with statistical analysis on the related data. It should also be noted that the validity and accuracy of non-statistical analysis crucially depend on the skill and experiences of the evaluator.

c) Analysis of Long term Program Effects

The central interest of a program evaluation is to measure direct results of the program. More often than not, evaluators and stakeholders are interested in examining longer-term or broader impacts of the program. Recall that the results of a program can be categorized into three distinct types; outputs, intermediate outcomes, and end outcomes. Outputs are the results that a program produces and therefore are often operational in nature. Intermediated outcomes are those which are expected to lead to the ultimate objects desired but are not themselves the final objects. They include benefits to clients and sometimes unintended negative effects. The end outcomes are closely linked to the program objectives and usually to the broad benefits sought by the public agency operating the program. The general format to analyze longer-term program effects uses an established analytical model to trace those three categories of program results from outputs to intermediate outcomes and finally to end outcomes.

Good reading skills are generally presumed to result in better job opportunities that many governments institute many programs to promote reading skills. Consider the program that teaches reading skills to kids from poor families. The program logic can be illustrated as follows. A program to teach reading skills to poor kids may result in increased reading skills among those kids participating in the program and finally better employment prospects and higher income. Here, reading classes offered by the program are outputs and increase in reading skills the intermediate outcomes, and better job prospects and higher income are the end outcomes. An evaluation strategy to assess the incremental impacts of the reading program on reading skills of poor children starts with using an established model to transform the observed changes in reading skills among the kids participating in the

program into projected prospects of employment and income. That is, an increase in reading skills is translated into an increase in probability of employment or an increase in expected lifetime income, based on existing researches that relate reading skills to employment and income.

The evaluator may resort to direct assessment of the long-term effects rather than utilizing the established results. For example, the evaluator might use a quasi-experimental design to examine whether the treatment group has experienced better job opportunities or higher income relative to the control group. However, there are several reasons direct assessment is not desirable in measuring long term effects. First, it may take a very long time for program effects to be realized in full scale and it is extremely difficult to measure those long term effects using direct assessment with a given amount of time and resources. Second, if the effects of the program are realized over a wide range of program aspects, an evaluator may face difficulties in measuring the outcomes. Rather the evaluator may reduce the risk and increase the reliability of the evaluation by examining immediate outcomes or primary effects. Third, it is most likely that there already exist many researches on broader and long-term effects of the program and the evaluator can save a lot of time and resources by utilizing them.

c. Use of Various Models

A model is the theoretical and conceptual framework that explains the causal relationship between a program and its effects. Based on a model explicitly or implicitly, the evaluator reaches the conclusion that the program resulted in certain outcomes. There are several models frequently used in social sciences including simulation models, input-output models, economic models, and statistical models.

a) Simulation Models

A simulation model is a model that transforms inputs and outputs and can be a useful tool for evaluation purposes. Suppose a new set of questions are introduced to strengthen customs control at every highway entry points across the border. When it takes on average 10 more seconds to fill out the new set of questions than the previous set of questions, then the evaluator may assess its effects on the average waiting time of entrants at a border checking point.

A simulation model consists of three components; input data, a mathematical model, and output data. Simulation models are either stochastic or deterministic. Stochastic models generate probabilistic environment by incorporating random number generators. Simulation models become very popular among program evaluators due to the widespread utilization of various spreadsheet programs that provide both stochastic and deterministic simulation models.

Simulation models are used in combination with statistical techniques. For example, a model is constructed by regression analysis and then simulations are conducted using the model. In addition, simulation models are also utilized in a risk model based on a cost-benefit analysis. When the inputs are given in ranges or probabilities rather than numbers in a cost-benefit analysis, simulation models can be employed to produce the range or probability distribution of the variable of interest such as net present value. This information on range and probability can be very useful to a manager or an evaluator seeking to assess the risk or uncertainty surrounding a program.

b) Input-Output Models

An input-output model is a static economic model depicting the mutual interdependence among the different parts of an economy. It describes how an industry in the economy uses outputs from other industries as inputs to produce its own outputs, which subsequently are used as inputs for other industries. In other words, an input-output model can be used to derive internally consistent multi-sector projections of economic trends and detailed quantitative assessments of both the direct and indirect effects of any single program or combination of programs. Specifically, an input-output model can produce a detailed description of the way a government program affects the production and consumption of goods and services.

The production process in an economy is expressed in terms of technical coefficients and capital coefficients. The technical coefficient indicates the amount of goods and services including labor required to produce one unit of a certain product. The capital coefficient describes the amount of capital required to transform a proper combination of inputs into outputs.

The input-output models possess the fundamental limitation in that they are not effective in projecting policy effects in the future since they are descriptive and static one-period models. Another limitation is that the model may not reflect technical advancement or changes in relative prices since it is totally based on the past data. In addition, an input-output model may not be an adequate measure of the program effects when expenditures are done on a small scale like most public expenditure programs since it is essentially a macroeconomic model. Moreover, many unfortunate cases are found where input-out models are misused. In particular, in examining the effects of program expenditures in one sector, crucial mistakes are frequently made such as not taking into account possible offsetting effects generated by increased tax or government borrowing to finance the program.

c) Economic Models

Economic models are theoretical frameworks from economic discourse and can be divided into two categories, microeconomic and macroeconomic models. Microeconomic models describe economic behaviors of individual economic agents such as households and firms and are typically represented by equations describing the demand and supply functions for a good or service, which in combination deliver the equilibrium price and quantity.

Macroeconomic models are based on several important assumptions. For example, every firm is assumed to maximize profits given the prices of output as well as all inputs. Under these fundamental assumptions, microeconomic models can be used to optimal input combination to produce optimal level of outputs. The basic units in most programs are individual, household, or firm that microeconomic models can be very useful in analyzing program effects.

Macroeconomic models deal mainly with aggregate economic phenomena such as inflation, and unemployment. Various macroeconomic models attempt to explain and predict the relationships among these aggregate variables and can be used to predict the impacts of a program on the macro variables. For instance, in evaluating the impacts of export subsidy program on employment, the evaluator can measure the increase in employment due to export subsidy by utilizing a macroeconomic model that explains the relationship among various macroeconomic variables including export subsidy, export, and employment. Significant limitation exists in using macroeconomic models for program evaluation. Omission of an important variable from the model may lead to a misleading conclusion and uncertainty surrounding the conclusion is significant due to the fact that most inputs to macroeconomic evaluation models are frequently outputs from other macroeconomic models. Moreover, many macro-economic models have poor predictive capability, especially in the short run. They can be quite useful if the program effects are realized in the longer term or the size of program is large enough to have impacts on macroeconomic variables.

d) Statistical Models: Regression Models

Various statistical models can be used in evaluation studies. The simplest one is a tabulation of data for a single variable, which shows the frequency distribution of the variable of interest. Cross-tabulations can be utilized to convey similar information in case multiple variables are under investigation. It is always a good idea to summarize and report data in the form of cross tabulations because they are very transparent and easy-to-understand tools to which even decision makers with little statistical skills can easily access. Analysis of variance models are sometimes used to identify program effects when the evaluator faces small sample problems typically found in clinical programs in health or education.

Regression models are the most popular choice among program evaluators. Regression models are extraordinarily powerful tools in describing relationships among variables, test theories, and for making predictions with data from experimental or observational studies. There are various regression models available to program evaluators; linear or non-linear models, continuous or categorical dependent variable models, cross-sectional, times series, or panel data models, etc. The evaluator must select specific regression models that are appropriate to data structure and research questions.

Many practical questions involve the relationship between a dependent variable, y, and a set of k independent or explanatory variables, $(x_1, x_2 \cdots x_k)$ where scores on all variables are measured for N cases. In the evaluation setting, the dependent variable y measures the status of an outcome variable and the set of independent variables consists of all variables that are thought to play roles in determining the status of the outcome variable including the variable representing the program treatment. A linear multiple regression model can be expressed as;

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \qquad i = 1, 2, \cdots, N$$

One can construct an estimate of the parameters, $(\beta_0, \beta_1 \cdots \beta_k)$, using a standard statistical technique like the ordinary least squares. We consider three important examples of regression models in program evaluations.

Statistical examination of the equivalence of the two group means can be easily carried out with standard tools such as "t-test for mean comparison". Under the assumption that the two groups are constructed by random samples from the populations with normal distribution, one can construct the following t-statistic to check the equivalence of the means of the two groups;

$$t_{MD} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1}} + \sqrt{\frac{\hat{\sigma}_2^2}{N_2}}}$$

where $\bar{y}_i, \hat{\sigma}_i^2$, and N_i are the sample mean, sample variance, and the sample size of group i, respectively. Under the null hypothesis that the group means are equivalent, the test statistic, t_{MD}, follows the t-distribution with the degrees of freedom, $(N_1 + N_2 - 2)$. If the normality assumption is dropped, the distribution of the test statistic asymptotically approaches to the standard normal. The same kind of statistical inference on the group means can be carried out with a linear regression model specified by

$$y_i = \beta_0 + \beta_1 D_i + \varepsilon_i \qquad i = 1,2,\cdots,N$$

where D_i is the dummy variable taking the value 1 if the observation i belongs to group 1 and 0 otherwise. Taking conditional expectations on both sides results in.

$$E(y_i|D_i = 1) = \beta_0 + \beta_1 \text{ and } E(y_i|D_i = 0) = \beta_0$$

Therefore, β_1 represents the difference in mean values between members of group 1 $(D_i = 1)$ and $2(D_i = 0)$. The most popular estimator for the regression model above is the ordinary least squares (OLS) estimator and it is readily available from a standard statistical package like STATA or SAS. The test statistic is defined as

$$t = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$$

where $\hat{\beta}_1$ is the OLS estimate of β_1 and $s.e.(\hat{\beta}_1)$ is the standard error of $\hat{\beta}_1$. Under the null hypothesis that β_1 is not different from zero (no difference in group means), the distribution of the test statistic is given as the t-distribution with the degrees of freedom, $(N - 1)$. It is not difficult to show that the t-statistic is statistically equivalent to the test statistic, t_{MD}.

The previous test statistics for mean comparison assume the availability of the samples from classical randomized experiment designs. However, random samples are a rare luxury to social scientists and the best most evaluators can expect is the sample from quasi-experimental designs. Consider a quasi-experimental design where the comparison group is constructed based on the pre-program observational similarities on important variables that are thought to affect the outcome variables. We can capture the program effects by specifying the following regression model.

$$y_{i,t} = \beta_0 + \beta_1 D_t^1 + \beta_2 D_i^2 + \beta_3 D_t^1 D_i^2 + \varepsilon_{i,t}$$

where $y_{i,t}$ is the status of the outcome variable observed from individual i at time t. D_t^1 is the dummy variable taking 1 if the observation is taken after the program and 0 before the program. In addition, D_i^2 is the dummy variable taking value 1 if the observation is taken from the treatment group and 0 from the comparison group. Note that

$$\begin{aligned} DF_1 &= E(y_{it}|D_t^1 = 1, D_i^2 = 1) - E(y_{it}|D_t^1 = 1, D_i^2 = 0) \\ &= (\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_1) = (\beta_2 + \beta_3) \\ DF_2 &= E(y_{it}|D_t^1 = 0, D_i^2 = 1) - E(y_{it}|D_t^1 = 0, D_i^2 = 0) \\ &= (\beta_0 + \beta_2) - \beta_0 = \beta_2 \end{aligned}$$

DF_1 is the mean difference between the treatment and the comparison groups before the program and DF_1 measures the difference in expected values between the treatment and the comparison groups after the program. Then the quantity, $(DF_1 - DF_2) = \beta_3$, can be regarded as capturing program effects on the outcome variable after controlling for the time-invariant unobservable differences between the two groups. In other words, the estimate for the coefficient of the interaction term, $D_t^1 D_i^2$, in the regression model captures the program effects in a quasi-experimental design. The estimator, called a difference-in-difference estimator, is one of the most popular estimators in program evaluation studies if the comparison group is constructed from a quasi-experimental design. The usual significance test on the estimated coefficient, $\hat{\beta}_3$, will tell us whether or not the program has achieved the desired effects. The model can be easily extended to the case where observations on other important explanatory variables are available.

$$y_{i,t} = \beta_0 + \beta_1 D_t^1 + \beta_2 D_i^2 + \beta_3 D_t^1 D_i^2 + \gamma_1 x_{i,t}^1 + \gamma_2 x_{i,t}^2 + \cdots + \gamma_k x_{i,t}^k + \varepsilon_{i,t}$$

Including additional explanatory variables do not change the nature of the coefficient on the interaction term as the program effect and simple significant test on the estimated coefficient can be conducted to check the existence of the program effect. One supplementary note is that most inferences are based on asymptotic distribution of the test statistic rather than exact distributions like t-distribution since few statistical inferences rely on the normality of the error term in modern econometrics.

A fundamental assumption supporting the validity of the OLS estimator is that all explanatory variables are statistically orthogonal to the error term. In the nomenclature of evaluation literature, unobserved characteristics affecting the outcome variable should not be correlated with the participation decision. If the assumption is not satisfied, the OLS estimator for the program effect as well as all other coefficients is not consistent. In measuring the effect of a job training program on the probability of finding a new job, a typical approach is to specify a difference-in-difference model to capture the program effect along with several explanatory variables such as age, education level, sex, marital status and others. One important variable most models fail to consider in the specification is the innate productivity of the unemployed and therefore the effects of unobserved productivity of an unemployed worker are relegated to the statistical error term in the regression. If an unemployed worker choose to participate in the training program at his own will, it is highly likely that the unobserved productivity of an unemployed worker is correlated with the decision to participate in the program. That is, an unemployed worker with higher productivity may show the tendency to participate more in the training program than the one with lower productivity. If this is the case, the inferences based on the OILS estimator lead

us to the wrong conclusion on the program effects. Several advanced statistical techniques to overcome such problems are available in the evaluation literature[9].

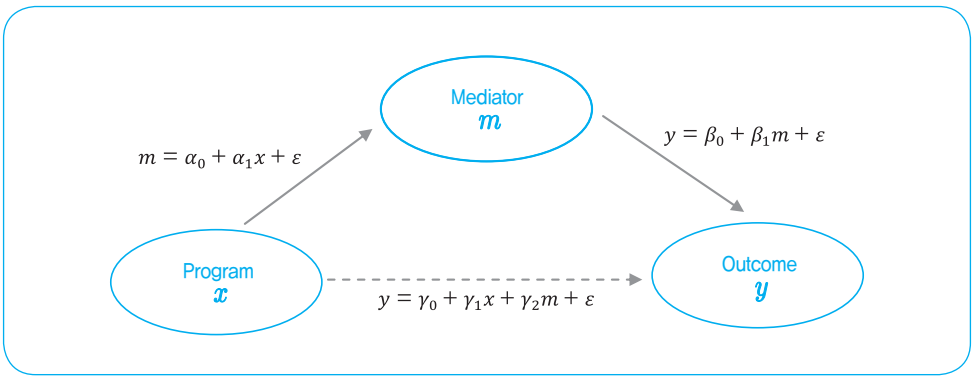
Regression models can be also used to describe and test conceptual models of how a program works, providing a useful framework for examining the validity of intervention logic. A simple test of the effects of a program might be focusing on the relationship between the level of program implementation (x) and an outcome (y). A more sophisticated theoretical analysis identifies the path or the logical sequence by which the program is presumed to have effects. This analysis can be conceptualized as a causal model where the program (x) has an impact of an intervening mediator variable (m), which in turn has an impact on the outcome (y), as shown in [Figure 2-7]. If the entire effect of the program operates through the mediator, the regression coefficient γ_1 is zero. If γ_1 is smaller than the regression coefficient δ_1 in the regression model, $y = \delta_0 + \delta_1 x + \varepsilon$, then the mediator m is said to partially mediate the effects of x on y. Mediation analysis can help us understand how programs work and guide development and modification of programs to make them more effective.[10]

The amount of mediation is measured by the difference between γ_1 and δ_1. This difference is also equal to the product of the paths to and from the mediator. Thus, $\gamma_1 - \delta_1 = \alpha_1\beta_1$, which can be rewritten as $\delta_1 = \gamma_1 + \alpha_1\beta_1$. We can interpret the relationship to indicate that the total effect of x on y can be decomposed into a direct component and an indirect component. A common test of statistical significance of the indirect component $\alpha_1\beta_1$ uses an approximation of the standard error of $\alpha_1\beta_1$ as $S_{\alpha_1\beta_1} = \sqrt{\hat{\beta}_1^2\left(s.e.(\hat{\alpha}_1)\right)^2 + \hat{\alpha}_1^2\left(s.e.(\hat{\beta}_1)\right)^2}$ where $\hat{\alpha}_1$ and $\hat{\beta}_1$ and their standard errors are taken from the regression models illustrated in [Figure 2-7]. The ratio $\frac{\hat{\alpha}_1\hat{\beta}_1}{S_{\alpha_1\beta_1}}$ is distributed approximately as a standard normal variable under the null hypothesis of no statistical significance of the indirect component. A more casual test that has a common application and intuitive appeal is simply to test the significances of both $\hat{\alpha}_1$ and $\hat{\beta}_1$. If both paths are statistically significant, we conclude that there is mediation.

9. For an introductory discussion on various statistical tools to overcome the problem, see Angrist and Pischke (2008).

10. For further discussion on mediation and moderation analysis in program evaluation, see Donaldson (2001).

Figure 2-7 | Mediation Model



d. Cost-Effectiveness and Cost-Benefit Analysis

A significant challenge for program evaluation is the comparison of total program costs to total program benefits. Developing accurate costs and relating those costs to specific measures of effectiveness or to total benefits, can greatly help decision makers but can prove difficult for the program evaluator.

Cost-effectiveness analysis identifies and provides information on the full costs of a program and relates these costs to specific measures of program outcomes, such as so many lives saved per unit of cost or the reciprocal, so many units of program cost per life saved. The users can then compare the cost-effectiveness of various similar programs to determine which program is most cost-effective, that is, which program cost less per unit of outcome or achieves the most outcomes per unit of cost. Cost-benefit analysis also identifies and provides information on the full costs of programs and further weighs those costs against the money value of all program benefits. The evaluator can then calculate the net benefits of the program, examine the ratio of benefits to cost, and determine the economic rate of return to society on the program investment. Users can then compare the program's benefits and costs with those of other programs or proposed alternatives.

These analysis can take place at different points in policymaking. As a program is being considered, an ex-ante analysis of costs and benefits can be evaluated to see if a program should be undertaken or to compare alternative prospective programs aimed at a common policy objectives. At any point during a project, costs and benefits can be compared. A current year or snapshot analysis provides data on whether the program's current benefits are worth the costs. Finally, an ex-post analysis provides decision makers with total program costs and benefits to date so they can evaluate a program's overall success. Each of these types of analysis has its usefulness, peculiarities, and issues. For example, in the analysis of

proposed programs, the estimation of costs and benefits is most difficult because they have not yet occurred. Those costs and benefits to attribute to the project is often challenging because the observed outcomes may have been the result of programs or events other than the one being analyzed.

a) Framework for Analysis

In conducting a cost-effectiveness or cost-benefit analysis as part of a program evaluation, whether *ex-ante* or *ex-post*, the first step is to identify all of the known costs and benefits of the program. There are several categories of costs and benefits, real versus transfers, direct versus indirect, and tangible versus intangible. For each benefit or cost, direct or indirect, it is important to clearly state its nature, how it is measured, and any assumptions made in the calculations of the monetary value involved. The statement of the assumptions is particularly critical because the decision maker needs to understand the analysis behind the numbers. Those assumptions need to be made clear to decision makers and also subject to a sensitivity analysis to determine to what extent the outcome of the analysis is controlled by certain assumptions made.

Real benefits and costs represent net gains or losses to society, whereas transfers merely change the distribution of resources within the society. Real benefits include money saved and money earned, lives saved and lives enriched, increased earnings and decreased cost to the taxpayers, and time saved and increased quality of life. In contrast, some societal gains are directly offset by other losses and are considered transfers. For example, a tax abatement program for the elderly may provide a tax-saving benefit to some but a cost to others in the form of higher taxes or lower services. Transfers also occur as a result of a change in relative prices of various goods and services as the economy adjusts to the provision of certain public goods. Transfers are often important to policymakers. Many government programs involve subsidization of one group by another and thus should be clearly identified where possible. But from an overall perspective, transfers do not increase total welfare so that they should be counted as neither benefits nor costs.

Real benefits and costs are those that are closely related to the primary objective of the program. Indirect or secondary benefits and costs are by-products, spillovers, or investment effects of the program. Direct costs include the costs of personnel, facilities, equipment and material, and administration. Indirect costs are intended and unintended costs that occur as a result of a program. For example, a dam built for agricultural purposes may flood an area used by hikers, which results in loss to them. This loss might be partially offset by new benefits to those using the lake created by the dam for recreation. In all cases, the benefits and costs must be traced to the program. For instance, if a claimed benefit is the creation of new jobs, the benefit is the new jobs created at the margin over what new jobs would have been created without the program.

An important distinction for the program evaluator is the difference between total and marginal benefits and costs. In assessing overall profitability of a program, an analyst needs to consider the total cost in getting the program started through its operational cycle. But at any point when the agency decides whether to continue or discontinue a program, it will consider marginal rather than the total costs and benefits. Marginal cost is defined as the incremental cost of producing one more unit of output. Analogously, marginal benefit is the incremental benefit generated by the last one unit of output. An analyst should discontinue the program of the marginal cost of the program exceeds the marginal benefit of the program. In considering a program, the evaluator must always start with status quo, that is, no change in the current level of expenditure for a program. In case of a new program, the evaluator should always contain a "do nothing" option to provide a baseline. So the only costs and benefits that would be considered are those that would occur in addition to those that would have occurred anyway without the program.

Fixed costs are the portion of the total costs that do not vary with the amount of services provided by the program and variable costs are those that vary depending on the size of the program. Sometimes it is useful to consider the distinction between fixed costs and variable costs. The distinction may be particularly important in sizing a program, as marginal benefits may increase or decrease with increasing program size.

Tangible benefits and costs are those that the evaluator can readily identify in unit terms and can convert to monetary units for cost-benefit analysis. In contrast, intangible benefits and costs include such things as the value of wilderness or an increased sense of community tie. It is especially difficult to place monetary value on many intangible benefits and costs. This is perhaps the most problematic area in cost-benefit analysis and why cost-effectiveness analysis might be more appropriate for some types of benefits and costs.

b) Estimating Costs

The first thing to do in cost-effectiveness or cost-benefit analysis is to account for all program costs. In examining various types of program costs, the evaluator often employs several broad categorizations. One-time or up-front costs such as expenditures involved in planning, R&D, pilot study, and computer software, ongoing investment costs such as costs of land, buildings and facilities, equipment, vehicles, and other expenditures with longer than one year life, recurring costs such as expenditures for operations, maintenance, personnel salaries, wages, fringe benefits, materials, supplies, and overhead, indirect costs including mitigation measures and compliance cost and costs to other government agencies or to the third parties.

Accounting or budgetary information should be the primary source for calculating various types of costs. Nevertheless, some costs will not be as easily identified from

program documents but should be developed using estimation methods, or shadow pricing. For example, in a dropout prevention program, the school providing services may use a dedicated classroom, whether during the school day or after school, there is no cash outlay for the school, but the classroom use would represent an opportunity cost. The use of the classroom for the program may mean it cannot be used for alternative academic activities. Then, the opportunity cost should be measured by the value of the best alternative use for the classroom used for the program. If the best alternative is to rent the classroom for other after-school activities, the opportunity cost would be measured by the rental income foregone. If the classroom would otherwise be vacant, the opportunity cost is zero.

The cost of capital assets should be spread out over their expected life. There are many standard depreciation schedules for buildings and other capital equipment. For government programs, an estimate needs to be made of the useful life of the asset considering physical deterioration, potential for obsoleteness, and salvage value at the closing of the program. In addition to depreciation, the government loses opportunity to use the money that is tied up in the un-depreciated parts of the assets. The opportunity cost is expressed as an interest rate times the un-depreciated portion. Land is consumed as other capital facilities and equipment, and it is not depreciated. However, it has alternative uses. Land dedicated to one activity cannot be used for another, and it cannot be sold to raise funds for other activities. Its value for a particular program is its opportunity cost, normally expressed as the market value of the land times the prevailing rate of interest cost for the government, such as yield on long-term treasury bonds. The cost of interest payments is sometimes counted as the program cost if the program required the issuance of debt to finance it. This is particularly true if the program is designed to be self-sufficient, with revenues paying for total costs. From a budgetary perspective, interest payments clearly constitute cost. However, if the evaluator is conducting a comparison of programs across jurisdictions, the inclusion of interest payments would give a faulty comparison of program efficiency.

Indirect costs are by-products of the program. They include intended items such as overhead and unintended items such as environmental costs. Many organizations employ a standard formula in allocating indirect costs on top of their direct cost, often computed at 30 to 60 percent of the total direct costs or a subset of direct costs. Government also uses a similar rule in allocating indirect costs and the specific figure is determined based on total administrative overhead costs compared with all other expenditures from all other programs. The major controversy with indirect cost allocation is whether a specific program really adds marginal cost to overhead agencies. That is, addition of a new program may not cause an increase in administrative cost since a large portion of overhead costs are sunk costs. However, an additional program sometimes does cause additional workload on some of those agencies that may lead to an increase in personnel and other administrative needs.

The application of the overhead rate is a significant judgment call for the evaluator and the results of the evaluation projects might be very sensitive to the treatment of the overhead costs. Government often shifts costs to the private sector, especially in regulatory activity. Though the shifted cost does not fall on the government, it should be counted as a part of indirect costs since it should be borne by somebody in the economy. Sometimes costs to the private sector are easy to identify, such as the increased cost to car manufacturers when a passive restraint system is required to be installed by a new regulation. At other times, regulations may impose additional reporting requirements, causing an increase in staff for the preparation of the reports or a loss of time to individuals who must wade through the additional bureaucratic red tape. These costs should be identified and valued in monetary value to be included as indirect costs. One other indirect cost of a program often excluded from identification of cost components is the cost to participants. Although these are not cash layout from the program agency's perspective, they should be considered costs of the program. For example, in the dropout prevention program offering an academy for at-risk students after school, program participants bear the costs in the form of the value of best foregone opportunity. Indirect costs to the private sector and to participants are controversial and their valuation sometimes problematic. Because of this, it is useful to separate costs to the government from costs to others, which may enable decision makers to determine the most important costs to consider.

Sunk costs are defined as investments that have been already made in a program but are not recoverable. In an ex-post evaluation of total costs of a program, the evaluator will consider all previous costs. However, when recommending future actions on a program, sunk cost should be ignored, because they have no impact on the marginal costs of the continuation of the program.

c) Cost-Effectiveness Analysis

Cost-effectiveness analysis related the cost of a given alternative to specific measures of program outcomes, for example, costs per life saved on various highway safety programs, and is sometimes regarded as the stepping stone to more elaborate work known as cost-benefit analysis.

The major advantage of cost-effectiveness analysis is that it frees the evaluator from having to express all benefits in monetary terms as required in cost-benefit analysis. Program outcomes can be addressed according to their multiple attributes. For an education program, for instance, a student learning can be assessed in terms of improved test scores, a physical education program can be assessed in terms of improvements in various physical skills of the participants, and programs to increase college placement can be assessed in terms of the number of students placed in colleges. In none of these cases does the evaluator

have to weigh the costs against monetary value of benefits. The evaluator simply presents the results to the decision maker, who then decides whether the various outcomes are worth the costs. This is often a very effective and inexpensive method of providing comparative program cost data to decision makers.

Many government programs, however, generate more than one type of cost and benefit. Comparison is unavoidable to reach a reasonable decision making. When conducting a cost-effectiveness analysis comparing programs having multiple objectives, the evaluator may need to assign weights on the relative benefits.

Two points should particularly attract attention from the evaluator in cost-effectiveness analysis especially compared to cost-benefit analysis. First, in considering programs with multiple benefits, unless the evaluator assigns weights to each benefit to obtain a common denominator for comparison purposes, the comparison may be of less use to decision makers. A cost-effectiveness analysis does not produce a single bottom-line number, with benefits exceeding costs or costs exceeding benefits. Thus, the final decision on the worthiness or desirability of a program should be made by the decision makers.

d) Cost-Benefit Analysis

Cost-benefit analysis is an economic technique that attempts to assess a government program by determining whether social welfare increases because of the program. Cost-benefit analysis provides information on all costs of a program and weighs those costs against the monetary value of all benefits generated by the program. Based on the information, the evaluator can calculate the net benefits of the program, determine the rate of return on the investment, and compare the program's net benefits with those of other programs. Cost-benefit analysis consists of three steps. First, the evaluator determines the benefits of a program and places monetary value on them. Second, the total cost of the program is calculated. Third, the evaluator compares the benefits and the costs. The information produced during the analysis can be used in various stages of program administration. In planning stages, cost-benefit analysis is utilized to inspect whether the program is expected to produce net social benefits and based on the information the decision on whether to continue the program is made. When the program has already started, the analysis can provide useful information to examine whether the program is producing net benefits as expected and to explore the way to improve the program's performance. Moreover, when multiple programs are proposed to achieve the same outcomes, cost-benefit analysis offers a methodologically sound tool to select the best alternative.

Both benefits and costs of a program occur for a long period of time after the initiation of the program. In order to compare cash flows in different points of time, costs and benefits at every point of time should be evaluated at the same point of time. The usual practice is to

evaluate all cash flows in terms of the value at the start of the program. In other words, all costs and benefits are discounted to be evaluated in terms of the present value. Net present value is the difference between the present value of all benefits from the program and the present value of all costs.

$$NPV = \sum_{t=0}^{T} \frac{B_t}{(1+r)^t} - \sum_{t=1}^{T} \frac{C_t}{(1+r)^t} = \sum_{t=1}^{T} \frac{(B_t - C_t)}{(1+r)^t} = \sum_{t=1}^{T} \frac{NB_t}{(1+r)^t}$$

where 0 is the starting time and T is the closing time of the program, B_t and C_t are benefits and costs of the program at time t, respectively, and r is the discount rate.

Positive net benefit means that the value of all benefits generated by the program exceeds the value of all costs expensed to carry out the program when all cash flows are evaluated in terms of present value. Therefore, it is beneficial to the society to carry out all programs with positive net benefit.

The evaluator may be confronted with a delicate issue when he simultaneously conducts cost-benefit analysis on multiple programs. Without budget constraints, accepting all programs with positive net present value is the optimal strategy in that net gain in social welfare is maximized by doing so. However, in practice, the government constantly faces a tight budget constraint that it is most likely that the total budget allocated to programs would run out before all programs with positive net present value are accepted. Then, the optimal strategy under budget constraint is to list all programs according to the size of net present value and then accept as many programs as possible starting from the program with the biggest net present value as long as the budget constraint allows. The decision rule, however, has a problem of favoring programs requiring large investment since larger programs tend to yield larger net present values. The situation is illustrated in <Table 2-12>. All listed programs yield positive net present value so that all should be accepted unless budget constraint is not binding. Facing budget constraint of 1,000, we should choose Program A if we follow the rule that programs with larger net present value should be accepted. Consider an alternative strategy of choosing Program B and Program C. The strategy is feasible with the total investment of 1,000 and yield net present value worth of 268, which is larger than net present value of Program A. Examining more carefully the figures in the table, one may notice an oddity that benefits per unit of investment is larger for Program B than Program A. In some sense, we may say that Program B is more efficient than Program A since we expect more benefits from B for the same amount of investment. Isn't it true that we should choose Program B rather than Program A even though the size of net present value is larger for Program A? Based on the example, we may argue that we should select programs according to the size of benefits per unit cost rather than the size

of net present value. The size of benefit per unit cost can be calculated by dividing present value of benefits with present value of cost - B/C ratio;

$$B/C = \left(\sum_{t=0}^{T} \frac{B_t}{(1+r)^t}\right) \Big/ \sum_{t=1}^{T} \frac{C_t}{(1+r)^t}$$

Table 2-12 | Cost-Benefit Analysis

	Cost	Benefit	Net Benefit	B/C
Program A	1,000	1,200	200	1.2
Program B	800	968	168	1.21
Program C	200	300	100	1.5

B/C ratio and NPV convey the same information since B/C ratio is greater than 1 if and only if NPV is positive. As far as a single program is concerned, the two rules lead to the same conclusion. However, in selecting a subset of programs from many potential candidates, we may want to apply the rule that programs with higher B/C ratios should be chosen over those with lower B/C ratios.

3.2.5. Reporting and Dissemination of Evaluation Results

Evaluations should be useful. The usefulness of an evaluation depends on its findings, conclusions and recommendations. It also crucially depends on the way findings, conclusions, and recommendations are reported and disseminated. Reporting is the process through which the evaluator transmits the findings and conclusions from the evaluation to the evaluation sponsors and other interested parties within the government. Dissemination refers to the activities through which information on the program produced by the evaluation is made available to all stakeholders including customers and the general public.

a. Maximizing the Use of Evaluation Results

In order to maximize the use of evaluation results, we need to pay close attention to the three important facts. First, the reports should be conformable to the needs of clients. Second, the reports should be presented in a timely manner. Third, the evaluator should seek to involve all stakeholders in the design of the evaluation.

The first requirement for the maximum use of evaluation results is that the reports should correspond to the needs of potential users of the evaluation. Program evaluations are conducted to improve the management the program, to enhance accountability of the program agency, and to assist the allocation of budgetary resources. If the primary objective

of the evaluation is to improve program management, the main target audiences of the evaluation reports are experts and specialists that the reports can be very brief and technical. However, it is always useful to provide a non-technical summary, perhaps written in a more discursive style, available for evaluation users who are not directly involved in program management and lack specialist knowledge on the program. Program evaluations carried out for the purpose of enhancing accountability or assisting the allocation of budgetary resources have more diverse potential users and the evaluation reports should be easy and non-technical to accommodate a relatively low level of knowledge of potential audiences.

The second requirement is to ensure the timeliness of evaluation reports. Evaluation reports should be completed in time to contribute to the material decisions on the program. This involves planning backward in time and making realistic projections of evaluation schedule.

Finally, one should seek to involve stakeholders in the design of the evaluation. The evaluator and the sponsors can increase the potential usefulness of an evaluation by ensuring wide participation in the evaluation design. The aim is not only to ensure sensitivity to the interests of different stakeholders, but also to make them aware of future plans for utilizing and disseminating the evaluation results.

b. Presentation of the Evaluation Reports

Evaluation reports should follow a logical structure. The terms of reference for the evaluation project usually specifies the structure of final reports. [Figure 2-8] illustrates an exemplary structure of an evaluation report. The evaluator should exercise the discretion in writing evaluation reports that they can accommodate the needs of the evaluation sponsors as well as the major stakeholders.

Though there is no universally accepted structure for the evaluation reports, it is nevertheless important that all reports contain an executive summary with reasonable length. While the executive summary is expected to feature towards the opening section of the evaluation report, it is also expected to be circulated as a separate stand-alone document.

In order for an evaluation report to be useful, it must be understood. This is the primary responsibility of the evaluator. The issues that should be clearly understood include the purpose of the evaluation, the subject of the evaluation, evaluation design, main findings and conclusions, and major recommendations. In addition, an evaluation report should provide sufficient information in an analytically rigorous way to constitute sound foundation for conclusions and recommendations. Moreover, an evaluation report should be comprehensible to a non-specialist with reasonable intelligence, which requires keeping technical language to a minimum and being friendly enough to readers by providing explanations on concepts and technicality.

It is likely that only a small proportion of the target audience will read the full report. It is therefore essential that the executive summary is well written. A frequent problem is that executive summaries are hastily prepared and so only give the reader a poor idea of the arguments and analysis contained in the main report.

Figure 2-8 | An Exemplary Structure of Evaluation Report

Title page:
- Title and nature of evaluation
- Title and duration of program
- Identification of author, date of submission, commissioning service

Table of contents:
- Main headings and sub-headings
- Index of tables of figures and graphs

Executive summary:
- Overview of the entire report with reasonable length
- Discussion of the strengths and weakness of the chosen evaluation design

Introduction:
- Description of the program in terms of needs, objectives, delivery systems etc.
- The context in which the program operates
- Purpose of the evaluation in terms of scope and main evaluation questions
- Description of other similar studies which have been done

Research methodology:
- Design of research
- Implementation of research and collection of data
- Analysis of data

Evaluation results:
- Findings
- Conclusions
- Recommendations

Annexes:
- Terms of reference
- Additional tables
- References and sources
- Glossary of terms

c. Dissemination of Evaluation Results

Dissemination encompasses the whole range of activities by which the information contained in evaluation reports is made available to a wide range of audiences. The list of wider audiences of an evaluation report include various stakeholders such as key policy makers, program and evaluation sponsors, program customers, program management, other interest groups, and the academic community.

Given the diversity of potential audiences, it is very important to convey the evaluation findings in a way appropriate to each audience. Aside from circulating the full report, communication can take place through the circulation of the executive summary or through oral presentations based on audio-visual material. When evaluators or sponsors wish to ensure dissemination of the information derived from an evaluation other than through distributing the report itself, their most important task is to target the presentation to match the audience. Different audiences are likely to react in different ways to a presentation of evaluation findings. Program customers present particular problems. They are often unorganized and geographically segmented. In the case of some programs, beneficiaries may even be unwilling to identify themselves. Where they do make their voices heard, it may be through organizations which purport to represent their interests.

Finally, it is important to remember that conflicts of interest are, to some extent, inevitable where there is a multiplicity of stakeholders. Conflicts of interest can best be tackled at the outset by having an inclusive management structure. By clearly separating findings, conclusions and recommendations, the evaluator can draw a line between the evidence that was found about a program and his own personal opinions. Thus, even if some stakeholders choose to reject certain recommendations, they may be less inclined to dispute findings and conclusions. Moreover, program managers can formulate their own observations on reports prepared by external experts. In addition, by no means should evaluators become entangled in negotiations. The professional expertise and conscience of an external evaluator should be a sufficient guarantee for the impartiality and credibility of his findings and conclusions.

2012 Modularization of Korea's Development Experience
Performance Management System of Budgetary Programs in Korea

Chapter 3

Performance Management Systems in the World

1. The United States of America
2. The United Kingdom
3. Australia
4. Japan

Performance Management Systems in the World

1. The United States of America

1.1. Brief History of Performance Management in the U. S. A.

In the United States, there have been many attempts to link budget to performance in the public sector since the Hoover Commission proposed a performance budget system in 1949. In 1950, the Budget and Accounting Procedure Act was enacted, and it was mandated to submit the budget bill drafted based on functions and activities. The Planning Program Budget System (PPBS) was enacted in 1965 by President Johnson, the Management by Objective (MBO) in 1973 by President Nixon was implemented to enhance the management of federal governments, and Zero-Based Budgeting (ZBB) was enacted in 1977.

Performance budget system started to be implemented in 1993 with the introduction of the Government Performance and Results Act (GPRA). Based on GPRA, government organizations submitted their annual performance plans from 1997 to early 1999 and performance reports in 1999 to the Office of Management and Budget (OMB). Existing performance reports imposed too much of a burden on program agencies by requiring excessive reporting on the financial aspects of government activities and performance. The Reports Consolidation Act of 2000 intended to enhance the utilization of performance data, requiring program agencies to report only items pertaining to program performance. Therefore, government organizations conferred with OMB to set the date and frequency of reporting and submitting the annual report containing performance information as well as financial data. This report was to be submitted within 150 days after the end of corresponding fiscal year, and it must include the evaluations of inspecting officers on the organizations' efforts to resolve management issues and performance problems in hand.

The organizations are also obligated to explain the irregularities and reliability problems with data in the report and propose remedial measures.

In 2001, the Bush administration, announcing the Presidential Management Agenda (PMA), laid out plans to reform into a results-oriented government. Budget and Performance Integration was one of the most important agendas in the plan. Program Assessment Rating Tool (PART) was introduced to utilize performance records and performance information of GPRA in the budget process. PART is composed of four parts; program's purpose and design, strategic planning, program management, results and responsibility of the program. It required departments to evaluate their own programs with check-lists containing thirty questions and OMB uses these results for improving a program's performance and for the budget allocation process. Feedback from PART, implemented to enhance performance management of government agencies and to utilize the results of the budget process, turned out ineffective mainly because it was *ex-post* assessment of performance and effective feedback was difficult. Moreover, it was not easy to find measures to link budget allocation with the actual performance, and the performance results could not be applied effectively in the budget process because of the Congress' indifference to performance information.

The Obama administration, inaugurated in 2009, aimed to improve timely performance management by enhancing program management processes, paying attention to the deficiency of the previous regime's PART from the Bush administration. Performance management system of the Obama administration consists of four parts. The first is to introduce the system which manages and sets High Priority Performance Goals (HPPG) to attain a high level of performance achievement. The second is to conduct the in-depth program evaluation on on-going programs and utilize the results in restructuring of government expenditure programs. The third component is to conduct cost-benefit and cost-effectiveness analysis and use the results in budget allocation among competing programs. The fourth is to improve competency of public sector employees to accomplish better performance results of government programs.

1.2. Performance Management System in U. S. A.

1.2.1. Institutions

On both the federal and state level, the United States government possesses a wide range of power and responsibility in the budget process. The Constitution of 1787 clearly states that the Congress possesses the power to collect taxes and annual government expenditure must be approved by the law enacted by the Congress in order for the revenue to be expensed from the exchequer. The Constitution, however, does not clearly state the president's role or

the legislative power of the Congress on the budget. The gap has been bridged by various supplementary laws since then.

There are two kinds of laws regarding the budget system - process-oriented, and performance (or results) - oriented. Most of the process-oriented laws specify players and their roles participating in the budget process. For example, the Budget and Accounting Act of 1921 requires the president to obtain annual congressional approval on the federal budget. OMB was established to assist the president to review budget demands of each public agency. GAO was also established by the same law and the institution is responsible for the audit and investigation independent of party influences. Congressional Budget and Impoundment Control Act of 1974 includes the process of budget resolution in the Congress, the establishment of the Congressional Budget Office (CBO), and restrictions on delaying or abolition of expenditures by the administration. In addition, the Inspector General Act of 1978 and Federal Manager's Financial Integrity Act 1982 were enacted to improve financial management of the state, and in 1993, GPRA was introduced to improve the performance of federal government.

The United States have a two-tier political system, federal and state. Each of the 50 states possesses its own constitution and is integrated into the federal system with a separate federal constitution. State governments are not affiliated to the federal government, but maintain independent status. The president of the United States, as the head of the administration, with a four-year term to manage the administration, has prerogative of commander-in-chief, power to appoint cabinet members, and authority to ratify treaties. The United States federal government consists of 15 administrative departments, and each secretary is appointed by the president and approved by the Congress. The legislature is divided into the Senate and the House of Representatives, and has six year and two year terms each. The Congress has power to impose and collect taxes and duty, budget decision, and enactment of statutes. Congressional committees related to financial matters are the Ways and Means Committee, Finance Committee, and Appropriations Committee. States governments also have administrative, legislative, and judicial branches.

1.2.2. Legal Framework for Performance Management

In the United States, GPRA, enacted in 1993, is thought to be the starting point of the performance budget system. During the Bush administration, the Federal Program Performance Standards and Goals Act were proposed. After many reviews and hearings, the bill was renamed GPRA and passed both chambers of the Congress, enjoying bi-partisan support.

The fundamental objective of GPRA is to enhance the efficiency and effectiveness of federal programs by establishing a system which sets performance objectives and evaluates

the results. GPRA mainly comprises a strategic plan, the performance plan, and the program performance report. Federal organizations must submit a strategic plan to Congress and the OMB. Strategic plan should include a five-year plan and be renewed every three years. More specifically, the plan should include broad and inclusive descriptions on core functions of the federal organization in addition to goals and objectives as well as the means to attain them. In drafting the strategy, the federal agencies and organizations must consult the Congress and incorporate opinions and suggestions from stakeholders. Annual performance plan is required to specify the objective and measureable performance indicators as well as target levels for performance indicators that should be achieved in the corresponding year. Annual program performance reports are the vehicle through which program performances are reported to the President and the Congress. The reports should include the results of comparison between the target level and the measured performance of the programs. They should also offer the explanation on reasons of unsatisfactory performances and necessary measures to improve it.

In 2009, President Bush announced the Presidential Management Agenda consisting of five key tasks to promote management innovation of the administration and improve the performance of federal programs. One of them was the integration of budget and performance. As GPRA was thought to have limitations in providing information regarding fiscal management and budget allocation, PART was introduced as a diagnostic tool to improve program performance and link the performance with the budget allocation process. While GPRA institutionalized the performance budget system through legislative measures, PART was an administrative initiative under the OMB to reinforce GPRA. PART consists of a series of questions to evaluate execution and performance of government programs. The purpose of PART is to provide performance information based on which a consistent budget allocation process is carried out rather than manufacturing new performance data. Decisions related to the government budget are not made automatically based on PART, but in consideration of policy priority and other factors. In most cases, measuring the effects of additional budget on program performance is generally inaccurate and meaningless. Especially, tying short-term performances to the budget can be problematic. Therefore, a program achieving good results did not necessarily receive favors in budget allocation.

1.2.3. Elements of Performance Management System

a. Government Performance and Results Act (GPRA)

The fundamental purpose of GPRA is to improve efficiency and effectiveness of government programs by establishing a system in which the mission or objectives of the program is clearly defined and performances are accurately measured. GPRA proposed a

performance management system consisting of three components; strategic plan, annual performance plan, and annual program performance reports.

According to GPRA, each organization and agency of the federal government must report respective strategic plans to OMB and the Congress and the plan should include a five-year plan that should be revised every three years. Strategic plan must also include organizations' general and specific objectives of each federal organization and measures to achieve them along with an overall framework to execute the plan. It should be remembered that drafting the strategic plan, each federal organization must confer with congress and make efforts to incorporate opinions or suggestions of stakeholders into the plan. Annual performance plan is a report on the next year's performance plan and must be filed every year. The report should specify performance goals that are quantitatively measurable. In addition, it should describe human and material resources required to accomplish the performance goals, and provide the ground on which a prescribed level of performance goals and the results of the program are compared. Annual program performance reports compare the performance goal in the annual performance plan with measured performances of the program. In addition, they should also offer the explanation on reasons of unsatisfactory performances and necessary measures or alternative methods to improve the results.

Drafting of these three reports is closely related to the public interests and considered intrinsic to the function of the government so that it is not a good idea to commission the work to external experts and should be carried out by the members of the organization in charge of the program. <Table 3-1> compares the three reports of the GPRA.

Table 3-1 | Three Reports in GPRA

	Strategic plan	Annual performance plan	Annual program performance reports
Frequency	For next five years Revised at least every three years	Annually, on next year	Annually, on next three years
Contents	Mission statement on organization's major functions and tasks Strategic goal Resources, means to achieve the goal, and procedures of the work Relationship between strategic goal and annual performance plan Uncontrollable external factors affecting achievement of the goal Explanation on previous program evaluations consulted in setting and revising strategic goal	Program's performance goals Definition of performance goal in objective, quantitative, and measurable manner Resources, means to achieve the goal, and procedures of the work Definition of performance indicators to measure and evaluate the outcome and results of service standard Standards for comparison between the actual project outcome and goals Explanation on method used to check validity of measured value	Explanation on success or failure of performance goal Evaluation of current year's performance plan based on past records Explanation on goals not achieved and future plans to achieve them Suggest reason and alternative policy when the goal is unrealistic or infeasible Summarization and explanation of current year's evaluation of programs

b. Performance Improvement Initiative (PII)

PII was suggested as an important component in the Bush administration's PMA. It aims to achieve the most cost-effective outcomes. It was designed to overcome issues raised in GPRA. In particular, the target was set at too low of a level and utilization of performance information was extremely unsatisfactory. Federal organizations and OMB exchanges a series of discussion on which programs showed unsatisfactory performance and what measures should be taken to improve it. Then they redirect budget resources from ineffective programs to more effective ones. Even though the final decisions on programs are made in the Congress and high ranking officials in the administration, the performance information facilitates decision makings for both administrative and legislative branches. PII conducts cost analysis to produce information on efficiency used in decision makings.

PII's performance can be assessed based on two criteria. Improvement of program outcomes is the first criteria. PII is expected to contribute to improving program outcomes through an active and flexible accommodation of the results from performance evaluations.

PII requires each organization to identify its weaknesses in planning and management of programs and revised the plan to manage the organization and programs more effectively. The second criterion is the strength of the link between budget and performance. The efficient use of limited resources requires that programs with higher performance should attract more resources and efforts should be made to induce better results from essential programs with unsatisfactory performances. Budget allocation does not totally depend on performance but the performance information provided by PPI can be used as a useful guiding stick in discussion among decision makers.

Assessment on the first aspect of PII is generally positive. Federal programs became more effective and efficient through the execution of the improvement plan developed in conjunction with OMB. For example, the Social Security Administration was able to improve its productivity in 2007 by 15.5% compared with 2001 through an improvement in information technology and the program process. Without improved productivity, additional 980 million dollars would have been required to perform the same tasks. Also, the Administration of Aging expanded its services for senior citizens suffering from diseases and disabilities. In 2006, 18 states expanded their support for senior citizens under the poverty line to offer program services to 80,000 more seniors. This enabled more than 345,000 disabled senior citizens, 52,000 more compared to 2003, to receive in-home treatment instead of being sent to sanatoriums. Assessment on the second aspect is rather mixed, but the administration showed a more active attitude in reallocating budget to better performing programs. In 2008, seven programs were terminated and six were scaled down due to unsatisfactory performances, which saved almost 1.3 million dollars.

Experience from PII identified four important factors in maximizing a program's performance; regular performance evaluation based on PART, issuance of scorecard[11] for each federal organization to ensure the responsibility for PART results, announcement of evaluation results to all stakeholders and the general public, efforts to improve performance of inter-agency programs.

c. Program Assessment Rating Tool (PART)

PART is a diagnostic tool developed by the OMB in 2004 to evaluate the performance of programs and systematically utilize the results in the budget process. For each target program under evaluation, PART asks 25 common questions and an additional 20 questions depending on the types of the program. Common questions are divided into four categories; program purpose and design, strategic planning, program management, and program results

11. Each federal organization is assessed on a quarterly basis under PII and assigned the scorecard summarizing its performance with three different colors; green, yellow and red. It is known that the announcement of the overall performance of an organization through scorecard has been very effective in ensuring accountability of the organization.

and accountability. PART divides all programs into seven categories for the purpose of asking additional questions unique to a particular type of program. These types include Direct Federal Programs,[12] Competitive Grant Programs,[13] Block/Formula Grant Programs,[14] Regulatory-based Programs,[15] Capital Assets and Service Acquisition Programs,[16] Credit Programs,[17] and R&D Programs.[18]

The list of common questions includes;

Program purpose and design

- Is the program purpose clear?
- Does the program address a specific and existing problem, interest or need?
- Is the program designed so that it is not redundant or duplicative of any other federal, state, local or private effort?
- Is the program design free of major flaws that would limit the program's effectiveness or efficiency?
- Is the program effectively targeted, so that resources will reach intended beneficiaries and/or otherwise address the program's purpose directly?

Strategic planning

- Does the program have a limited number of specific long-term performance measures that focus on outcomes and meaningfully reflect the purpose of the program?

12. Programs where services are provided primarily by employees of the federal government, like the State Department's Visa and Consular Services program.
13. Programs that provide funds to state, local and tribal governments, organizations, individuals and other entities through a competitive process, such as Health Centers at the Department of Health and Human Services (HHS).
14. Programs that provide funds to state, local and tribal governments and other entities by formula or block grant, such as the Department of Energy's (DOE) Weatherization Assistance program and HHS' Ryan White/AIDS program.
15. Programs that accomplish their mission through rule making that implements, interprets or prescribes law or policy, or describes procedure or practice requirements, such as the U.S. Environmental Protection Agency's Mobile Source Air Pollution Standards and Certification program.
16. Programs that achieve their goals through development and acquisition of capital assets (e.g. land, structures, equipment, and intellectual property) or the purchase of services (e.g. maintenance, and information technology). Program examples include Navy Shipbuilding and the Bonneville Power Administration.
17. Programs that provide support through loans, loan guarantees and direct credit, such as the Export Import Bank's Long Term Guarantees program.
18. Programs that focus on knowledge creation or its application to the creation of systems, methods, materials, or technologies, such as DOE's Solar Energy and NASA's Solar System Exploration programs.

- Does the program have ambitious targets and timeframes for its long-term measures?
- Does the program have a limited number of specific annual performance measures that can demonstrate progress toward achieving the program's long-term goals?
- Does the program have baselines and ambitious targets for its annual measures?
- Do all partners (including grantees, sub-grantees, contractors, cost-sharing partners, and other government partners) commit to and work toward the annual and/or long-term goals of the program?
- Are independent evaluations of sufficient scope and quality conducted on a regular basis or as needed to support program improvements and evaluate effectiveness and relevance to the problem, interest, or need?
- Are budget requests explicitly tied to accomplishment of the annual and long-term performance goals, and are the resource needs presented in a complete and transparent manner in the program's budget?
- Has the program taken meaningful steps to correct its strategic planning deficiencies?

Program management

- Does the agency regularly collect timely and credible performance information, including information from key program partners, and use it to manage the program and improve performance?
- Are federal managers and program partners (including grantees, sub-grantees, contractors, cost-sharing partners, and other government partners) held accountable for cost, schedule and performance results?
- Are funds (federal and partners') obligated in a timely manner and spent for the intended purpose?
- Does the program have procedures (e.g. competitive sourcing/cost comparisons, IT improvements, appropriate incentives) to measure and achieve efficiencies and cost effectiveness in program execution?
- Does the program collaborate and coordinate effectively with related programs?
- Does the program use strong financial management practices?
- Has the program taken meaningful steps to address its management deficiencies?

Program results and accountability

- Has the program demonstrated adequate progress in achieving its long-term performance goals?

- Does the program (including program partners) achieve its annual performance goals?
- Does the program demonstrate improved efficiencies or cost effectiveness in achieving program goals each year?
- Does the performance of this program compare favorably to other programs, including government, private, etc., with similar purpose and goals?
- Do independent evaluations of sufficient scope and quality indicate that the program is effective and achieving results?

The lists of specific questions by program types include;

Competitive grant programs

- (program management) Are grants awarded based on a clear competitive process that includes a qualified assessment of merit?
- (program management) Does the program have oversight practices that provide sufficient knowledge of grantee activities?
- (program management) Does the program collect grantee performance data on an annual basis and make it available to the public in a transparent and meaningful manner?

Block/formula grant programs

- (program management) Does the program have oversight practices that provide sufficient knowledge of grantee activities?
- (program management) Does the program collect grantee performance data on an annual basis and make it available to the public in a transparent and meaningful manner?

Regulatory-based programs

- (strategic planning) Are all regulations issued by the program/agency necessary to meet the stated goals of the program, and do all regulations clearly indicate how the rules contribute to achievement of the goals?
- (program management) Did the program seek and take into account the views of all affected parties (e.g., consumers; large and small businesses; State, local and tribal governments; beneficiaries; and the general public) when developing significant regulations?

- (program management) Did the program prepare adequate regulatory impact analysis if required by Executive Order 12866, regulatory flexibility analysis if required by the Regulatory Flexibility Act and SBREFA, and cost-benefit analysis if required under the Unfunded Mandates Reform Act; and did those analysis comply with OMB regulations?
- (program management) Does the program systematically review its current regulations to ensure consistency among all regulations in accomplishing program goals?
- (program management) Are the regulations designed to achieve program goals, to the extent practicable, by maximizing the net benefits of its regulatory activity?
- (program results and accountability) Were programmatic goals (and benefits) achieved at the least incremental societal cost and did the program maximize net benefits?

Capital assets and service acquisition programs

- (strategic planning) Has the agency/program conducted a recent, meaningful, credible analysis of alternatives that includes trade-offs between cost, schedule, risk, and performance goals and used the results to guide the resulting activity?
- (program management) Is the program managed by maintaining clearly defined deliverables, capability/performance characteristics, and appropriate, credible cost and schedule goals?
- (program results and accountability) Were program goals achieved within budgeted costs and established schedules?

Credit programs

- (program management) Is the program managed on an ongoing basis to assure credit quality remains sound, collections and disbursements are timely, and reporting requirements are fulfilled?
- (program management) Do the program's credit models adequately provide reliable, consistent, accurate and transparent estimates of costs and the risk to the Government?

R&D programs

- (strategic planning) If applicable, does the program assess and compare the potential benefits of efforts within the program and (if relevant) to other efforts in other programs that have similar goals?

- (strategic planning) Does the program use a prioritization process to guide budget requests and funding decisions?
- Program management) For R&D programs other than competitive grants programs, does the program allocate funds and use management processes that maintain program quality?
- (program results and accountability) Were program goals achieved within budgeted costs and established schedules?

For each question on the program, the evaluator answers yes or no[19] to each question and the total score is calculated by weighted[20] average of scores achieved in each question. However, despite the guidelines from OMB, there still remains the possibility of inconsistent and subjective evaluation results. In order to cope with this problem, OMB has made efforts to set clear standards, to improve questionnaires, and to put more resources on educating evaluators. Evaluation results are presented in four different levels of performances; effective, moderately effective, adequate, and ineffective. If measurable performance indicators are not available or a program agency does not provide information on program performance, then the evaluator labels the case as "result not demonstrated". OMB collects and summarizes results from PART assessments to publish "Performance and Management Assessments" as a part of the presidential budget bill and announces it on its website.

Since the introduction of PART, performance and transparency of programs have steadily improved. The number of programs receiving a grade no worse than "adequate" has considerably increased.[21] The main driving forces behind the improvement in program performances are that the evaluation was conducted in a transparent and consistent manner and program agencies made great efforts to get high scores.

d. Performance Improvement Officers (PIO)

On September 13, 2007, President Bush signed the Executive Order 13450 to improve the government program performance. The Order clearly announced that more efficient use of taxpayers' money is an official policy of the federal government. Following the Order, Performance Improvement Officer (PIO) was appointed in each program agency and Performance Improvement Council (CIO) was established as a consultative group mainly

19. In some cases, more options for the answer are allowed like yes, somewhat yes, somewhat no, or no.
20. Currently, the following weights are assigned; 20% to program purpose and design, 10% strategic planning, 20% program management, 50% to program results and accountability.
21. By 2008, OMB and federal organizations have completed evaluations on 1,015 programs, equivalent to 98% of programs run by the federal budget. Seventy-five percent of them achieved performance goals in 2007, and 63% in 2008. Moreover, in 2008, 57% of programs had improved their performance, which is an increase of 12% points from 2007.

consisting of PIOs from program agencies. PIO is appointed by the head of program agency and should be given authority to take actions and measures to accomplish the mission of performance improvement.

The following duties are assigned to PIOs;

- Advise and assist the head of the agency and the Chief Operating Officer to ensure that the mission and goals of the agency are achieved through strategic and performance planning, measurement, analysis, regular assessment of progress, and use of performance information to improve the results achieved;
- Advise the head of the agency and the Chief Operating Officer on the selection of agency goals, including opportunities to collaborate with other agencies on common goals;
- Assist the head of the agency and the Chief Operating Officer in overseeing the implementation of the agency strategic planning, performance planning, and reporting requirements, including the contributions of the agency to the federal government priority goals;
- Support the head of agency and the Chief Operating Officer in the conduct of regular reviews of agency performance, including at least quarterly reviews of progress achieved toward agency priority goals, if applicable;
- Assist the head of the agency and the Chief Operating Officer in the development and use within the agency of performance measures in personnel performance appraisals, and, as appropriate, other agency personnel and planning processes and assessments;
- Ensure that agency progress toward the achievement of all goals is communicated to leaders, managers, and employees in the agency and Congress, and made available on a public website of the agency.

PIC, consisting of high ranking officer of OMB and PIOs from program agencies, is responsible for establishing standards of program performance and evaluation criteria, facilitating inter-agency information exchange, coordinating performance evaluation, and deciding the policy on information disclosure and information gathering from stakeholders. As of September, 2010, 49 of the federal agencies appointed their own PIOs, aiding to improve performance management system in federal agencies.

e. Crosscutting

Crosscutting is another feature of the performance management system in the United States. Its major use is to improve performance of programs that share program objectives

or belong to similar types. Crosscutting facilitates coordination and communication among managers of similar programs and enables them to reach an agreement on common objectives and to handle common difficulties cooperatively. Both the OMB and program agency participate in the process to decide on whether programs with similar characteristics should be combined into one program for the purpose of PART evaluation or treated as independent programs with similar objectives to be inspected through crosscutting.

There are two kinds of crosscutting; internal crosscutting and external crosscutting. The former is conducted on multiple programs from a single agency and the latter from different agencies. The fundamental goal of crosscutting is to discover best practices to be used for other programs with similar characteristics, to establish common performance indicators, resolving problems in the decision making process, and to coordinate actions and measures among different program agencies.

Unlike PART that treats programs with similar objectives or characteristics as a single subject of evaluation, crosscutting reviews programs on an individual basis. Crosscutting identifies the features of programs with similar objectives or characteristics through evaluations on individual programs and clarifies similarities of those programs by asking questions like;

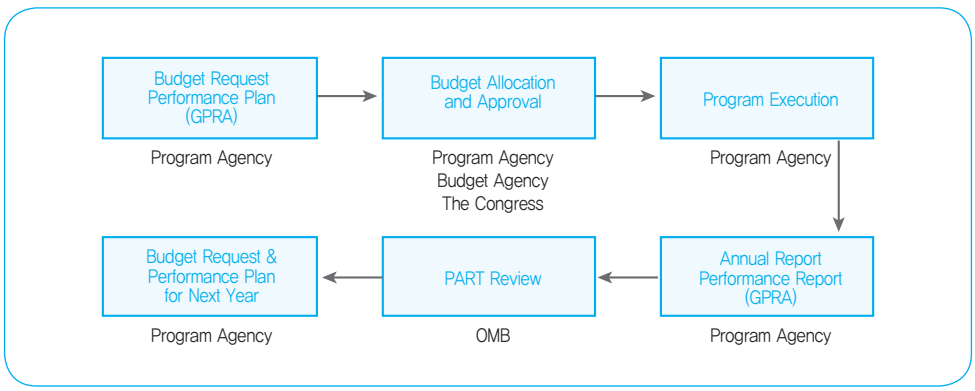
- What are the targets for each program?
- What kind of outputs and services are provided? And, is it appropriate to build common performance indicators?
- How efficiently are outputs and services provided? And, is it appropriate to set up common output indicators?
- What kind of outcomes did the programs achieve? And, is it appropriate to accept common outcome indicators?

Internal crosscutting does not require additional data since the above questions are already asked in PARTs of individual programs, and a summary and improvement plan from PARTs of individual programs can be used as useful references in crosscutting. However, external crosscutting requires identification of common strengths and improvement factors from PARTs of individual programs. Fundamental purpose of crosscutting analysis is to examine the existence of common system to collect performance information and to secure responsibility of program agencies and provide the foundation for the common system if it does not exist.

1.3. Operation of New Performance Management System

Performance management system of Bush administration sought to link performance information and evaluation results with budget allocation. The generic procedure is illustrated in [Figure 3-1].

Figure 3-1 | Performance Management System under Bush Administration



The cycle of performance management system starts with each program agency submitting a budget request and annual performance plan based on GPRA. Budget requests from all agencies are compiled, adjusted and prioritized to become the President' Budget. The Budget bill is sent to the Congress for debate and voting. Once the Bill passes the Congress, program agencies execute the program in compliance with the approved budget and performance plan. At the end of the fiscal year, program agencies submit to OMB annual reports on the financial status of the program and annual performance report as required by GPRA. Then, OMB conducts performance evaluations according to PART and the evaluation results are announced to the public as well as stakeholders. The evaluation results are incorporated into next year's budget process.

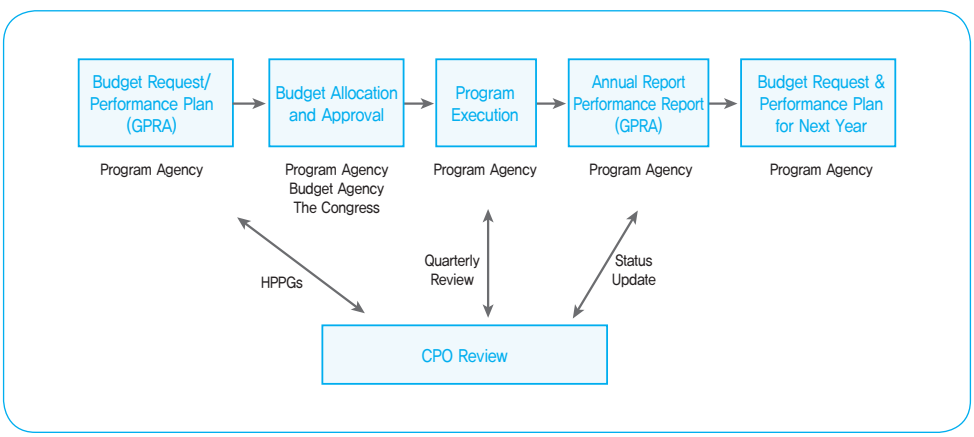
Performance management system becomes an important institution in that it is an attempt to reinforce fiscal efficiency and accountability by incorporating results of performance evaluation into the budget process. The system, however, has limitations. It is very difficult to establish a direct link between performance and budget due to various external factors like the political climate. Moreover, the link is indirect in the sense that evaluation results are reflected not on a current year's budget but on the next year's budget. To address these problems, the Obama Administration suggested a new form of the performance management system that can consistently monitor the program performances and promptly incorporate performance information into the budget process.

The key elements of the Obama administration's new performance management system consist of four parts. The first element is the introduction of the system in which core performance goals are established and managed with a view to attaining a high level of program performance. [Figure 3-2] illustrates the flow chart of the Obama administration's new performance management system.

Each program agency drafts next year's budget request taking into consideration strategic goals, long-term and medium-term strategic plan, and three to eight high priority performance goals (HPPGs) attainable within the next 24 months and sends it to the Chief Performance Officer (CPI) of OMB who reviews the reports from each program agency and attaches them to the final draft of the budget bill. CPI calls for quarterly meetings of under-secretaries of departments in administrative branches on reported HPPGs and crosscutting management goals. Liaison officers from each department should post recent performance information on the Federal Performance Portal to maintain the transparency of the status of the program performance. Performance reports will be published at the end of the fiscal year and they should be disclosed to the public as well as the Congress and all stakeholders through the Federal Performance Portal.

The Federal Performance Portal[22] was introduced in 2011 as the central information hub on the federal government's efforts to improve performance and accountability of the public agencies. The Portal provides updated information on performance information classified by theme, agency, and program type as well as individual programs.

Figure 3-2 | Performance Management System under Obama Administration



22. http://www.performance.gov/

The second feature of Obama administration's performance management system is the emphasis on in-depth program evaluations on on-going programs or inter-agency programs. The information collected from the program evaluation is utilized in restructuring the structure of government expenditure. The results of the program evaluations are made available through both on-line and off-line accesses, calling attention to program evaluations among experts and practitioners. The third characteristic is that each department is required to conduct cost-benefit analysis to evaluate efficiency and effectiveness of programs and utilize the information from the analysis in budget allocation among program agencies. The fourth feature is to enhance competency of federal workers in performance management to achieve better performance.

2. The United Kingdom

2.1. Brief History of Performance Management in U. K.

The concept of performance management was first introduced in the mid-1960s, affected by the planning-programming budget system (PPBS). Yet, performance evaluation had not been actively pursued until the 1980s. However, between the end of the 1960s and early 1970s, consensus was being reached on the necessity for systematic tools to control government expenditures and prioritize diverse government policies. Based on such consensus, the Central Policy Review Staff (CPRS) was created within the Cabinet Office in 1971 with a view to strengthening evaluation capacity of government departments and public agencies.

The Heath Cabinet, formed in 1970, created the CPRS and conducted Program Analysis and Review for large and important administration agencies. However, when Margaret Thatcher took over the Cabinet, Program Analysis and Review was abolished and CPRS was disbanded for the lack of professionalism and expertise. Instead, the Thatcher Cabinet introduced Efficiency Strategy and Financial Management Initiatives (FMI). Efficiency Strategy was introduced to enhance the efficiency of the public sector and a thorough examination on the process of policy implementation in public agencies was conducted. Those measures enabled the government to broaden the realm of performance evaluation, which led to the establishment of FMI in the 1980s. Under FMI, strategic goals of each public agency were clearly defined and discretionary power of the budget was delegated to the program agencies to maximize "value for money (VFM)" of government expenditure programs.

Led by Efficiency Strategy and FMI, the British government made significant efforts to enhance performance of the public sector from the 1980s with special emphasis on results-

oriented management. Consequently, a paradigm shift had occurred in the performance management system from control and management of inputs to emphasis on results or outcomes. As a consequence of these efforts, the National Audit Act was enacted in 1983 to solidify the foundation for realization of value for money in government expenditure programs. Later on, in 1988, the British government launched Next Step Initiatives (NSI) with the purpose of improving efficiency of the public sector combining the public nature of the government activities and market discipline. One important feature of NSI was to confer both discretion and responsibility on the manager of a program agency. In addition, with the establishment of the Citizens' Charter in 1991 and its revision in 1998, all public agencies were encouraged to publish administrative information and set the standards for services, which ultimately resulted in better accessibility to public sector services. Since the general election in 1997, various measures to improve the performance management system was taken in the context of fiscal reform. *Ex-ante* review on the government budget was firmly institutionalized and Spending Review (SR) by HM Treasury as an ex-post examination on government expenditure was introduced in 1988. SR was conducted every two years and provided performance goals and performance indicators for the next three years for program agencies under review. In 1999, Public Service Agreement (PSA) was adopted to lay down a foundation for the principle of result-oriented performance management.

From 2000, PSA, a revised version of old SR, started to be implemented in all departments including HM Treasury, along with Service Delivery Agreements (SDA) and Technical Notes (TN). Those efforts helped modernize performance management and facilitate reform of the public sector. The British government made continuous efforts to accomplish goals of PSA. Not only did each program agency establish policy objectives and performance goals, but HM Treasury and the Cabinet Office oversaw the execution of the performance management system by conducting regular SR and performance evaluations. In 2000, the Freedom of Information Act was legislated and additional devices such as the Local Public Service Agreement (LPSA), Service Delivery Agreement (SDA), and Best Value were explicitly introduced into the performance management system. Moreover, the Office for Public Services Reform and Delivery Unit were established to promote reform measures including setting up national standards for public services and accountability of the public agencies for their performance. Comprehensive Performance Assessment (CPA) was implemented in 2002 to promote cost efficiency of public expenditure programs. Also, Capability Review was introduced as a measure to reinforce the performance management system.

2.2. Performance Management System in U. K.

2.2.1. Performance Based Budget System in U. K.

Public expenditures in the UK are based on the following principles;

- Long-term consistency in fiscal management, and transparent system
- Policy decision making based on results rather than inputs
- Mid-term and long-term plans to provide public services more efficiently and promotion of inter-agency cooperation
- Appropriate cost accounting and management of capital assets for public investment

The UK government announced two fiscal rules all public agencies should abide by in formulating expenditure programs, namely, the Golden Rule and Sustainable Investment Rule. The former requires that the budget balance or a reasonable amount of budget surplus should be maintained irrespective of fluctuations in revenue and expenditure due to a business cycle, which allows the government to borrow only for a mid-term and long-term investment. The latter stipulates that the public sector debt should be maintained at a stable and manageable level, which is currently targeted at 40% of GDP.

Departmental expenditure limit (DEL) for each government department is another important element in the performance based budget system in UK. Each government department sets the expenditure limits on the 3-year expenditure plan over the SR term, which should be compliant with two fiscal rules of the golden rule and the sustainable investment rule. DEL consists of two parts; resource budget and capital budget. Once DEL is fixed, it is allowed to carry forward unused budget to the next year and such end-year flexibility has the advantage of preventing the public agencies from spending unused resources without due regard to efficiency at the end of the fiscal year. Since its introduction in 1997, SR has been conducted 6 times[23] so far. The 3-year DEL, included in the conclusion of SR, should be considered as a political promise without parliamentary approval or legal ground. Nonetheless, each department is expected to abide by the 3-year DEL and required to obtain approval from HM Treasury before submitting an annual budget request.

In general the DEL will cover all administration costs and most program expenditures. In a few exceptional cases, spending is not recorded in DEL but is instead in departmental Annually Managed Expenditure (AME) because it cannot reasonably be subject to close control over the period longer than a year. In general, AME deals with large-scale, fluctuant, and demand-driven expenditures such as social security spending. Screening on AME is conducted twice a year as a part of the budget process and pre-budget report process.

23. Those are the years of 1998, 2000, 2002, 2004, 2007, and 2010.

Total Managed Expenditure (TME) is currently the preferred measure of total public expenditure. TME includes the current and capital expenditure of the public sector but not financial transactions such as government lending or buying of shares. Public sector capital expenditure includes fixed capital formation (expenditure on capital assets, both tangible such as buildings, machines, and vehicles, and intangible such as computer software) and capital grants to the private sector. TME covers expenditure by the whole public sector; central government, local government, and public corporations. TME can also be presented in terms of the way the government sets budgets for spending by the public sector. TME is the sum of Departmental Expenditure Limits (DEL) and Annually Managed Expenditure (AME).

Since reforms of the government budget and accounting standards, the terms, resource budgeting and resource accounting started to be widely used. They actually mean that the accrual principle is fully incorporated into budgeting and accounting practices. Resource Accounting and Budgeting (RAB) was launched in 1993, with a commitment to introduce resource accounting. This was followed by a White Paper in 1995, which gave a commitment to use resource accounting as the basis of public expenditure planning and control. RAB was implemented in full scale from 2001. Resource budgeting is the application of accruals accounting for reporting on the expenditure of central government and a framework for analyzing expenditure by departmental aim and objectives, relating these to outputs where possible. It can provide practical information for Parliament and HM Treasury by eliminating limitations on information in the accounting practice based on cash based accounting. However, it needs to be noted that in spite of all that, the British government has not completely abandoned cash basis accounting. In other words, budgetary documents submitted to Parliament for approval are prepared based on accrual as well as cash principles. Requests for resources that include depreciation and capital charge are prepared based on accrual principle but cash requirement and financing requirement derived from requests for resources are prepared following the principle of cash basis accounting.

The Spending Review (SR) is a Treasury-led process to allocate resources across all government departments, according to the government's priorities. SR sets firm and fixed spending budgets over several years for each department. It is then up to departments to decide how best to manage and distribute this spending within their areas of responsibility. In addition to setting departmental budgets, SR also examines non-departmental spending that cannot be firmly fixed over a period of several years, including social security, tax credits, some elements of local authority spending and spending financed from the proceeds of the National Lottery. SRs have been an important part of governmental planning since the late 1990s. Prior to the introduction of SR, departmental budgets were set on a year-by-year basis which made multi-year planning more difficult.

Based on PSA, SR sets departmental expenditure limits valid for the next 3 years and defines core improvements expected. It includes information on the factors of financial pressure, the possibility of efficiency improvement, and the expected cost of new expenditure programs. While HM Treasury is responsible for the entire procedure, it is each department that actually collects information, conducts analysis, and prepares the reports. Although SR is based on a bottom-up procedure, the budget request should be made in consideration of the total managed expenditure (TME) of the department.

The core of the performance management system in the UK is PSA. PSA aims to improve performance of the public sector by providing performance indexes related to the results or outcomes of government expenditure programs and is conducted every three years.[24]

2.2.2. Public Service Agreement (PSA)

PSA, first introduced in CSR in 1998, is the three-year agreement on departmental activities and presents objectives of public expenditure programs. That is, it contains strategic goals of each government department as well as performance goals of the programs managed by the department and is established based on agreement between HM Treasury and each department during the SR process.

PSA consists of the following elements;

- The description of the government's aims in major area and departments and senior officials in charge of them.
- Lower level objectives contributing to achieving higher level aims.
- Performance indicators and performance targets that represent the level of achievement of objectives.
- Description on how measure and manage performance indicators and targets.

For the operation of PSA including delivery strategy to accomplish aims, objectives, and performance targets or indicators, the following process is;

- HM Treasury examines the initial expenditure plan submitted by each government department.
- The Delivery Unit in the Prime Minister's office monitors delivery strategy of each government department.
- Each government department should submit reports on objectives in PAS at least twice a year and post the relevant information on the web-site on a regular basis.

24. A detailed discussion on PSA is provided in the next section.

- The National Audit Office monitors reliability and adequacy of performance indicators and targets.

PSA was revised in 2007 to solidify its success by incorporating experiences and important features of the new system which include;

- Each PSA is reinforced with Delivery Agreement (DA) which describes the plan and the role of major cooperators.
- Result or outcome-oriented performance indicators at national level were suggested in PSA and specific contents of each PSA are described in DA.
- New PSA demonstrates result-oriented performance index on a national level, DA specifies its content.
- Departmental Strategic Objective (DSO) was introduced to simplifythe complex system of objectives and targets and establish more efficient system at departmental level.

The most revealing change in the British performance management system is a transformation of the PSA system toward a simpler and result-oriented direction. To be more specific, the change in the PSA system involves the number, characteristics, and the scope of performance targets. The number of performance targets under PSA significantly decreased from 600 in 1998, to 130 in 2002, and to 30 in 2007. Drastic decrease in the number of performance targets was related to the transformation of the performance management system from input or process oriented to result or outcome oriented ones. For instance, while outcome performance indicators accounted for only 15 percent of 600 performance indicators in 1998, 68% of all performance indicators were related to results or outcomes in 2000 and all of the performance indictors in PSA were outcome indicators. The drastic decrease in the number of PSA performance indicators in 2007 is also related to the changes in the basic unit of the PSA itself from an individual department to the program that may require the involvement of several departments. Like performance targets, the

number of PSAs decreased drastically from 160 in 2000, to 130 in 2002, to 110 in 2004, and to 30 in 2007[25].

2.3. Institutional Arrangement

The UK government system does not have an independent organization in charge of policy evaluation. Instead, several organizations such as the Strategy Unit and the Delivery Unit in the Cabinet Office, Division of Public Service in HM Treasury are conducting evaluation related tasks.

Evaluations are conducted at the departmental level and each department has different organizational structures for various works for evaluations. Some have a permanent organization and others have a temporary one as necessary. The Strategy Unit and the Delivery Unit in the Cabinet Office also take part in the performance evaluation process and the supreme organization related to the budget is the Division of Public Service in the HM Treasury.

25. The current 30 PSAs are; 1) Raise the productivity of the UK economy. 2) Improve the skills of the population, on the way to ensuring a better skills base by 2020. 3) Ensure controlled, fair migration that protects the public and contributes to economic growth. 4) Promote science and innovation in the UK. 5) Deliver reliable and efficient transport networks that support economic growth. 6) Deliver the conditions for business success in the UK. 7) Improve the economic performance of all English regions and reduce the gap in economic growth rates between regions. 8) Maximize employment opportunity for all. 9) Halve the number of children in poverty by 2010-11, on the way to eradicating child poverty by 2020. 10) Raise the educational achievement of all children and young people. 11) Narrow the gap in educational achievement between children from low income and disadvantaged backgrounds and their peers. 12) Improve the health and wellbeing of children and young people. 13) Improve children and young people's safety. 14) Increase the number of children and young people on the path to success. 15) Address the disadvantage that individuals experience because of their gender, race, disability, age, sexual orientation, religion or belief. 16) Increase the proportion of socially excluded adults in settled accommodation and employment, education or training. 17) Tackle poverty and promote greater independence and wellbeing in later life. 18) Promote better health and wellbeing for all. 19) Ensure better care for all. 20) Increase long term housing supply and affordability. 21) Build more cohesive, empowered and active communities. 22) Deliver a successful Olympic Games and Paralympic Games with a sustainable legacy and get more children and young people taking part in high quality PE and sport. 23) Make communities safer. 24) Deliver a more effective, transparent and responsive Criminal Justice System for victims and the public. 25) Reduce the harm caused by Alcohol and Drugs. 26) Reduce the risk to the UK and its interests overseas from international terrorism. 27) Lead the global effort to avoid dangerous climate change. 28) Secure a healthy natural environment for today and the future. 29) Reduce poverty in poorer countries through quicker progress towards the Millennium Development Goals. 30) Reduce the impact of conflict through enhanced UK and international efforts.

2.3.1. The Cabinet Office

The Cabinet Office is one of two secretarial organizations to the Prime Minister in the UK. The other organization is the Prime Minister's Office assisting the Prime Minister by providing advices and services, facilitating communications with other organizations inside and outside the government, and carrying out administrative works on behalf of the Prime Minister.

The Cabinet Office performs several important tasks. The Office supports the Prime Minister by coordinating policies among different government departments, analyzing policy issues involved in several departments, and monitoring the implementation of the Cabinet's decision. In addition, it is in charge of controlling and reforming the public sector organizations as an aid to the Prime Minister. Moreover, the director and the staff members of the Office do not promptly resign even when a new Prime Minister is elected. They provide services to the new Prime Minister at least for a reasonable period to ensure the continuity of the government.

2.3.2. The Strategy Unit

The Strategy Unit was formed in 2006 by merging the Performance and Innovation Unit, Forward Strategy Unit in the Prime Minister's Office, and a part of the Centre for Management and Policy Studies. The major roles of Strategy Unit include;

- Conducting strategic policy evaluation and making policy recommendations to the Prime Minister.
- Supporting government departments to develop effective and efficient policies and strategies along with strategic capability.
- Carrying out strategic audits if necessary.
- Identifying core tasks the government should pursue.

The Unit reports to the Prime Minister and the Cabinet through the Cabinet Officer, and most of members are public officers, not civilian special advisers. The Executive Office of the President (EOP) in the United States has similar roles and functions as the Strategy Unit.

2.3.3. The Delivery Unit

The main function of the Delivery Unit is to help the government department provide public services in a more effective and efficient manner. The Unit examines the core policies and reports the results to the Prime Minister. It also helps enhance governmental capability to provide public services by identifying main obstacles and actions required. Moreover, it aids the development of PSA targets that can improve the public service.

The Unit maintains a cooperative relationship with other government departments and public agencies to evaluate the performance of the providers of the public services and strengthen the performance management system. In addition, the Unit shares responsibility on PSA targets with HM Treasury.

Efficient division of labor between the Unit and HM Treasury is a crucial factor for successful operation of the performance management system in the UK. The Unit selects core policy objectives and performance targets on which the government will concentrate and HM Treasury deals with the rest. The responsibility of the Delivery Units limited to some core sector and departments that are high in the Prime Minister's priority. Therefore, there is little tension between the Unit and HM Treasury because their roles and responsibilities are clearly delineated.

2.3.4. HM Treasury

HM Treasury reviews and assesses effectiveness of government programs and activities. The Treasury obtains information necessary for future policy development and budget allocation thorough these activities. The target of evaluation is PSAs, the documents signed by the Treasury and government departments. PSA is the budget plan that includes the mission, objectives, and performance goals of a government department. The Delivery Unit in the Cabinet Office is in charge of education, consultation, and research on evaluation related works and conducts policy analysis and performance evaluation on selected programs.

The Treasury produces and distributes the guidelines for evaluation of public services. Following the guidelines, each government department constructs performance indicators and conducts evaluations on its own programs. The results of performance evaluations are reported to the Treasury. Performance evaluations by the Treasury are carried out through Comprehensive Spending Review (CSR) and SR. The difference between the two schemes is that CST analyzes and evaluates the current government activities from zero-base while SR tries to evaluate the marginal changes in the performance of government expenditure programs based on the current status. More specifically, SR evaluates program performance in the past two years based on PSA and SDA and modifies and further develops program objectives for the next three years based on the evaluation results. SR is divided into two parts, departmental review and cross-cutting review. Departmental review is the evaluations on the activities of individual departments and cross-cutting review is a system under which evaluation on inter-departmental programs or policies are carried out. The cross-cutting review is a characteristic feature in the British performance evaluation system and was first conducted in 2000 as a part of SR to overcome limitations of the departmental review. From 2000, the Treasury collected reports of performance evaluation from government

departments and prepared annual reports on expenditure evaluations and public service. The report is submitted to the Prime Minister and Parliament.

2.3.5. Evaluation Units at Department Level

Each government department has different organizational structures for performance evaluation. While some departments have permanent organizations or units for performance evaluation of their own activities and programs, others form an evaluation unit on a temporary basis when it is necessary. In practice, most of the government departments do not have an independent organization for performance evaluation but the treasury division of each department assumes the responsibility of evaluation tasks. During the budget process, each government department carries out evaluations on PSA and its expenditure programs and the treasury division is the main channel through which Division of Public Service in HM Treasury communicates with the department during the evaluation process. Without a permanent organization for evaluation, most government departments are active in utilizing internal and external experts to carry out in-depth evaluations on program performance.

2.3.6. National Audit Office

The National Audit Office (NAO) in the UK scrutinizes public spending of the central government on behalf of the Parliament. NAO's audit of central government has two main aims. By reporting the results of audits to Parliament, NAO holds government departments and bodies to account for the way they use public money, thereby safeguarding the interests of taxpayers. In addition, NAO aims to help public service managers improve performance and service delivery. The Audit and inspection rights are vested in the head of the National Audit Office, the Comptroller and Auditor General (C&AG). The staffs of the NAO carry out these tasks on his behalf. The Comptroller and Auditor General is an Officer of the House of Commons. Both he and his staff at the NAO are totally independent of government. They are not civil servants and do not report to any Minister of the government department. Oversight of the NAO is carried out by a Parliamentary committee, the Public Accounts Commission, which appoints external auditors and scrutinizes the NAO's performance.

Established based on the National Audit Act 1983, the NAO has been conducting audits of fiscal values in addition to the traditional audit activities like year-end audits of the government budget and financial audits of the government, public enterprises, and social security organizations.

Regarding performance audits, the NAO participated in the development of forms of performance reports and standards of writing performance reports along with the Treasury and other government departments. It also carries out verification tests on performance data system of the government departments and publishes the results of the test.

The verification test on the performance data management system focuses on adequacy of control and degree of danger in handling the data. The test is carried out in five steps;

- Understanding of performance management system in PSA of the department.
- Identification of risk factors in performance data.
- Assessment on the importance of identified risk factors.
- Assessment on the adequacy of internal control on important risk factors.
- Examination of evaluation results and preparation of the report.

3. Australia

3.1. Brief History of Performance Management in Australia

Since the 1970s, low saving rate, high tariff barriers, and strong regulation on product market had caused serious economic malaise such as slow growth, high unemployment, and high inflation. In particular, the average growth rate of per capita income during the 1970-1980s was merely 1.8% while the unemployment rate soared to 8% to 10% during the 1980-1990s. Reform of the public sector in New Zealand in 1984 stimulated the Australian government to launch its own reform program.

Since 1983, the Australian government started to push forward the reform program of the government budget and fiscal management to reduce the budget deficit by overall spending cut and enhance the efficiency of the budget management process. The size of total budget as well as government expenditure was tightly controlled. On the other hand, a new paradigm of budget management was introduced. Flexibility and incentives were granted to enhance the efficiency of budget management and more authority to make financial decisions was delegated to the budget managers at the departmental level. The budget managers were made accountable for the program performance and result-oriented performance management system was firmly established. The Program Management and Budgeting (PMB) was launched in 1987, which turned the focus of performance management from inputs to results and outcomes. The main features of PMB include establishing performance objectives at the departmental level, developing tools to measure program performance, and publishing annual budget reports containing performance objectives along with performance indicators. Each government department was asked to make a Portfolio Evaluation Plan[26] and submit it every November to the Department of Finance (DOF) that was in charge

26. Portfolio indicates a group of the public agencies under the authority of a minister in the Cabinet and the Department oversees coordination among the agencies in a portfolio.

of coordination and support for performance evaluation activities. In addition, the major evaluation on each program was carried out every three to five year.

The new conservative government was formed in 1996 to turn the system to a more market oriented direction and to delegate more authority to the Department Head. Subsequently, the budget and accounting system based on accrual principle was introduced to promote a competitive atmosphere in the federal government in 1999. Various duties of government departments related to performance evaluation under the old regime were abolished and a new performance evaluation system was introduced. Department of Finance and Administration (DOFA) published a manual called "Good Practice Principles for Performance Management" to guide each department to carry out performance evaluations on its own programs. Each government department manages its own performance information, performance measurement and evaluation, and performance reporting. Performance information is contained in two important documents, portfolio budget statement (PBS) and annual report. PBS is an accountability document that contains information on internal assessment and discussion on program performance. PBS is written by the government department and submitted to the Parliament. Annual report is published by the Department of Prime Minister and Cabinet and provides information on the budget of the cabinet departments. As a consequence of reform efforts by the Australian government, the role of DOFA was reduced while the autonomy of each department was strengthened and the roles of the Secretarial Office and private experts became more important as an advisor on the public service.

The new performance evaluation system supplemented the Financial Management and Accountability Act 1997, the Commonwealth Authorities and Companies Act 1997, Auditor General Act 1997, and the Public Service Act 2000. Public Service Act 2000 is a significant departure from the past laws in terms of its contents and style. The law brought considerable changes into the public sector by extending the role and power of the organizational heads and assigning independent screening power in appointing staff members to each department. In addition, through the Joint Committee of Public Accounts and Audit, Senate Committees, Senate Finance and Public Administration Committees, monitoring the government departments by the Parliament was strengthened to increase accountability of the Cabinet members.

Yet, the verdict on the reforms on the performance management system up to now is not clear. As the criticism has been raised from the lower echelons of the government employees that performance information submitted by each department is highly heterogeneous and hence not so useful in decision making, a new trend that puts more emphasis on stronger intervention on the performance management system by the budget authority has surfaced. In sum, the reforms in the performance management system emphasizing results and

outcomes based on the accrual principle have affected both the Australian government and the Parliament and made all stakeholders note the importance of monitoring and continuous improvement of the performance management system.

3.2. Performance Management System in Australia

The current performance management system in Australia started with fiscal reform initiated by the National Commission of Auditing 1996. One of the important characteristics in the Australian system is that the performance management system has strong legal foundations. Four of them are particularly important; Financial Management and Accountability Act 1997, Commonwealth Authorities and Companies Act 1997, Auditor General Act 1997, and Public Service Act 2000.

Financial Management and Accountability Act 1997 was enacted primarily to delegate authorities on fiscal management to lower ranked officials in program agencies. More precisely, the Act deals with the roles and responsibilities of the Secretary to DOFA, management and control of assets and financial resources in the public sector, authority on financial transactions of the ministers of DOFA and the Department of the Treasury, and financial audit. As for the performance management system, it provides standards for efficient management of public assets and delegates the fiscal authority as well as accountability to the Secretary of each department. Each department is responsible for submitting an annual financial report to the Auditor General.

Commonwealth Authorities and Companies Act 1997 stipulates the qualifications required for the organization dealing with non-public funds. The Act regulates reporting rules and responsibilities of the Commonwealth Authorities and private companies. The Commonwealth Authorities should prepare the annual reports and submit it to the Secretary of the relevant government department by the fifteenth day of the fourth month after the fiscal year. The Secretary should immediately be transferred to the Parliament except for the case when she allowed late submissions of the report due to special reasons.

To enhance accountability of the Administration, the Australian Parliament passed the Auditor General Act 1997, which clearly describes the roles of the Auditor General as the supreme monitor and reporter on the performance and accountability in using public resources by the Commonwealth government. The Act specifies that the Auditor General is responsible for financial audit, performance audit, crisis management, and internal audit on the Commonwealth Authorities. The Auditor General possesses the authority to conduct performance audits on public agency, the Commonwealth Authority, and its subsidiaries. The audit report should be submitted to the Parliament and Secretaries of the relevant departments.

The purposes of the Public Service Act 2000 are to establish an efficient and effective platform to provide the public services and to offer a legal framework for fair and efficient employment of workers of the public service sector. In addition, the Act specifies the roles and responsibilities of the Head of the public service agency, the Public Service Commissioner and the Merit Protection Commissioner along with rights, duties, compensation, and code of conduct for workers of the public service sector. The Act also includes clauses on conversion to accrual basis budgeting and accrual appropriation, performance management system based on results and outputs, and extended responsibility and flexibility in budgeting process of the Head of the government department.

Australian fiscal reform in the 1990s intended to strengthen the accountability and responsibility of the public agencies by holding the heads of them accountable for the performance of the organization or programs it manages. The reforms promoted the development of the political environment under which a public agency can establish the strategic priority among competing programs or policies. A significant degree of flexibility and discretion was allowed to the public agency and its member to achieve policy goals effectively. The head of the public agency are held accountable for outputs and results of the programs. The central government's grip on comprehensive evaluation plans was loosened and the demand for accurate and measurable performance information increased. Every public agency was required to submit the evaluation plan on the programs for the next five years. Public agencies started to use new techniques of performance evaluation and monitoring and The Australian National Audit Office (ANAO) and DOFA jointly selected and publicized the best practices to promote the new result-oriented performance management system.

The experience in the 1990s led the Australian government to introduce Budget Estimates and Framework Review which pursued to achieve more focus on program information, more timely and detailed reports on fiscal information, stronger monitoring on financial performance and cash flow of the program agency, better registration and database systems to cope with expanding information stocks and better analytical skills for the employees of the program agency.

We can list the important elements in the Australian performance management system as follows;

- Efficient program evaluation thorough accrual basis accounting; all government departments have prepared all financial reports following accrual basis accounting since 1999. The practice enabled the program agency to estimate the total costs, including depreciation and indirect costs of the public service, to evaluate financial stability, and to achieve enhanced accountability of the performance management system.

- Mid-term fiscal plan; a fiscal plan for at least next three years should be submitted to the Department of the Treasury to accommodate the budgetary requirements of the programs lasting longer than a year. The conflict between the budget and program agencies can be minimized and predictability and stability of the budget process can be achieved through the mid-term plan.
- Stronger monitoring; to strengthen the monitoring and evaluation on financial and performance reports in each public agency, Australian National Audit Office was established as an independent unit in 1997 in addition to the parliamentary supervision over the administration.
- Outcomes and outputs framework; two issues have been repeatedly raised in establishing performance management system in Australia. They are the quality of performance information to identify contribution of the program agency in achieving outputs and outcomes and limited utilization of performance information. These problems led the Australian government to reach the conclusion that it is very important to secure the measurability of the linkage among inputs, activities, outputs, and outcomes. Australia used to have output oriented performance management system along with New Zealand but recently turned the attention to outcomes. Significant efforts are made to draw clear distinction between outputs and outcomes and to measure quantity, quality, and costs of outputs accurately. In addition, performance indicators are consistently updated for better measurement of program performance.

3.3. Institutional Arrangement

Three important players, the Department of Finance and Administration (DOFA), the ANAO, and the Australian Parliament, are worth noting in the Australian system of performance management.

DOFA provides guidelines for performance management and makes recommendations on the priorities for government expenditure programs. It also prepares reports on the performance of program agencies as well as government programs, which would be submitted to the Cabinet and its members. DOFA is responsible for various institutions related to the performance management system in Australia. Also, it acts as a consultant on government expenditure programs. In this context, DOFA offers advice to the Prime Minister and the Cabinet on the performance information when new programs are suggested during the budget process. Moreover, the Department is in charge of reviewing the annual budget plan as well as performance monitoring and evaluation on program agencies and government departments. DOFA introduced the Strategic Review Unit in 2006 to provide

information to important decision makers such as heads of government departments during the budget process. Strategic Review Unit examines large and complex expenditure programs with high priority and produces useful information for strategic management of fiscal resources.

The Auditor-General is responsible, under the Auditor-General Act 1997, for providing auditing services to the Parliament and public sector entities. The Australian National Audit Office (ANAO) supports the Auditor-General, who is an independent officer of the Parliament. The ANAO's primary client is the Australian Parliament and its purpose is to provide the Parliament with an independent assessment of selected areas of public administration, and assurance about public sector financial reporting, administration, and accountability. The ANAO does this primarily by conducting performance audits, financial statement audits, and assurance reviews but does not exercise management functions or have an executive role. The ANAO performs the financial statement audits of all Australian Government controlled entities and seeks to provide an objective assessment of areas where improvements can be made in public administration and service delivery. The ANAO has extensive powers of access to Commonwealth documents and information, and its work is governed by its auditing standards, which adopt the standards applied by the auditing professions in Australia. In accordance with these standards, performance audit, financial statement audit and assurance review reports by ANAO are designed to provide a reasonable level of assurance. The ANAO adopts a consultative approach to its forward audit program, which takes account of the priorities of the Parliament, as advised by the Joint Committee of Public Accounts and Audit, the views of entities and other stakeholders. The program aims to provide a broad coverage of areas of public administration and is underpinned by a risk-based methodology. The final audit program is determined by the Auditor-General.

The Budget Bill is prepared based on outputs and outcomes of each government department and presented to the Parliament for approval. Indirect costs are inputted to outputs or outcomes and the Parliament votes only on budgets for outcomes. Each department manages the budget allocated to them within the boundary of the authorization by the Parliament. Information on output and programs are provided to the Parliament through the Portfolio Budget Statements, which is a core component of the budget documents.

4. Japan

4.1. Brief History of Performance Management in Japan

Japanese economy had been stagnant ever since the famous collapse of asset price bubbles in the 1990's and the Japanese government responded to the depression by increasing government expenditure to stimulate the economy. An inevitable consequence of prolonged depression and ever increasing expenditure was the accumulation of a budget deficit and government debt at an alarming speed. The government debt reached well over 108% of GDP in 1997 and the sheer size of the government debt was enough to trigger a cry for fundamental fiscal reform. In addition, a rapidly aging population combined with low fertility also casted negative implications on long term economic growth and fiscal stability. Confronted with grim prospects on future economic conditions, the Japanese government launched a series of fundamental reforms covering all aspects of the ailing economy; economic structure, fiscal structure, administration, financial system, social security, and education.

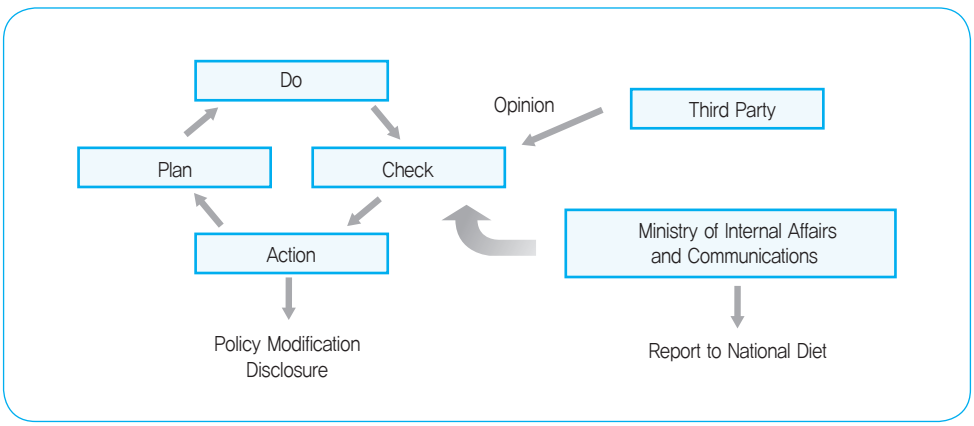
During the reform process, one area that drew special attention in the context of government accountability is policy evaluation. In December 1997, citing the following three important reasons, the Special Commission for Government Reform published the final report and recommended introducing the "Policy Evaluation System" at the central government level. First, past practice in the administration was to put too much emphasis on making rules and securing budgetary resources that assessment on policy effects was relatively neglected and re-examination of policy processes based on response to changes in social and economic environment drew little attention. Second, it is always very important, especially during the implementation stage to examine effects of policies and modify the process to improve performance so that rigorous and objective evaluations, *ex-ante* or *ex-post*, should be carried out and the evaluation results should be utilized to improve the performance. Third, policy evaluation could be utilized as a vehicle through which communication between policy making and implementation is significantly improved. Moreover, transparency and fairness can be promoted by releasing the information on policy evaluation.

Following the recommendation of the Commission, Japanese National Diet passed the Central Government Reform Act in June 1998 to provide a legal foundation for policy evaluation. However, it was not until the enactment of the Policy Evaluation Act in June 2001 that a full-scale policy evaluation started to be designed and carried out.

4.2. Performance Management System in Japan

Standard Guideline for Policy Evaluation, first drafted in 2001, is the most important document occupying the central position in Japanese performance management system and prescribes the principles and standard procedures all public agencies should follow in policy evaluation. The Guideline defined policy evaluation as "the process through which a government agency policy measures and analyzes policy effects and produces objective and scientific information useful for design and implementation of policies". Policy evaluation can be best understood as a component of policy management cycle consisting of four steps – plan, do, check, and action.

Figure 3-3 | Policy Management Cycle and Policy Evaluation in Japan



The Guideline lists three main purposes of policy evaluations. First, policy evaluation contributes to enhancing the quality and effectiveness of administrative services. Policy evaluations may help the government define the boundary of public sector involvement in providing various services and leave out private sector players the services that do not require involvement of the public sector. Efficient division of labor between private and public sectors makes it possible for the public sector to provide necessary services for low costs. Second, policy evaluation may lead the administrative system to outcome oriented framework. Policy evaluation provides information on how much input was used, how much output was produced, and how the outcome of the policy affected people's life. In particular, the information on effectiveness of government policies can be utilized in designing an outcome sensitive incentive scheme so that the administration put more emphasis on the outcomes rather than inputs. Third, policy evaluation may be a very useful tool to promote accountability of the administration.

The subjects of policy evaluation can be divided into three groups; schematic policy, action policy, and action plan. Schematic policy indicates activities of the administration to realize the basic guideline in carrying out tasks assigned to the administration. Action policy is concrete action or measure to accomplish objectives of schematic policy. Action plan is the basic unit of administrative activities and indicates individual steps and procedures to achieve the objectives of action policy.

Policy evaluations consist of three different kinds of evaluations – program evaluation, result evaluation, and comprehensive evaluation. Each department or agency in the administration selects the type of evaluation considering the nature and characteristics of the policy.

4.2.1. Program Evaluation

Program evaluation can be categorized into *ex-ante*, *interim* or *ex-post* evaluation according to the timing of evaluation. Selection of evaluation timing should be made by taking into accounts various factors such as the purpose of evaluation, characteristics of the policy, and availability of information required for evaluation.

Ex-ante evaluation is useful when the evaluator intends to provide information on whether to start the program or policy at all or which alternative to choose among many competing programs or policies. Japanese law on policy evaluation requires *ex-ante* evaluation on the programs or policies that are expected to have significant implication on the economy or people's welfare. In addition, *ex-ante* evaluation should be carried out for R&D programs and official development assistance (ODA) programs that require a large amount of budgetary resources. *Ex-ante* evaluation should examine;

- whether the objectives of the program is well aligned with those at higher level.
- whether it is adequate for the government to intervene to provide service from the perspective of the proper roles of the government.
- whether the program contributes to accomplishing the higher level objectives.
- whether it is possible to obtain benefits that exceed the costs of the program.
- whether there is a more effective alternative.
- whether the benefit and cost of the program can be shared fairly.
- whether the program should be launched before other programs in terms of priority.

Ex-post evaluation is conducted after the program is terminated or a significant amount of time has already elapsed since it was started. The purpose of the evaluation is to examine the effectiveness of the program empirically and to provide information for improvement

of the program implementation. *Ex-post* evaluation is carried out when there arises the need for re-examination or improvement of the program as the social and economic environments change. Head of any public agency planning to conduct *ex-post* evaluation should announce the plan at the start of the fiscal year.

The purpose of interim evaluation is to examine progress of the program under implementation by checking whether it is on the path to fulfilling the goals. Also, the evaluation is conducted to provide information that can be used in re-design or modification of the program in response to environmental changes.

4.2.2. Result Evaluation

In result evaluation, the evaluator compares the actual performance of the program with the target level prescribed when the program started. Result evaluation can be done on a program while it is being implemented or when it is terminated. The information from the result evaluation can be used for the improvement of the program performance or enhancement of accountability of the program agency.

The evaluator, first of all, should select an outcome indicator as the object of comparison. When it is impossible or inappropriate to select an outcome indicator that is easy to measure and understand, an output indicator could be a practical alternative. In selecting indicators as the reference for evaluation, the most important criterion is the measurability. Therefore, they should be quantitative in nature or at least qualitatively assessable by experienced evaluators. The target level should be set up in a clear and concrete manner and the way to determine the level of achievement should be announced in advance. Since the achievement of the performance goal can be affected by external factors that the program agency cannot control, it is always a good practice to specify the external factor that may have affected the result of the program in the evaluation report. One more important thing to note when the evaluator selects the indicator for the evaluation purpose is that he may not want to select an indicator that requires significant amount of resources in measuring it. It is also recommended to describe the logical and practical reason why the indicator was selected. In-depth evaluation like program evaluation or comprehensive evaluation can be employed to scrutinize the performance of the program that showed unsatisfactory results in the result evaluation.

4.2.3. Comprehensive Evaluation

Comprehensive evaluation is an integrated and inclusive assessment procedure through which we can identify the problematic issues in program performance and uncover the causes. Therefore, comprehensive evaluation is generally conducted on the programs that a reasonable amount of time has already elapsed since the program started. The fundamental

purpose of the evaluation is to provide information that is useful in solving the problem. When selecting the program for comprehensive evaluation, the evaluator should establish the priority based on urgency, seriousness of the problem, and workload of the evaluator. Comprehensive evaluation examines the adequacy of program objectives, necessity of government intervention, relative magnitude of benefits and costs of the program, and coherency with other programs.

The following programs are most frequently selected for comprehensive evaluation;

- Programs that require improvement and re-examination in response to changes in social and economic environments.
- Programs that are in high demand for evaluation from the perspective of national interests.
- Programs that have significant implication on people's welfare and the like.
- Programs that are replaced with new ones.
- Programs that were evaluated a long time ago.

4.3. Institutional Arrangement

Each government agency in the administration can be the sponsor of any form of policy evaluation. When necessary, a government agency initiated a policy evaluation and in most cases the evaluation is conducted internally. A policy or program is evaluated in terms of necessity, efficiency, and effectiveness. To ensure the objectiveness and comprehensiveness, each government agency including the cabinet department and the Ministry of Internal Affairs and Communication share the roles in the process of policy evaluation.

Each department is authorized to design, plan, and implement policies and programs by the Government Organization Act and the Cabinet Office Act. The same laws grant each department to conduct internal evaluations on its own policies or programs. Annual evaluation plan should be drafted by the department focusing on the following programs;

- Programs that were newly introduced.
- Programs that have not started or completed even after significant amount of time has elapsed.
- Programs that have not yet been the subject of evaluation since introduced.
- Programs that require modification or review due to changes in economic and social environment.

Policy evaluation division determines the evaluation procedure and method, which are disclosed to the public to ensure objectivity and transparency of the evaluation results. What evaluation division should decide and announce include the purpose of evaluation, step-by-step explanation on the evaluation procedure, basic perspective of the evaluation, evaluation technique, feedback of evaluation results, and communication with outside stakeholders on policy evaluation. In particular, each department offers technical assistance on evaluation technique to program agencies that sponsor and conduct evaluation projects.

In addition, the head of a department in the administration is supposed to make and announce the "basic plan" and "implementation plan" every three to five year as required by the Policy Evaluation Act. Each government organizations and agency should conduct ex-post program evaluation following the plans and disclose the evaluation results to all stakeholders and the general public. The evaluation report should include a detailed description of the evaluation method, data, expert opinion, as well as the evaluation results.

The Ministry of Internal Affairs and Communications (MIAC) is in charge of the overall management of the policy evaluation system in Japan and examines the procedure and results of policy evaluations completed by individual departments in the administration. Division of Administration Evaluation in MIAC oversees the operation of the policy evaluation system that includes policy evaluation, administration evaluation and monitoring, evaluation of independent administrative agency, and administration consulting. Administration evaluation refers to examination on evaluation organization and staff in each department to facilitate smooth execution of policy evaluation. Administration evaluation and monitoring refers to the task of monitoring each department's activities to improve management of administrative tasks. Evaluation of independent administrative agency refers to support for objective and fair evaluation of independent administrative agencies by offering an opinion on the evaluation results reported by the Evaluation Committee in each department. Administration consulting tries to improve the administrative system and its operation with opinion polls on the Division of Administration Evaluation's work.

MIAC plays important roles in ensuring objectivity and fairness of policy evaluations by;

- Examining the level of the objectivity and fairness achieved in the policy evaluation carried out by individual departments.
- Identifying the cases for which a new policy evaluation should be carried out to cope with changes in social and economic environment.
- Conducting policy evaluation to ensure objectivity and fairness of the policy evaluations that were selected for new evaluations for the reasons mentioned above.

- Conducting policy evaluation when joint evaluation between the department and MIAC is decided at the request of the department.

Policy evaluations are carried out based on the following fundamental perspectives;

- Necessity; whether policy objectives are reasonable considering demand from the people and society and whether government intervention can be justified logically or practically.
- Efficiency; whether we can expect policy outcomes corresponding to inputs or costs, whether the planned outcomes can be achieved with less inputs or lower costs, whether better outcomes can be accomplished with the same inputs or costs.
- Effectiveness; whether the desired outcomes can be achieved through the policy or program.
- Fairness; whether the benefits and costs of the policy are fairly shared.
- Priority; whether a particular policy should be evaluated over other policies.

The Board of Audit of Japan is an organization in the administration. It is, however, an independent agency not included in the cabinet. The Japanese Constitution orders the establishment of the Board and examines the annual budget report before being sent to the National Diet. It also conducts audits on the annual accounting report of the administration, public corporations and state-owned enterprises.

Audits by the board are carried out in four steps; planning, preliminary audit, on-site audit, and reporting and feedback. Planning is essential for more efficient and effective audits since the Board has limited resources. Audit plan provides the basic scheme and principle of auditing. In planning the audit, the Board considers budget size of the audited, records on internal and external audits, and the importance of auditing in terms of the public interests. Once selected as a subject of auditing by the board, the public organization or agency should submit various documents that can demonstrate accuracy, legality, and adequacy of budgetary accountings. The Board examines the submitted documents to get a general idea and possible issues before the on-site audit is launched. Completing the preliminary inspection on documents, the Board sends out its staff members to headquarters of the subject of the auditing to conduct on-site auditing. The Board possesses the authority to administer audits on local public organizations that have received subsidies from the central government. The final audit report should include the Board opinion and judgment on accuracy, legality, and adequacy of budgetary accountings. The Board sends the final audit report to the Ministry of Finance and the National Diet along with supplementary documents.

Japan has a parliamentary government and a member of the ruling party holds a position as the minister of a department so the parliamentary monitoring on the administration does not have so much meaning or implications as other countries where the separation of the parliament and the administration is the fundamental principle in the government organization. The administration reports the results of the policy evaluations and their feedback to the budget process.

The Ministry of Finance is not directly involved in the policy evaluation and the feedback of evaluation results into the budgeting process is carried out when each department drafts an annual budget request and modifies rules and regulations related to the programs by taking evaluation results into consideration.

2012 Modularization of Korea's Development Experience
Performance Management System of Budgetary Programs in Korea

Chapter 4

Performance Management System of Budgetary Programs in Korea

Performance Management System of Budgetary Programs in Korea

1. Background

The history of the performance management system in Korea dates back to 1999 when the Ministry of Planning and Budget launched the pilot project to introduce performance based budget by asking 16 selected ministries in the administration to submit performance plans along with budget requests for the next year. Unsatisfactory experience from the pilot project prompted the Korean government to push forward the process of establishing the performance management system with stronger and more effective legal and institutional foundations. Ever since, a series of important institutions related to performance management in the public sector have been introduced, such as the Performance Goal Management of Budgetary Programs in 2003, Self-Assessment of the Budgetary Programs in 2005, and In-depth Evaluation of Budgetary Programs in 2006.

The fundamental driving forces behind the introduction of the new institutions throughout the 2000's are the changes in social and economic environments and budget office's responses to establish a more efficient and effective budget system. Korea experienced a dramatic increase in public debt after the Asian financial crisis in late the 1990s. The growing debt was mainly driven by a rapid increase in public expenditures to strengthen the social safety net which became an urgent policy agenda in response to widening income disparities resulting from the economy-wide restructuring. Moreover, the Korean economy faced huge challenges of rapidly aging population and slow economic growth, which casts seriously negative implications on the long term sustainability of the budget. The grim prospects of the future fiscal conditions propelled the Korean government to initiate a fundamental reform process.

The reform of performance management system was pursued in the context of a larger reform framework known as the Four Major Fiscal Reforms, which provided an extraordinarily favorable environment for building up an effective performance management system. Due to such a big push forward on a large scale, the Korean government was able to establish a comprehensive and robust performance management system in a short period of time. The Four Major Fiscal Reform consisted of the establishment of a medium-term expenditure framework known as the National Fiscal Management Plan, introduction of top-down budgeting, establishment of the performance management system, and building of a digital budget information system. The medium-term fiscal plan puts government spending decisions in a five-year framework. Based on prudent projections on future economic growth, the plan determines the level of annual overall expenditure over the medium term and allocates the total amount available among major sectors of government spending. Consistency between such medium-term resource allocation decisions and annual budget appropriations is enforced through the top-down budgeting system. The system assigns firm spending ceilings on the expenditure of each ministry according to the medium-term fiscal plan, but delegates lower-level budgeting decisions to ministries, provided that the latter's aggregate expenditures remain within their assigned ceilings. The greater autonomy given to the ministries in turn requires greater accountability on their part. This is ensured through the performance management system, which was introduced to monitor and analyze the performance of government spending programs and thus strengthen the link between budgeting and performance. The digital budget information system allows the budget office to monitor ministries' spending in real time.

Performance management system was introduced to Korea in four phases. The first phase was the experimental pilot project carried out during 2000-02. The project experimented with a modified version of GPRA (Government Performance and Results Act) in the United States. While the GPRA requires each agency to submit strategic plans, annual performance plans and annual performance reports for every single program, the Korean version requires performance plans and reports only for major budgetary programs over USD 1 million in size. The twenty-two ministries and program agencies that participated in the project were asked to develop annual performance plans.

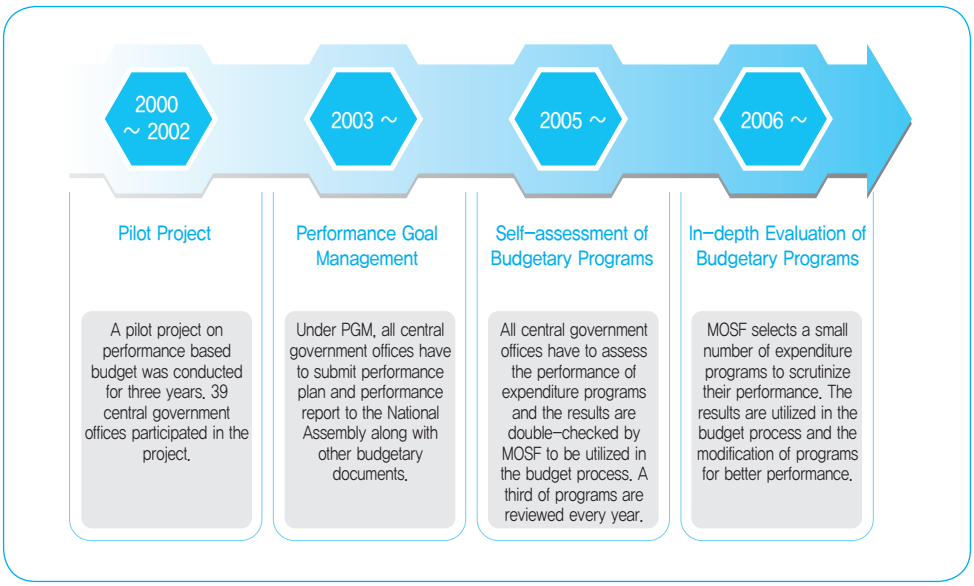
The pilot project was terminated as the new administration was inaugurated. Building on that experience, the second phase began as a core component of the Four Major Fiscal Reforms in 2003. Twenty-two ministries and agencies were selected and asked to submit their performance plans along with their annual budget requests. The initiative was named the performance goal management of budgetary programs (PGM) and fully implemented from 2006 for all government expenditure programs managed by the central government offices. PGM is a modified version of performance monitoring and occupies the upper echelon

of the performance management system in Korea. Under PGM, each central government office is requested to produce both the performance plan of the next year and performance report of the previous year that would be sent to the National Assembly as a component of budgetary documents. Based on layers of hierarchical concepts of mission/vision, strategic goals, performance goals, and tasks, each central government office constantly monitors the performance of expenditure programs under its jurisdiction by regularly measuring performance indicators and comparing them with the target levels.

The third phase took place in 2005 along with the introduction of the self-assessment of budgetary programs (SABP). The initiative was basically a modified version of performance review based on the Program Assessment Rating Tool (PART) of the United States. Under the SABP, about a third of all government expenditure programs have been reviewed every year jointly by the program offices and the Ministry of Strategy and Finance (MOSF), which would allow MOSF to review every major budgetary program over a three-year cycle. Each central government office was asked to fill out the checklist that includes questions on planning, management and results of a government expenditure program. The size and priority of the program can be adjusted reflecting the results of SABP. In some extreme cases, programs are completely restructured or terminated. The primary purpose of SABP was to hold central government offices accountable for the performance of government expenditure programs they are responsible for and to supplement PGM by establishing an explicit link between performance and budget of the central government offices.

The fourth phase started in 2006 with the launch of In-depth Evaluation of Budgetary Programs (IEBP). IEBP is a Korean version of program evaluation that examines the performance of government expenditure programs with analytical and scientific methods typically by external experts. IEBP is a comprehensive and analytical investigation on various aspects of budgetary programs such as effectiveness, relevance, and efficiency. The results of IEBP are incorporated into the budget process to improve program performance.

Figure 4-1 | Four Phases of Performance Management System in Korea



2. Performance Goal Management

2.1. Introduction

In 2003, MOSF launched the Performance Goal Management (PGM); a modified version of performance monitoring that benchmarked the performance management system introduced by the Government Performance and Results Act (GPRA) in the United States in 1993.

Under PGM, all central government offices including ministries in the administration and independent government bodies such as the National Assembly, the Supreme Court, the National Election Commission, and the Board of Inspection and Audit, first of all, are asked to set up performance goals and performance indicators that can contribute to the accomplishment of long term strategic goals, which in turn is believed to have a close relationship with the mission of the office. Then, each office should make the performance plan for the next fiscal year and submit it to MOSF along with the budget request. MOSF is responsible for the overall management of the system including examination of the performance plan, communication with the central government offices, and incorporation of the results of PGM into the budget process. The final step in PGM is to measure the current state of performance indicators and compare them with the target levels. Like other

components of the performance management system, the ultimate purpose of PGM is to incorporate the performance information into budgetary progress so that effectiveness and efficiency of government expenditure programs are enhanced. The results of PGM are presented to the National Assembly as well as MOSF as a component of official budgetary documents. All government activities should be included in PGM except for the ones that require no expenditure or assure little benefit of performance evaluation like salaries and general administrative expenditure.

PGM started with 22 ministries[27] of the administration in 2003 and each participating ministry was asked to establish strategic goals and performance goals as well as performance indicators for 30% of expenditure programs that each was in charge of. The first measurement of performance was done in 2004. In addition, the construction of performance goals and performance indicators for the remaining programs in those 22 ministries was completed in 2004 and the measurement started the next year. Other central government offices started to follow suit from 2004. In principle, all programs accompanying government expenditure should be the subject of PGM. The exceptions include grants to local government, concession funding for local government and local education authority, maintenance expenditure, internal transactions, reserves, and other minor expenditures. Administrative costs are also excluded from PGM but expenditure that is not current expenditure and requires regular performance management should be included in PGM. In addition, expenditure programs on public relation activities, small size research funds for policy development, and routine expenditure on information technology such as replacing the obsolete computer equipment are also exempt from PGM. Programs related to national security are the subjects of PGM but the results are not made public.

In 2004, the program budget budgeting system was introduced as a part of the budget reform launched back in 1999. Program budgeting is the budgeting system that, contrary to conventional budgeting, describes and gives the detailed costs of every activity or program that is to be carried out in a budget. Objectives, outputs and expected results are described fully as are necessary resources and costs, for example, raw materials, equipment and staff. Sum of all activities or programs constitute the program budget. A program budget is the basic unit of the performance management system, which facilitates the important elements in the system such as strategic budget allocation and performance evaluation. The introduction of program budgeting made it crucial for Korean government to establish a

27. They are the Ministries of Education, Public Administration and Local Government, Science and Technology, Culture and Tourism, Agriculture and Forestry, Industry and Resources, Information and Communication, Welfare, Environment, Construction and Transportations, Maritime and Fishery, Patriot and veterans Affairs, Public Relations andRand Development Administration, Forestry Service, Intellectual Property Office, Public Procurement Service, National Policy Agency, Coast Guard, Meteorological Administration, Statistical Office.

robust and efficient performance evaluation system including performance monitoring and program evaluation.

In 2006, MOSF drafted the Guideline for Performance Goal Management that describes the procedures of PGM that all central government offices should follow. Moreover, the coverage of PGM was significantly extended from 26 to all 48 central government offices and from 46 to all 61 public funds managed by the central government offices. Thanks to the continuous efforts to extend the coverage since its introduction in 2003, all government expenditure programs became the subject of PGM in 2006.

The legal foundations for PGM are provided by the National Finance Act and the National Account Act, enacted in 2007 and 2009, respectively. Article 8 of the National Finance Act makes the heads of central government offices and managers of public funds responsible for establishing the performance management system for the expenditure programs they manage and requires them to prepare the performance plan and performance report and submit them to MOSF and the National Assembly. Moreover, Article 34 and 71 of the National Finance Act explicitly make the performance plan an official budgetary document as a part of the budget bill submitted to the National Assembly. As for the performance report, Article 14 of the National Account Act requires it as a part of official documents in the report in the final account of the budget.

Table 4-1 | Legal Foundations of Performance Goal Management

Act	Articles
National Finance Act	Article 8. ① The head of each central government office and the manager of each public fund designated by law should establish the performance management system. ② The head of each central government office should submit the performance plan of the next year and the performance report of the current year to the Minister of MOSF when he submits the budget request and the manager of each public fund should submit the performance plan of the next year and the performance report of the current year when he submits the management plan of the public fund. ③ The head of each central government office and the manager of each public fund should make performance report as required by the National Accounting Act. ④ Deleted. ⑤ The Minister of MOSF should inform the head of each central government office and the manager of each public fund of the guidelines for the performance plan in ②.

Act	Articles
National Finance Act	⑨ The consistency of program expenditure as well as core program contents should be maintained in the performance plans based on the budget bill in Article 33, the revised budget bill in Article 35, management plan of public fund in ① of Article 68, and revised management plan of public fund in ② of Article 70. Article 34. The following documents should be accompanied by the budget bill submitted to the National Assembly as directed by the Article 33. 8. the performance plan in ② of the Article 8. Article 71. The Administration or the manager of each public fund should submit the following documents when submitting management plan and revised management plan of the fund to the National Assembly as directed by ① of Article 68 and ② of Article 70. 4. the performance plan in ② of the Article 8.
National Account Act	Article 14. The budget report on final accounts consists of the following. 4. performance report. Article 15. ④ The performance report in 4 of Article 14 should be prepared by comparing the performance goals as directed by Article 8 of the National Finance Act and the actual accomplishments.

Source: http://www.law.go.kr/main.html. Translation is provided by the author.

It was not until 2007 that a comprehensive PGM system covering all expenditure programs with some significance was first established. All central government offices and public funds with expenditure programs submitted performance reports to MOSF and the National Assembly for the first time in 2007.

Table 4-2 | Introduction of Performance Goal Management

Year	Measures
2003	• Performance goals and performance indicators for 30% of expenditure programs in 22 ministries were developed.
2004	• Program budgeting system was introduced to strengthen the efficiency and accountability of government budget and expenditureprograms. • The first measurement of performance indicators by the leading 22 ministries was conducted and performance plans for 2005 were prepared.
2006	• The Guideline for PGM was drafted and distributed among relevant government bodies. • The coverage of PGM was extended to include all central government offices and public funds with government expenditure programs.

Year	Measures
2007	• The National Finance Act was enacted and the legal foundation of the performance management system including PGM was laid.
2008	• A comprehensive set of the performance report for 2007 for all central government offices and public funds was submitted to MOST and the National Assembly.
2009	• The National Accounting Act was enacted.
2010	• The performance reports for fiscal year 2009 were submitted.

2.2. Structure of PGM

As of December 2012, 50 central government offices and 65 public funds managed by the central government are required to submit the performance plan as directed in the National Finance Act when they make their budget requests to the National Assembly.

The performance plan is a component of official budgetary documents that describes the execution plan to accomplish the organization's strategic goals and performance goals. Therefore, the document should explicitly contain the mission and strategic goals of the government office and performance goals and performance indicators of the current year as well as the performance information of the three previous years. The plan also may include a medium- or long-term plan of the office incorporating the medium-term fiscal management plan according to the National Finance Act.[28] More specifically, in preparing the performance plan, each government provides information on important performance results achieved in previous years, policy directions of the current year, and current status of organizational and fiscal conditions. The plan should also describe the office's system of performance goals and include a detailed description of the programs.

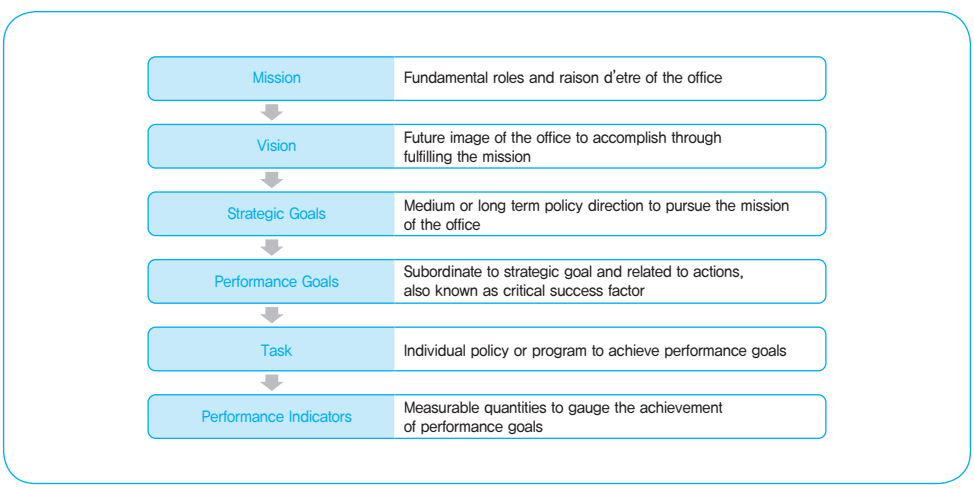
It is very useful to understand the performance plan as a schematic design consisting of six hierarchical concepts such as mission, vision, strategic goals, performance goals, task, and performance indicators. The mission is the raison d'etre and important functions of the government office. It should be specified in a result-oriented way and comprehensive to include all important roles and functions of the office. However, defining the mission too broadly that it overlaps with those of other government offices should be avoided. Vision is the long term objective of an organization and epitomizes the future image to be accomplished. The vision should be drawn up based on a precise diagnosis of the current state of the organization. The value of vision is to provide motivation to the members of the organization. Therefore, it should be written in easy to understand, comprehensive and encouraging expressions.

28. Article 7 of the National Finance Act postulates that the government should prepare the Fiscal Management Plan for the next five or longer years and submit to the National Assembly by 90 days before the next fiscal year starts.

Strategic goal summarizes medium- or long-term policy direction that the central government office should pursue to fulfill the mission and the vision. Logical and practical linkage between strategic goals and the mission of the office should be clearly established. Strategic goals should be defined in a concrete and clear manner so that the public can grasp the overall picture of the future direction for the next five years. It is strongly recommended to use a short but strong slogan-like expression for strategic goals to convey core messages. MOSF explicitly asks all government offices to incorporate the relevant issues from the Medium Term Fiscal Management Plan into the strategic goals so that the consistency among plans with different time horizons is preserved. The number of strategic goals should be large enough to cover all important objectives and functions of the office but small enough to make it possible to conduct efficient and effective performance management.

Performance goal is a bridge that links the mission or strategic goals with program activities. First of all, performance goal is a lower level target that leads to the accomplishment of strategic goals and therefore should be result oriented to make the identification of the program's impacts easy. Next, performance goals can be utilized as interim targets that should be reached to accomplish longer term objectives like strategic goals. Third, there are no restrictions on the number of performance goals for each strategic goal but special care should be taken to avoid too much asymmetry in the distribution of performance goals across strategic goals. Finally, performance goals should be written as specific and concrete as possible that their achievement can be identified without significant difficulty. The usefulness of performance indicators crucially depends on the quality of definitions on concepts in higher layers in the system of performance goals illustrated in [Figure 4-2].

Figure 4-2 | System of Performance Goals



Task means individual policies or programs that are conducted to accomplish the performance goal. In principle, one task is supposed to be assigned to one program and it should be avoided to assign multiple tasks to a program except for the case when it is absolutely necessary. It is needless to say that a reasonable and detailed action plan to carry out the task should also be provided.

Performance indicator constitutes the component at the lowest layer of the system in [Figure 4-2]. It should be defined as objectively and quantitatively as possible that the government office in charge of the program can routinely measure the level of attainment with ease. It is strongly recommended that at least one performance indicator is assigned to each performance goal and task. Both quantitative and qualitative performance indicators can be utilized in practice but the program agency should avoid using qualitative ones when it can. In addition, outcome indicators should be selected as performance indicators as far as it is possible but process or output indicators are allowed when it is impossible to find appropriate outcome indicators. MOSF recommends all government offices to consult international evaluation indices suggested by IMD, WEF and UN to secure objectivity and reliability of performance indicators.

The operational procedure of PGM is illustrated in <Table 4-3>. Every government office prepares and submits to MOSF the performance plan for fiscal year N by the end of June of fiscal year (N - 1). Then, MOSF reviews the performance plan by the end of August and informs each government office that it should revise the performance plan by incorporating MOSF's review opinion. Each government office sends to the National Assembly the revised performance plan along with the budget bill 90 days before the fiscal year N starts. Government offices may again revise the performance plan before the fiscal year N starts to reflect opinions from MOSF and the National Assembly. Once the execution of the budget for the fiscal year N is completed, all government offices should prepare and submit the performance report to MOSF by the end of February of the fiscal year (N + 1) and MOSE relays it to the Board of Audit and Inspection (BAI) by the tenth of April. The performance report is the document that explains the results of self-assessment on the performance. Completing review and inspection of the performance reports, BAI sends them back to MOSF. Each government office revises the performance plan to incorporate BAI's opinion and submits it as well as the report on final accounts to the National Assembly by the end of May.

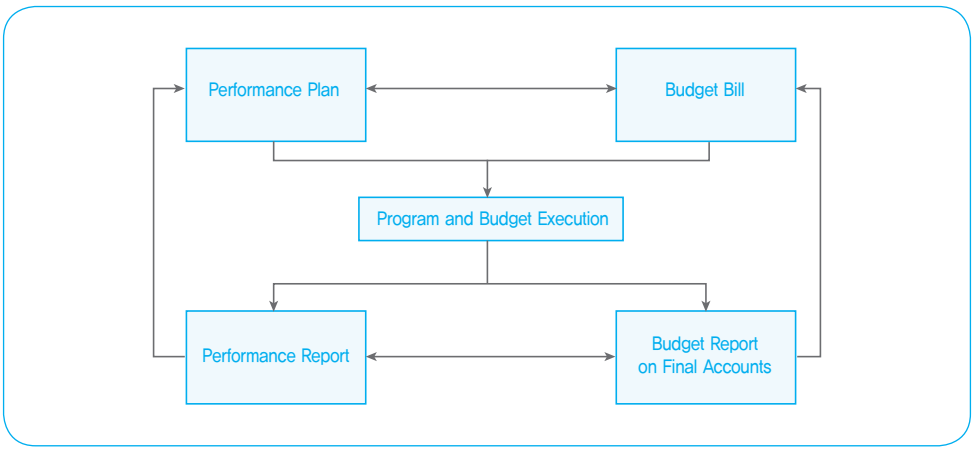
Table 4-3 | The Operational Procedure of PGM

Time		Actions
Fiscal Year (N-1)	May	• MOSF distributes the Guideline for PGM.
	by 30th of June	• Government offices prepare and submit performance plans to MOSF.
	July ~ August	• MOSF reviews and asks revision.
	by 90 day before new fiscal year	• Government offices revise performance plans incorporating MOSF's opinions and submit it with the budget bill to the National Assembly.
	by 31th of Dec.	• Government offices may revise submitted performance plan to incorporate the opinion of MOSF and the National Assembly.
Fiscal Year N	whole year	• Government offices execute the budget.
Fiscal Year (N+1)	January	• MOSF distributes the Guideline for the performance report.
	by 28th of Feb.	• Government offices prepare and submit the performance report to MOSF.
	by 10th of April	• MOSF relays the performance report to BAI for review and inspection.
	by 20th of May	• BAI send the results of review and inspection to MOSF.
	by 31th of May	• Government offices incorporate the opinion of MOSF and BAI into the performance report and submit it to the National Assembly along with the report on final accounts.

A complete set of the performance reports was submitted to the National Assembly on May 31st, 2010, for the first time that it became finally possible to establish a feedback system of performance evaluation into the budget process in the National Assembly. The National Finance Act and the National Accounting Act as summarized in <Table 4-2> demand MOSF and the National Assembly to institute a system that establishes an explicit linkage between the performance plan and the budget bill at the planning stage in addition to a strong linkage between the performance report and budget report on final account at the evaluation stage.

It is not easy to directly relate the performance information derived from the performance plan or the performance report to the budget process due to the difficulty in measuring performance and costs but much effort should be made to build a linkage between PGM and the budget process as explicitly as possible in order not to make the performance management system, including the PGM, an additional administrative burden on government offices.

Figure 4-3 | Linkage between PGM and the Budget Process



2.3. Assessment on the Performance of PGM

According to the performance plans for 2013 submitted to MOSF and the National Assembly, 179 strategic goals, 444 performance goals, and 2,138 tasks are currently operated by 50 central government offices, which is a significant decline from the previous year. There were 191 strategic goals, 479 performance goals, and 2,155 tasks in the performance plan for 2002. It is thought that the decline in the number of performance goals and related indices is mainly due to the effort by MOSF to streamline the performance management system by consolidating overlapping programs across different program offices and eliminating ineffective or irrelevant programs. Sixty-eight percent of all government expenditure is covered by the tasks reported in the performance plan for 2013. Despite continuous effort by the Korean government, the coverage of the performance management system does not seem to have reached the satisfactory level yet.

The number of performance indicators for performance goals in the performance plan for 2013 is 1,298 so that on average 2.92 performance indicators are assigned to a performance goal. The distribution of performance indicators shows that much emphasis is put on outcome indicators consistent with the aim of result oriented performance management the Korean government is pursuing. Out of a total 1,298 performance indicators 791 (60.9%) are outcome indicators and 395 (30.4%) are output indicators while only 112 indicators (8.7%) are either input or process indicators. As for performance indicators for tasks, 50 central government offices report 5,139 indicators, of which 2,877 (56.0%) are outcome indicators and 1,922 (37.4%) are output indicators.

Despite the short history, PGM is believed to have established itself as an indispensible element in the performance management system in Korea, contributing to enhancing the effectiveness and transparency of government expenditure programs by making central government offices accountable for their own decisions. However, several commentators[29] argue that there still is plenty of room for improvement in PGM in Korea.

First, more effort should be made to set the appropriate target level for performance indicators. In principle, the target level for a performance indicator should be feasible considering the constraints imposed by various policy environments. There is no doubt that performance targets can be lowered, if necessary. However, a typical practice is to raise the target to a higher level every year even if the program agency fails to achieve the target level for the current year or it is almost impossible physically or economically to accomplish the performance target. The main reason for the ever rising performance target is that the size of budget is linked to the target level of performance indicators in many cases. Though the linkage between the budget size and program performance is one of the most effective ways to motivate better performance of government expenditure programs, care should be taken in order not to distort the incentive structure of the program personnel by applying without due regard to particular circumstances each program is facing. Both too much conservatism and groundless optimism should be avoided. Recognizing the issue of the ever escalating performance target, MOSF requires all central government offices to provide an explanation or justification for the selection of performance targets in the performance plan.

Second, it turns out to be very important to get every program personnel involved in all stages of the performance management system. Most of the program offices designate a small group of staff members to handle various tasks related to the performance management system including planning, execution and monitoring. It is not unlikely that the group in charge of the performance management system made enough effort to solicit for inputs from other program personnel and set unrealistic performance targets in preparing the performance plan. The lack of communication may lead to setting up unrealistic performance targets, which in most cases results in an unenthusiastic attitude toward the performance management system among program personnel.

Third, much more effort should be taken to enhance the reliability of performance information provided by PGM. Program offices are constantly exposed to the temptation to distort the reality in preparing the performance plan and the performance report to make their performance look as good as possible. MOSF asks them to make both the performance plan and the performance report strictly following the guideline distributed in advance so that the risk of distorting information contents of PGM can be minimized.

29. For example, see NABO (2012), Park *et. al.* (2012).

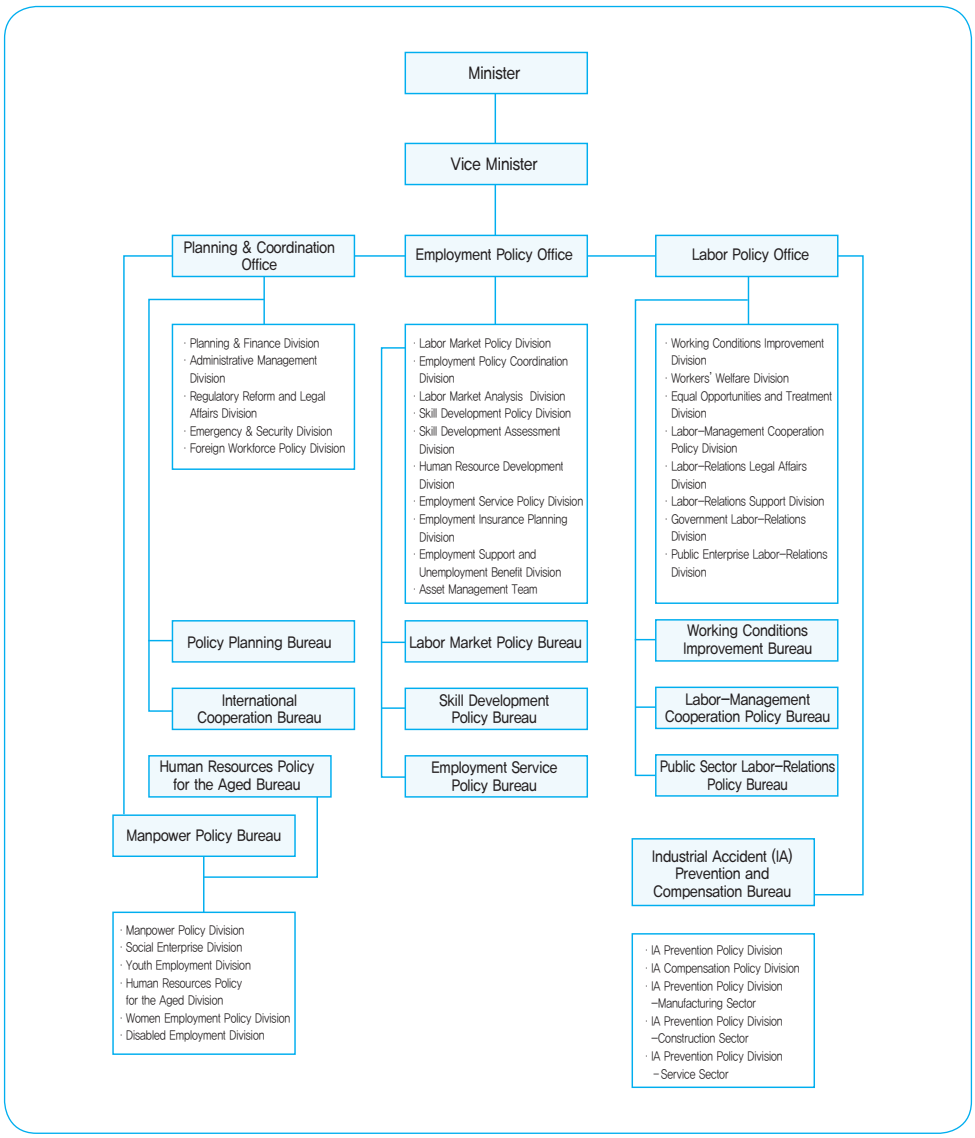
Fourth, the utilization of performance information should be further facilitated. Several measures can be taken to make performance information from PGM more useful. Program offices may provide education classes and consulting services to promote expertise of staff members on the performance management system. It would be useful to appoint performance officers to stimulate interest of all members on the performance management system. Each performance indicator is assigned to a member of the program office that takes the initiative in managing the indicator. In addition, a tighter review of the performance plan is suggested to reinforce the usefulness of performance information PGM provides. There is little formal monitoring on preparation of the performance plan even though it is the core document in PGM. A program office may establish a self-monitoring scheme in which a member of the office is selected to conduct a review of the performance plan or appoint an independent outside expert as the reviewer. Moreover, the program office may invite comments and assessment on the performance plan from the stakeholders and the general public.

Fifth, information exchange among program offices should be further promoted. Workshops might be a useful platform to share experiences and exchange ideas. Experts can be invited to train program personnel as well as performance managers to establish a better system for PGM.

2.4. An Example; Performance Plan of the Ministry of Employment and Labor

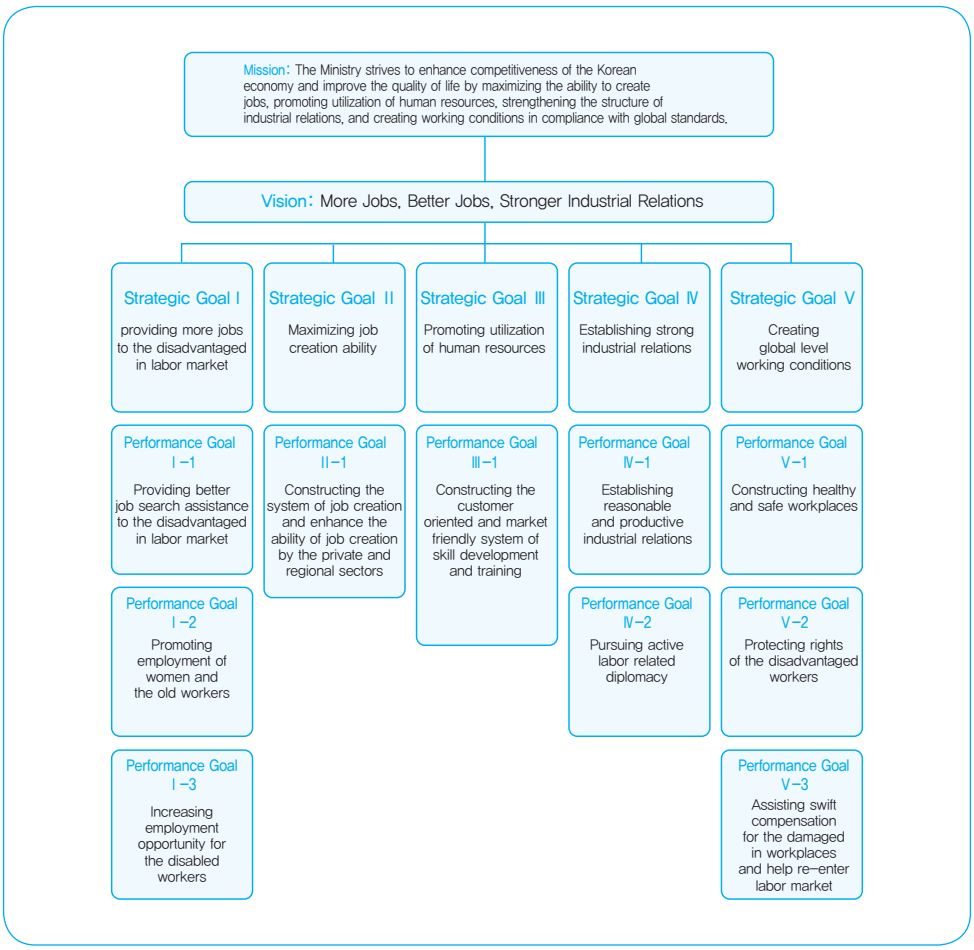
To provide an idea on the practice of PGM in Korea, we present an example of the system of performance goals extracted from the performance plan of 2011 for the Ministry of Employment and Labor (MOEL). MOEL is responsible for establishing and coordinating employment and labor policies, employment insurance, vocational skills development raining, equal employment, work-family balance, labor standards, workers' welfare, industrial relations adjustment, cooperation between labor and management, occupational safety and health, industrial accident compensation insurance, and other affairs. As of 2011, MOEL employs 541 civil servants and the budget was 12,659 billion Korean won. The organizational structure of MOEL is illustrated in [Figure 4-4].

Figure 4-4 | Organizational Structure of MOEL



The system of performance goals of MOEL is illustrated in [Figure 4-5] and <Table 4-4>. MOEL provides five strategic goals and 10 performance goals along with 107 tasks. Sixteen performance indicators for 10 performance goals are summarized in <Table 4-5>.

Figure 4-5 | The System of Performance Goals: the Ministry of Employment and Labor



Source: Performance Plan for 2013, the Ministry of Employment and Labor.

Table 4-4 | Performance Goals and Tasks; the Ministry of Employment and Labor

Performance Goal	Task
Strategic Goal I: Providing more jobs to the disadvantaged in labor market	
Performance Goal I-1: Providing better job search assistance to the disadvantaged in labor market	(1) Assistance of youth intern program
	(2) Assistance of the young to search for jobs
	(3) Assistance of youth intern program in SME
	(4) Fostering young global leaders
	(5) Assistance of career development program for the youth
	(6) Employment maintenance subsidy
	(7) Employment promotion subsidy for disadvantaged workers
	(8) Assistance of job creation
	(9) Employment subsidy for construction workers
	(10) Labor market analysis and guidance for career development
	(11) Unemployment insurance
	(12) Assistance of business start-up by the long-term unemployed
Performance Goal I-2: Promoting employment of women and the old workers	(1) Building better environment for gender equality in employment
	(2) Assistance of the stability of women employment
	(3) Assistance of motherhood and child rearing
	(4) Promoting employment of old age workers
	(5) Assistance of employment restructuring in SMEs
	(6) Loan program for improving work conditions
Performance Goal I-3: Increasing employment opportunity for the disabled workers	(1) Assistance of equipment for employment of the disabled workers
	(2) Subsidy to promote employment of the disabled workers
	(3) Assistance for standard workplace
	(4) Assistance for assistive technology equipment
	(5) Disabled employment maintenance program
	(6) Promoting employment of the disabled workers
	(7) Vocational education and training for the disabled workers
	(8) Skill competition of the disabled workers
	(9) Research on employment of the disabled workers
	(10) Assistance of improving the image of the disabled workers in workplace
	(11) Information technology improvement of Korea Employment Agency for the Disabled

Performance Goal	Task
Strategic Goal II: Maximizing job creation ability	
Performance Goal II-1: Construct the system of job creation and enhance the ability of job creation by the private and regional sectors	(1) Operation of employment infrastructures
	(2) Expansion of jobs in social enterprises
	(3) Assistance of job creation in corporate social responsibility sector
	(4) Assistance of management of foreign workers
	(5) Labor statistics
	(6) Construction of Korea Job World
	(7) Assistance of employment service in private sector
	(8) Assistance of creation of jobs with regional characteristics
	(9) Employment impacts assessment
	(10) Survey on turnover of non-regular workers
	(11) Youth Employment Academy program
	(12) Job Young Plaza program
	(13) Relocation of Korea Employment Information Service
	(14) Collection of employment insurance premium
	(15) Information management of employment insurance
	(16) Management of job security information network
	(17) Management of employment insurance information network
Strategic Goal III: Promoting utilization of human resources	
Performance Goal III-1: Constructing the customer oriented and market friendly system of skill development and training	(1) Vocational education and training for the unemployed
	(2) Job training for the unemployed in agricultural and fishery sectors
	(3) Assistance of skill development and training of workers
	(4) Contribution to Human Resource Development Service of Korea
	(5) Assistance of vocational education and training by Human Resource Development Service of Korea
	(6) Assistance of other skill development and training
	(7) Assistance of skill development and training by employers
	(8) Assistance of skill development and training by SMEs
	(9) Loans for skill development and training programs
	(10) Contribution to Korea Polytechnics
	(11) Contribution to Korea University of Technology and Education
	(12) Assistance of vocational education and training by Korea Polytechnics

Performance Goal	Task
Performance Goal III-1: Constructing the customer oriented and market friendly system of skill development and training	(13) Assistance of skill development and training by Korea University of Technology and Education
	(14) Job training for the national key and strategic industries
	(15) Information technology improvement of Human Resource Development Service of Korea
	(16) Information technology improvement of Korea Polytechnics
Strategic Goal IV: Establishing strong industrial relations	
Performance Goal IV-1: Establishing reasonable and productive industrial relations	(1) Establishment of strong industrial relations
	(2) Facilitation of corporation between employers and workers
	(3) Strengthening roles of National Labor Relations Commission
	(4) Fostering industrial relations experts
	(5) Management of National Labor Relations Commission
	(6) Information technology improvement of labor administration
Performance Goal IV-2: Pursuing active labor related diplomacy	(1) Corporation with international labor organizations
	(2) International corporation in labor affairs
Strategic Goal V: Creating global level working conditions	
Performance Goal V-1: Constructing healthy and safe workplaces	(1) Building clean workplace
	(2) Assistance of technology development for occupational safety and health
	(3) Loans to industrial accidents prevention program
	(4) Cultural development for occupational safety and health
	(5) R&D and international corporation in occupational safety and health
	(6) Construction of facilities to prevent industrial accidents
	(7) Assistance to pneumoconiosis workers
	(8) Prevention of pneumoconiosis
	(9) Construction of information system for occupational safety and health
Performance Goal V-2: Protecting rights of the disadvantaged workers	(1) Protection of working conditions
	(2) Assistance to strengthening retirement compensation system
	(3) Assistance to pro bono legal aids
	(4) Substitute payment and refund
	(5) Assistance to enhance workers' welfare
	(6) Subrogate payment of credit guarantee
	(7) Loan program to help workers under hardship

Performance Goal	Task
	(8) Labor supervision
	(9) Assistance of rents for unemployed workers to start a new business
Performance Goal V-3: Assisting swift compensation for the damaged in workplaces and help re-enter labor market	(1) Workers' compensation insurance payment
	(2) Rehabilitation of victims of industrial accidents
	(3) Welfare of victims of industrial accidents
	(4) Assistance to medical services for victims of industrial accidents
	(5) Loans to victims of industrial accidents
	(6) Administration of workers' compensation insurance
	(7) Collection of workers' compensation insurance premium
	(8) Information management on victims of industrial accidents
	(9) Construction of information management system for workers' compensation insurance

Table 4-5 | Performance Goals and Performance Indicators; the Ministry of Employment and Labor

Performance Goal	Performance Indicator
Strategic Goal I: Providing more jobs to the disadvantaged in labor market	
Performance Goal I-1: Providing better job search assistance to the disadvantaged in labor market	(1) the number of the disadvantaged (disabled workers, female household heads, old aged workers, North Korean defectors) under employment
	(2) the number of young workers under employment
Performance Goal I-2: Promoting employment of women and the old workers	(1) women accession rate
	(2) accession rate of old age workers
Performance Goal I-3: Increasing employment opportunity for the disabled workers	(1) the number of disabled workers employed in workplaces with duty of minimum employment
Strategic Goal II: Maximizing job creation ability	
Performance Goal II-1: Construct the system of job creation and enhance the ability of job creation by the private and regional sectors	(1) accession rate
	(2) the number of employment through WorkNet

Performance Goal	Performance Indicator
Strategic Goal III: Promoting utilization of human resources	
Performance Goal III-1: Constructing the customer oriented and market friendly system of skill development and training	(1) satisfaction level of job training program participants
	(2) job skill improvement of job training program participants
Strategic Goal IV: Establishing strong industrial relations	
Performance Goal IV-1: Establishing reasonable and productive industrial relations	(1) the number of work-days lost due to labor dispute
Performance Goal IV-2: Pursuing active labor related diplomacy	(1) the number of participation in diplomatic events related to labor affairs
Strategic Goal V: Creating global level working conditions	
Performance Goal V-1: Constructing healthy and safe workplaces	(1) rate of injured workers due to industrial accident
Performance Goal V-2: Protecting rights of the disadvantaged workers	(1) speed of dealing complaints by disadvantaged workers
	(2) funding rate of retirement corporate pension
	(3) rate of disadvantaged workers under welfare program
Performance Goal V-3: Assisting swift compensation for the damaged in workplaces and help re-enter labor market	(1) rate of back-to-work workers out of victims of industrial accidents

3. Self-Assessment of Budgetary Programs (SABP)

3.1. Introduction

Another important component of the performance management system in Korea is the Self-Assessment of Budgetary Programs (SABP) under which a third of all expenditure programs of the central government offices undergo review on performance. Benchmarking the Program Assessment Rating Tool (PART) in the United States, it was introduced to supplement PGM. Each central government office conducts a standardized assessment on the performance of expenditure programs under its own management based on the checklist provided by MOSF and the reviewed results are sent to MOSF for a double check. The

final results are sent back to each government office along with recommendations for improvements in program performance. Based on the recommendations, government offices modify their programs. In addition to the feedback to government offices, MOSF also uses the final results of SABP in budgeting. The size and priority of expenditure programs may be re-examined and adjusted based on the results of SABP. In some extreme cases, programs are completely restructured or terminated. According to MOSF, SABP was introduced to make use of review results to improve program performance by aligning the incentive structure of stakeholders by establishing a clear linkage between performance and budget and reshuffling or terminating programs that turned out to have similar or overlapping objectives.

SABP was first introduced in 2005 as MOSF issued the guideline for SABP and the legal foundation for the review was laid by the enactment of the Framework Act on Government Performance Evaluation and the National Finance Act. The National Finance Act stipulates that the Minister of MOSF should conduct evaluations on important budgetary expenditure programs and take into account the results in the budget process. The primary purpose of SABP is to hold the program offices accountable for program performance. Top-down budgeting system introduced in 2004 empowered each ministry with more autonomy in the budget process. In exchange for the increased degree of freedom, MOSF required more accountability from the ministries and the introduction of SABP was an example.

SABP is an important tool for MOSF to monitor program performance and feed the results back into the budget process. Another important reason MOSF took the initiative in introducing SABP was to supplement PGM that had already been in place. PMBP was institutionalized to enhance the performance awareness of the central government offices in charge of expenditure programs by making it a routine procedure to measure performance and compare it with targets. However, PMBP has a fundamental limitation in that it did not require an explicit link between performance and budget. SABP was introduced to bridge the gap. Third, SABP was established as an evaluation tool on the government expenditure programs as a part of the integrated evaluation system on government performance launched by the enactment of the Framework Act on Government Performance Evaluation.

3.2. The Procedures of SABP

At the start of the fiscal year, each central government office selects a third of expenditure programs under its management as the targets of SABP after consulting with MOSF. In selecting the target programs, government offices are advised to ignore expenditure programs that are believed to yield little explicit benefit from performance evaluation. For instance, recurring administrative costs such as wages and salaries, grants to local government, and expenditure-only items such as reserves can be dropped from the list of

target programs. It is also recommended by MOSF as a criterion in selecting the target programs to consider significance of the programs. The following are examples of the indicator of significance; duplication or overlapping with other programs, improvement of delivery system or resource allocation process, modification of performance objectives or indicators, establishment of monitoring system on on-going programs, program evaluation and performance analysis, and enhancement of program achievement.

Once programs are selected as the target for SABP, the evaluation division of each government office asks all divisions or agencies in charge of individual programs to conduct self assessments according to the guidelines and manuals provided by MOSF. MOSF is entrusted with the authority to oversee the overall operation of SABP by the National Finance Act and its Enforcement Decree as summarized in <Table 4-6>. When the self assessment is completed by each division and agency, the preliminary report on the results is submitted to MOSF. Then, MOSF examines the internal and external consistency of the report and conformity of the report with the Guideline distributed earlier in the evaluation cycle. In examining the reports submitted by the government offices, MOSF is aided by external institutions with expertise in performance evaluation. Currently, the Center for Performance Evaluation and Management (CPEM) of the Korea Institute of Public Finance (KIPF) and National Information Society Agency (NIA) participated in examining the preliminary SABP reports prepared by the government offices. The reports are then sent to the Advisory Board for Evaluation of Budgetary Programs for review. The Advisory Board consists of experts in performance evaluations from research institute, universities, and other relevant private sector. The opinion of the Advisory Board then is sent to MOSF that, in turn, incorporates the opinion of the Advisory Board to draft the revised SABP report. MOSF then solicits appeals and opinions on the revised report from the related government offices. The final SABP report is completed and made public to all relevant stakeholders. The results of SABP are incorporated in the budget process to improve program performance by strengthening the link between evaluation and budgetary resource allocation.

Figure 4-6 | Self-Assessment of Budgetary Programs

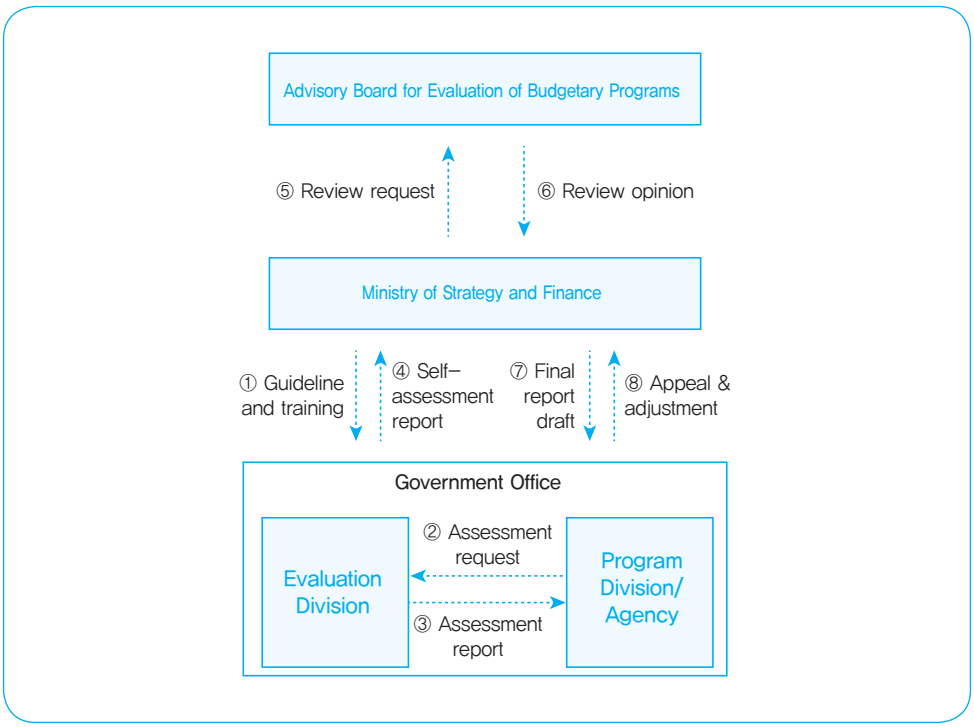


Table 4-6 | SABP and the National Finance Act

Act	Article
National Finance Act	Article 8 ⑥ The Minister of MOSF may conduct performance evaluation for important budgetary programs according to the enforcement decree and incorporate the evaluation results into the budget management. ⑦ When carrying out the evaluations in 6), the Minister of MOSF can let related institutions with expertise in the area conduct investigation and research, if necessary. ⑧ When carrying out the evaluations in 6), the Minister of MOSF can ask the head of the central government office under the evaluation to submit necessary data and opinion on the evaluation. The head of each central government office should accept the request from the Minister of MOSF without exceptional reasons.
National Finance Act Enforcement Decree	Article 3 The Minister of MOSF can ask the head of each central government office to conduct the Self Assessment of Budgetary Programs as stipulated by the guideline ~~.

3.3. The Structure of SABP

The evaluation criteria in SABP have been frequently updated to adjust the system in response to the changing environment and requests from the government offices and the National Assembly. The number of evaluation questions as of December 2011 is 13. Twelve of them are common questions applied to all programs under SABP and two criteria are applied only to programs in the information technology (IT) area. When it started in 2005, the number of evaluation questions was 15, which was reduced to 13 in 2007, and again to 11 in 2008. The evaluation system underwent a structural change in 2009 when all programs were divided into two sectors, general administration and IT, and then general administration sector are grouped into seven types, SOC investment, equipment and facility, direct service provision, equity investment, loan, grant to private sector, and grant to local government, and IT sector into two types, information system and supporting system for information. Eleven questions were asked for all programs under the self evaluation and one or two additional questions were assigned to programs belonging to each type. In 2010, the number of common questions was reduced to 10 and additional questions for each specific type were also revised. MOSF took the opposite direction in 2011 by giving up the complex categorization of programs under the old scheme and adopting a simpler scheme consisting of three types, general administration and two types of IT programs with 11 common questions and two questions specific to IT programs.

The evaluation questions consist of three sections – planning, management, and performance and feedback. Each section is assigned 20 points, 30 points, and 50 points, respectively. All questions are listed below in [Figure 4-7]. SABP is an internal evaluation and carried out by the program division or an agency designated to manage the program by the ministry in charge of the program. Left alone, program divisions or agencies may show the tendency to be overly optimistic for their own achievement that the information collected from self-evaluation like SABP may not be useful in figuring out what is going on inside the black box. In order to secure objectivity and accuracy of the evaluation, MOSF published "the Guideline for SABP" every year before a new evaluation cycle begins. The guideline provides not only evaluation questions as in [Figure 4-7], but detailed explanations on how to answer the questions and assign scores to each question. These are illustrated from <Table 4-7> to <Table 4-9>. The total score is obtained by adding all scores from each question. The maximum score is 100. Each program is assigned one of the five letter grades based on the total score. A program will get "very good" grade if the total score obtained in SABP is 90 or higher, "good" if greater than and equal to 80 and less than 90, "fair" if greater than and equal to 70 and less than 80, "unsatisfactory" if greater than and equal to 60 and less than 70, and finally "very unsatisfactory" if lower than 60.

Figure 4-7 | Evaluation Questions in SABP: 2011

- Planning
  1. Adequacy of program plan
     - 1-1. Are the program objectives clear and conformable to accomplishing performance targets? (2 points)
     - 1-2. Is the program unnecessarily similar or overlapping with other programs? (3 points)
     - 1-3. Does the program have adequate design and efficient delivery system? (5 points)

  2. Adequacy of performance plan
     - 2-1. Is there a firm link between performance indicator and program objectives? (5 points)
     - 2-2. Is the target for performance indicator reasonable and concrete? (5 points)
- Management
  3. Adequacy of program management
     - 3-1. Does the program agency do the best to expense the budget as planned? (15 points for general administration, 12 for IT programs)
     - 3-2. Does the program agency operate monitoring system and make efforts to improve it? (10 points for general administration and 5 for IT programs)
     - 3-3. Does the program agency improve efficiency in achieving program objectives? (5 points)
     - 3-IT1 Does the program agency adequately manage information management system? (8 points for IT programs)
     - 3-IT2 Does the program agency make effort establish to fair and competitive market environment? (3 bonus points for IT programs if yes)
- Performance and feedback
  4. Accomplishment of performance objectives and feedback of evaluation results
     - 4-1. Is the target level of performance indicator achieved? (30 points)
     - 4-2. Is the program carried out efficiently based on the evaluation results? (10 points)
     - 4-3. Are the feedbacks from evaluation results and other external opinion incorporated to improve program structure? (10 points)

Table 4-7 | Evaluation Criteria of SABP: Planning

	Points	Answer and Score	Criteria for Yes
1-1	2	Yes (2), No (0)	• Problems to be solved can be clearly defined, • The logical causality between program objectives and accomplishment of performance targets can be clearly established; and • Government expenditure is required.
1-2	3	Yes (3), No (0)	• Program objective is different from others, • Program customers are different even if program objective is similar; or • Problems have been resolved through coordination and cooperation even if program objective is similar.
1-3	5	Yes (5), No (0)	• Sub-programs are appropriate for achievement of program objectives; and • Program design is thought to be the best considering cost sharing, program agency, program customers, demand forecasting, program eligibility, etc.
2-1	5	Yes (5), No (0)	• Performance indicators represent the entire program, • Performance indicators consist mainly of outcome indicators; and • Definition and measurement of performance indicators are clear and reasonable.
2-2	5	Yes (5), No (0)	• Performance targets are accomplished before the scheduled completion date; and • Performance targets are set up reflecting efforts to improve performance.

Table 4-8 | Evaluation Criteria of SABP: Management

	Points	Answer and Score	Criteria for Answers
3-1	15 (12 for IT)	Yes (15, 12), Significantly (10, 8), Reasonably (5, 4), No (0, 0)	Yes • Budget is expensed as planned. Significantly • Budget is expensed as annual plan bit not quarterly plan. • Less than 100% of budget is expensed but spending record is improving significantly. Reasonably • 100% of budget is expensed but timing of spending is seriously misaligned. • Less than 100% of budget is expensed but spending record is improving. No • Budget is used for other purpose without compelling reasons. • Little efforts are observed to improve expense rate. • Expense rate decrease compared to the last year.
3-2	10 (5 for IT)	Yes (10, 5), Reasonably (5, 2.5), No (0, 0)	Yes • Monitoring system on budget expense and program management is well functioning to improve program performance and quality of outputs, • Feedback channel is firmly established; and • Monitoring system is well accepted by all stakeholders. Reasonably • Monitoring system is established but fails to solve all problems. • Monitoring system is established but not complete. No • Targets for budget expense and program management are in place but little effort to improve quality of service is made, • Monitoring system is inadequate to respond to issues raised during carrying out the program, or • Issues raised by NAO and the National Assembly are not resolved yet.
3-3	5	Yes (5), Reasonably (2.5), No (0)	Yes • Costs are saved through improving program management and the results are reflected in the next year's budget, • Budget incentive from MOSF is awarded or decided to be awarded, • Innovative program management is introduced to result in better program performance; or • The program achieves good results from external evaluation. Reasonably • Program efficiency is improved; or • Effort to improve efficiency is made but little improvement is observed. No • Costs are saved but not from explicit effort to achieve them; or • Inaccurate prediction of demand for programs services results in surplus.

Table 4-9 | Evaluation Criteria of SABP: Performance and Feedback

	Points	Answer and Score	Criteria for Yes
4-1	30	Yes (30), Significantly (20), Reasonably (10), No (0)	Yes • Yes to 2-2 and target is fully achieved. Significantly • Yes to 2-2 and target is significantly (90~100%) achieved, • Yes to 2-2 and target is fully achieved but the achievement is due to external factors or program is not carried out as planed or it is difficult to confirm the accomplishment; or • Yes to 2-2 and target is not fully achieved but effort is made to respond to changes in external environment. Reasonably • Yes to 2-2 and target is reasonably (80~90%) achieved, • Yes to 2-2 and target is fully achieved but unreliable data is presented as evidence; or • No to 2-2 but target is significantly (90~100%) achieved. No • No to 2-2 but target is achieved less than 90%, • Yes to 2-2 and target is achieved less than 80%; or • False report or data is presented. * With multiple indicators, the score is calculated as "sum of weights to indicators with YES*0.3 + sum of weights to indicators with SIGNIFICANT*0.2 + sum of weights to indicators with REASONABLE*0.1".
4-2	10	Yes (10), Reasonably (5), No (0)	Yes • External and comprehensive performance evaluation is conducted and effectiveness of the program is demonstrated through evaluation. Reasonably • Internal comprehensive performance evaluation is conducted and effectiveness of the program is demonstrated through evaluation; or • External performance evaluation is conducted but it is difficult to decide effectiveness of the program due to deficiency in evaluation. No • No performance evaluation is conducted; or • Performance evaluation is conducted but it turns out the program is ineffective.

	Points	Answer and Score	Criteria for Yes
4-3	10	Yes (10), Reasonably (5), No (0)	Yes • Modification of program or related procedures is done to solve the problems raised by internal and external evaluations, • Recommendations from MOSF through SABP or program evaluation are incorporated into the program and improvement in program management is demonstrated; or • Issues raised by NAO or the National Assembly are resolved and improvement in program management is demonstrated. Reasonably • Only part of the issues raised in "Yes" part is solved to improve the program management. No • No plan is established to improve program management by incorporating the results of performance evaluations; or • Little is demonstrated to show improvement by incorporating recommendations from MOSF through SABP or program evaluation and issues raised by NAO or the National Assembly.

Table 4-10 | SABP Grades

Grade	Total Score
Very Unsatisfactory	Less than 60
Unsatisfactory	Greater than and equal to 60 and less than 70
Fair	Greater than and equal to 70 and less than 80
Good	Greater than and equal to 80 and less than 90
Very Good	Higher than 90

The final results of SABP are explicitly embedded in the budgeting process. Programs with "Unsatisfactory" or "Very Unsatisfactory" grade would be penalized through budget cut amounting to 10% of the previous year's budget. However, the budget cut is not an automatic step since other elements such as fiscal conditions, importance of the program, future prospects, and political considerations should also be factored into the final decision. Incentives are offered to the programs demonstrating good performance. Programs with no worse grade than "Good" would be the primary candidates for budget increase. In addition, all the final reports of SABP are posted on MOSF's website and made available to the public.

The final report typically contains recommendations for better program management, which the government offices are supposed to incorporate into program management.

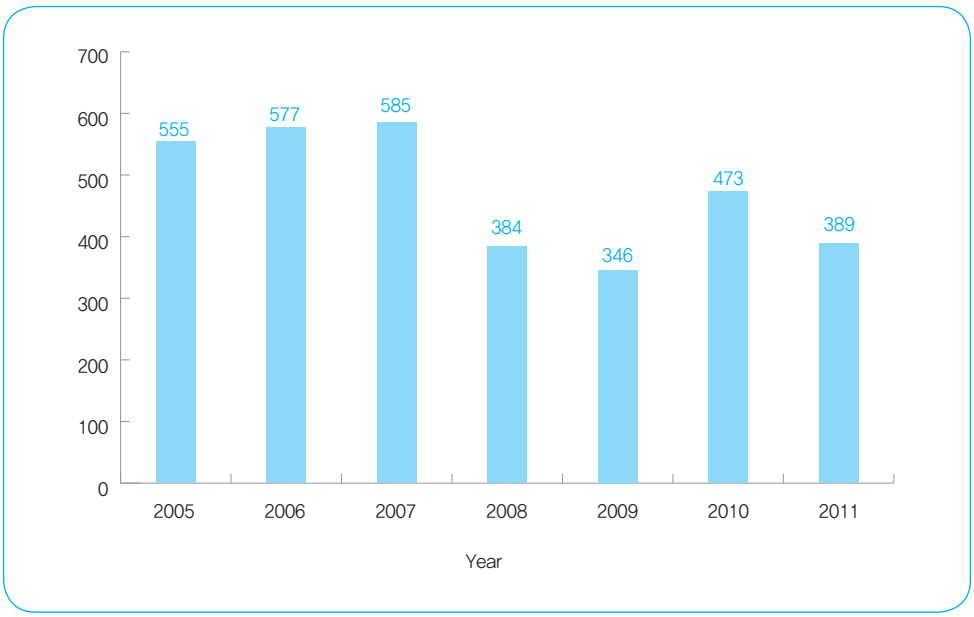
MOSF and the government offices check the implementation of the recommendations during the next evaluation cycle. Unsatisfactory implementation would result in an undesirable consequence such as penalty in next year's SABP to the government offices in charge of the program.

3.4. Assessment on the Performance of SABP

Park and Won (2012) analyzed the results of SABP from 2005 to 2011 and identified the factors that affected the performance of budgetary expenditure programs. [Figure 4-2] summarizes the number of budgetary expenditure programs evaluated by SABP each year. Up until 2007, SABP had been conducted for about 600 programs. The number was significantly reduced to 384 in 2008[30]. In 2010, MOSF asked each central government office to select a third of all programs for SABP not randomly but deliberately based on performance goals so that the link between individual programs and performance goals is strengthened. The sectorial distribution of the programs is reported in <Table 4-11> where we classify all programs into seven groups; SOC investment, equipment and facility, direct service provision, equity investment, loans, grants to private sector, and grants to local government. The central government offices themselves manage the programs in the first three groups; SOC investment, equipment and facility, and direct service provision, while they delegate to outside agencies such as public corporations, local governments, and non-government organizations (NGOs), programs in the other four groups; loans, grants to private sector, and grants to local government. One noticeable feature found in <Table 4-11> is that SOC investment programs drew less attention in SABP while more emphasis is put on programs of direct service provision. In 2005 and 2006, almost 10% of all programs evaluated under SABP were those in SOC investment group but the proportion of SOC investment programs was reduced to 0.8% in 2010 (4 out of 473) and 5.1% in 2011 (20 out of 389). On the other hand, the proportion of programs for direct service provision significantly increased from 25.9% in 2005 (144 out of 555) to 32.6% in 2011 (127 out of 389). It is also true that MOSF and government offices pay more attention to programs that provide grants to private agencies.

30. IT programs were excluded from the analysis by Park and Won (2012). The following discussion focuses only on programs classified as general administration in SABP framework by MOSF.

Figure 4-8 | The Number of Programs under SABP



Source: Park and Won (2012)

Table 4-11 | Distribution of Programs under SABP

(Unit: programs)

Year	SOC	Equipment & Facility	Direct Service Provision	Equity Investment	Loans	Grants to Local Gov't	Grants to Private	Sum
2005	54	8	144	49	68	109	123	555
2006	51	9	170	46	43	101	157	577
2007	22	6	187	69	42	103	156	585
2008	15	5	83	40	34	83	124	384
2009	7	6	112	16	23	64	118	346
2010	4	4	162	39	35	90	139	473
2011	20	0	127	25	24	53	140	389
Sum	173	38	985	284	269	603	957	3,309

Source: Park and Won (2012)

[Figure 4-9] illustrates the average score recorded in SABP from 2005 to 2011. The average score decreased slightly to 59.9 in 2006 from 60.1 in 2005, and then showed a significant improvement from 2007; 66.0 in 2007, 66.6 in 2008, and 65.9 in 2009. The average score dropped considerably in 2010 and 2011. Based on [Figure 4-10] we can conclude that the main driving force behind the fluctuations in average score is the changes in scores from performance questions which carry the largest weight of 50%. Note that both the average score and achievement in performance questions were relatively high between 2007 and 2009.

Figure 4-9 | Average Scores in SABP

Source: Park and Won (2012)

Figure 4-10 | Average Scores in SABP by Section

Source: Park and Won (2012)

Evaluation results can also be examined in terms of the distribution of the letter grade based on the total score achieved in SABP. Letter grades are assigned to each program following the scale explained in <Table 4-4>. We can point out two noticeable trends from the results in <Table 4-12>. One is that the ministries showed lukewarm responses to the request of assessing their own performance, which is typically observed in the self-evaluation like SABP. The overwhelming majority of the programs were graded as "Fair" consistently throughout the sample period. Second, the evaluation seems to have become stricter especially after 2008 when 26.8% of the programs evaluated in SABP were graded as "Unsatisfactory" or worse while in 2007 only 5.3% of the programs were assessed to have the same letter grades. It is not clear why, all of a sudden, ministries became so conservative in assessing their own performance. The conjecture is that based on the experiences for three years from 2005 to 2007, MOSF may have concluded that the ministries were too generous in evaluating their own performance and instructed the ministries that too generous of an assessment would not be acceptable anymore.

Table 4-12 | Distribution of Letter Grades in SABP

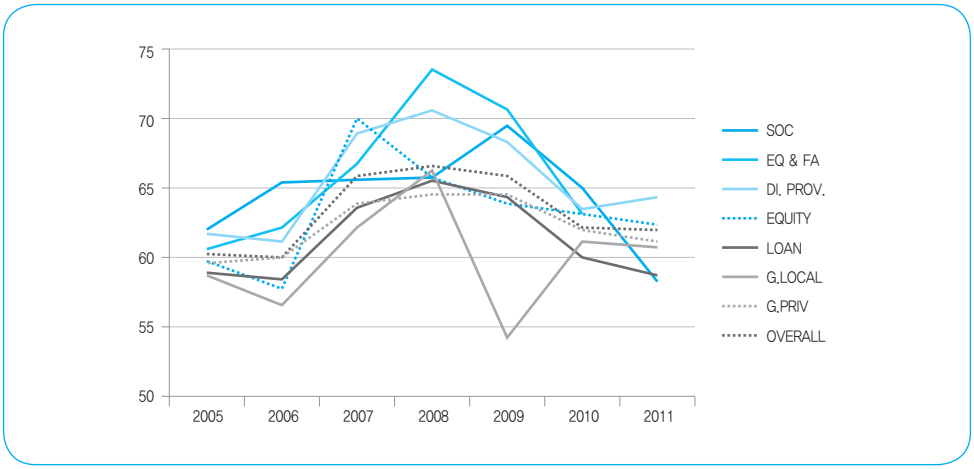
(Unit: programs, %)

	Very Unsatisfactory	Unsatisfactory	Fair	Good	Very Good	Total
2005	0 (0.0)	87 (15.7)	337 (60.7)	102 (18.4)	29 (5.2)	555
2006	0 (0.0)	65 (11.3)	338 (58.6)	94 (16.3)	30 (5.2)	577
2007	0 (0.0)	31 (5.3)	342 (58.5)	143 (24.4)	69 (11.8)	585
2008	0 (0.0)	103 (26.8)	226 (58.9)	44 (11.5)	11 (2.9)	384
2009	1 (0.3)	70 (20.2)	257 (74.3)	14 (4.0)	4 (1.2)	346
2010	30 (6.3)	86 (18.2)	335 (70.8)	22 (4.7)	0 (0.0)	473
2011	36 (9.3)	82 (21.1)	245 (63.0)	25 (6.4)	1 (0.3)	389

Note: The numbers in the parentheses are the percentage of the programs with the corresponding letter grade among the year's total programs in SABP

Source: Park and Won (2012)

Figure 4-11 | Average Scores in SABP by Program Group



Source: Park and Won (2012)

[Figure 4-11] disaggregates the results of SABP by program groups. It is worth pointing out that the performance of programs in SOC investment group had worsened most considerably and those in grants to local governments and private sector groups had shown relatively poor performance consistently. One more thing to note is that the performance of the programs in the direct service provision group has improved by a significant degree.

Park and Won (2012) investigated the factors that affect the performance of programs in SABP by multivariate analysis and offered three interesting findings. First, the size of program budget has a positive effect on performance assessed in SABP. The bigger the program is, the higher the score in SABP is. Second, programs carried out by the central government showed far better performance than those delegated to the local governments or external agencies. Third, programs with explicit customers performed worse than those with unspecified customers. Most welfare programs have well specified beneficiaries and their performance is worse than most of the programs belonging to the groups of equipment & facility or SCO investment whose customers are defined very vaguely like the general public.

Despite various difficulties encountered earlier, SABP has firmly established itself as an indispensible component in the performance management system in Korea. Problems were predominantly of a technical nature but political and cultural problems were also observed, especially when it came to utilizing performance information to improve the system. The limited technical capacity of the central government offices impeded efforts to produce essential performance information. Progress has been made through training

programs but deficiencies in data availability still seriously hinder the central government offices from providing performance information. It appears that a considerable amount of time and resource would be needed to overcome the difficulties completely and build a strong system with well-functioning information production mechanism. In addition, several policy-oriented central government offices were likewise struggling to identify concrete measures linking their efforts to outcomes. The technical difficulties became more challenging due to a distinctive human resource management system in which government employees, especially higher ranked ones, are placed on different posts and assigned to different tasks on a regular basis. This practice allows government employees to accumulate general knowledge and skills and also helps reduce the possibility of corruption. However, it has a critical deficiency that may hinder them from accumulating expertise in specific area like performance management. Cultural problem also posed a daunting challenge in the earlier stages. Government employees in Korea were not so accustomed to evaluation, not to mention self-evaluation like SABP. Naturally, they were very resistant to the new evaluation system that even linked the results of the evaluation to the amount of resources they can handle. However, it appears that continuous communication between MOSF and the program offices made government employees accept SABP as a normal part of the performance management system of budgetary programs. The strong support for the result-oriented performance management system from the public as well as the National Assembly also contributed to persuading all relevant stakeholders, including government employees.

It is too early to cast a definite verdict on SABP but important lessons can be learned from the Korean experiences, especially several mistakes committed in earlier stages. First, SABP contributed to calling attention to the importance of result oriented performance management by providing quantified and objective information on program outcomes in addition to the traditional information on inputs and activities. In fact, SABP was the first *bona fide* evaluation scheme that provided objective information on program outcomes in Korea. [Figure 4-10] clearly shows that in assessing their own achievement, the central government offices were more favorable to planning and management than to performance. This might be reminiscent of the traditional practices focusing on inputs and processes rather than outputs or outcomes. The traditional system was built on the presumption that controlling the flow of inputs or processes leads us to achieve the goals or purpose of government expenditure programs. One lesson we can learn from SABP is that it may not be necessarily a valid conjecture. The information obtained from SABP like [Figure 4-10] can be used to convince government employees that more effort should be made to construct a result oriented performance management system.

Second, to be effective, the results of SABP should be closely linked to the amount of budgetary resources each central government office can control. By doing so, the incentives

of the government offices and their employees are made to be compatible with the public interest so that they voluntarily make an effort to improve program performance. However, early experience from SABP was not satisfactory in that the principle relating performance in SABP with the budget was not strictly enforced. Since the introduction of SABP in 2005, MOSF has maintained the position that the results of SABP would be incorporated into budgetary decision makings. Aside from minor variations, MOSF has retained the principle that the next year's budget of the programs that attained the letter grade of unsatisfactory or very unsatisfactory in SABP should be cut by no less than 10%. However, the guideline for SABP also makes it clear that budgetary decision makings would take into account various factors such as fiscal condition and political consideration as well as results of SABP. The position is understandable in the sense that there are many unpredictable and unquantifiable elements that should be factored. Therefore, it is unavoidable to maintain a flexible position like the one MOSF currently maintains. However, it is also possible that the flexibility may weaken the incentive for the government offices to put more effort to achieve better performance. <Table 4-13> illustrates the joint distribution of SABP letter grades in 2008 and changes in budget size between 2008 and 2009. It is not clear whether there exists a discernible correlation between SABP grades and changes in budget size. But we can obviously confirm the tendency that a better grade in SABP is more likely to be associated with increases in the next year's budget. Pointing out the fact that 25.9% of the programs with "Unsatisfactory" grade experienced an increase rather than decrease in the next year's budget while 50% of the programs with "Very Good" grade experienced a decrease in the next year's budget, Park (2009) argues that a stronger tie between SABP and budget would be necessary to provide enough incentive to improve performance.

Table 4-13 | SABP Results and Budget Change

(Unit: programs, %)

SABP Grade in 2008	Budget Change between 2008 and 2009		
	Increase	No Change	Decrease
Very Good	4 (50.0)	0 (0.0)	4 (50)
Good	27 (65.9)	1 (2.4)	13 (31.7)
Fair	142 (67.6)	10 (4.8)	58 (27.6)
Unsatisfactory	22 (25.9)	3 (3.5)	60 (70.6)
Very Unsatisfactory	1 (7.7)	0 (0.0)	12 (92.3)

Source: Park (2009)

Third, there exists a significant gap between the preliminary results assessed by the program offices and the final results certified by MOSF. Under SABP, each central government office conducts self-assessment on the programs selected for evaluation in consultation with MOSF and sends the results to MOSF for confirmation. MOSF adjusts the evaluation results, if necessary, and finalizes them. In general, central government offices managing programs have a tendency to overestimate their achievement while MOSF tries to be as conservative as possible in evaluating the validity of self-assessments. Therefore, we can expect disagreements on performance of budgetary expenditure programs. This turns out to be indeed the case. For instance, out of 69 programs graded as "Very Good" by the program offices in SABP of 2008, 59 were assigned the final grade "Unsatisfactory" and 10 "Very Unsatisfactory". Moreover, the differences in opinions between the program offices and MOSF have not been reduced as more experience is accumulated. According to Lee (2010), the average difference between preliminary assessment of the program offices and the final confirmations by MOSF was 25.6 in 2005, 26.8 in 2006, and 24.6 in 2007, respectively. The majority of the gap in assessments stems from different opinions on the performance rather than planning or management of the programs. Lee (2010) reports that the different assessments on performance account for 58.5% of the average score gap. One may raise the question that the large explanatory power of scores from performance questions may be largely attributable to large weights assigned. However, even after controlling for the different weights, the differences in opinions on performance questions in SABP turn out to explain 33.8% of the total differences.

Table 4-14 | Different Assessments between the Program Offices and MOSF

(Unit: points, %)

Ministry	Total	Planning	MGT	Performance
STRATEGY & FINANCE	25.9	7.5	6.8	11.6
UNIFICATION	37.8	8.5	6.0	23.3
FOREIGN AFFAIRS	29.5	6.8	3.0	19.7
DEFENSE	20.1	3.2	7.3	9.6
JUSTICE	25.9	6.4	4.0	15.4
PUBLIC ADM. & SECURITY	34.5	9.4	6.0	19.1
EDUCATION	27.9	7.5	4.4	16.0
CLUTURE	28.5	8.5	4.8	15.3
SCI. & TECHNOLOGY	41.8	13.1	4.5	24.2
AGRI. & FORESTRY	21.1	7.6	3.0	11.5
INFO. & COMM.	26.7	4.0	5.4	17.3

Ministry	Total	Planning	MGT	Performance
INDUSTRY & RESOURCE	28.6	7.4	2.1	21.1
WELFARE	29.8	7.3	3.8	18.7
ENVORONMENT	26.6	5.1	5.3	16.2
CONST. & TRANS.	27.4	5.6	7.3	14.5
MARITIME & FISHERY	21.9	5.0	2.9	14.0
GENDER EQUALITY	37.5	11.7	5.0	19.9
Average	28.9	7.3 (24.4)	4.8 (24.0)	16.9 (33.8)

Note: 1) The table is based on the results of SABP in 2008

2) The numbers in parentheses are the differences in averages scores in terms of percentage

Source: Lee (2010)

Fourth, the National Assembly pays little attention to the results of SABP during the budget process, which may cause undesirable side effects of weakening incentive for improving program performance. Park and Park (2008) reports that for FY 2008 the National Assembly allocated more budget than the government requested to two programs that were assigned the grade "Unsatisfactory" by SABP in 2007. In addition, the budgets of FY 2008 for 11 programs with the grade "Very Good" were reduced compared to the budgets requested by the government while only 6 programs with the same grade were allocated more budgets than those requested by the government. SABP is the internal evaluation scheme introduced by the National Finance Act with a view to improving the efficiency of the administration. Since the National Assembly is not bound by the results of SABP in any sense, it does not have to take them into consideration when carrying out the budget process. On the other hand, the notion that both the administration and the National Assembly should take seriously the results of performance evaluations, including SABP, as important sources of information to improve the effectiveness of the budget process, has gained widespread support from various stakeholders. Therefore, a large body of literature on the performance evaluation recommends that the National Assembly and the administration work together to establish an explicit mechanism through which the information from SABP can be incorporated into the budget process.

Table 4-15 | Differences between Ministry and MOSF Evaluation: SABP in 2008

(Unit: programs, %)

SABP in 2007	Adjustment by the National Assembly for the Budget of FY 2008			
	Decrease	No Change	Increase	Total
Unsatisfactory	5 (17.2)	22 (75.9)	2 (6.9)	29
Fair	46 (13.7)	249 (74.3)	40 (11.9)	335
Good	16 (12.0)	96 (72.2)	21 (15.8)	133
Very Good	11 (17.2)	47 (83.4)	6 (9.4)	64
Total	78 (13.9)	414 (73.8)	69 (12.3)	561

Note: Four kinds of letter grades were given in SABP up to 2008

Source: Park and Park (2008)

4. In-Depth Evaluation of Budgetary Programs (IEBP)

4.1. Introduction

The in-depth evaluation of budgetary programs (IEBP) is a form of program evaluation discussed in Chapter 2. According to the OECD (1999), program evaluation is "a systematic and analytical assessment addressing important aspects of a program such as effectiveness". That is, program evaluation, first of all, investigates how much the program contributes to accomplishing the intended results in a scientific manner and then describes the factors of success or failure objectively to be used to improve policy design. Program evaluation provides information that can be utilized in decision making in the future by explaining how and why the program outcomes are obtained. In addition, program evaluation offers reliable assessment on the adequacy and appropriateness of program goals and implementation as well as performance indicators. However, due to its in-depth nature, it is impractical to apply complicated procedures and techniques of program evaluation to all budgetary expenditure programs. The usual practice is to carry out IEBP for only a small number of selected programs that merit scrutiny.

Table 4-16 | IEBP and Enforcement Decree of the National Finance Act

	Contents
National Finance Act Enforcement Decree	Article 3 The Minister of MOSF can ask the head of each central government office and manager of each public fund ~ to carry out in-depth evaluation of budgetary programs when one of the following conditions is satisfied. IEBP on R&D programs can be replaced with the evaluations according to the Act on Performance Management and Performance Evaluation of National R&D programs. 1. Further evaluation is needed considering the results of SABP. 2. Duplicate programs in different program offices or inefficient program implementation may lead to waste of the budget. 3. It is desirable to improve the efficiency of government expenditure through objective examination since continuous and significant increase in budget is expected. 4. The Minister of MOSF decides that it is necessary to carefully examine the program through IEBP.

IEBP was first introduced into the performance management system in 2006 based on the National Finance Act and its enforcement decree. IEBP shares ⑥, ⑦ and ⑧ of Article 8 of the National Finance Act with SABP as the legal foundation as shown in <Table 4-16>. The Article 3 of the Enforcement Decree of the National Finance Act specifies IEBP as a tool for performance evaluation along with SABP and stipulates four circumstances that may warrant IEBP initiated by MOSF.

4.2. The Procedures of IEBP

IEBP starts with the selection of programs to be evaluated. In consultation with the central government offices and Korea Development Institute (KDI), IEBP Committee at MOSF selects a small number of programs, typically about 10 programs a year, for which an intensive evaluation of IEBP is conducted. The programs that MOSF selects as the targets of IEBP should satisfy one of the following criteria;

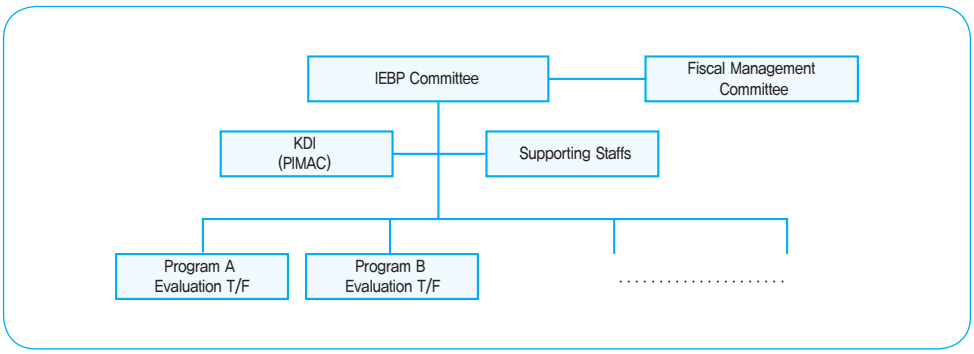
- Further evaluation is needed considering the results of SABP.
- Duplicate programs in different program offices or inefficient program implementation may lead to waste of the budget.
- It is desirable to improve the efficiency of government expenditure through objective examination since continuous and significant increase in budget is expected.
- The Minister of MOSF decides that it is necessary to scrutinize the program though IEBP.

In choosing the programs, IEBP Committee also takes into considerations various aspects of the programs such as budget size, importance of the program in terms of policy priority, and the results of evaluation by other authorities, especially the National Assembly and the National Audit Office. The preliminary list of the IEBP projects is reviewed by the Advisory Board for Evaluation of Budgetary Programs consisting of heads of divisions that possess interest in performance evaluation in MOSF and external experts in performance evaluation. Once confirmed by the Advisory Board, the list is sent to the Fiscal Management Committee (FMC) for finalization of the selection process. FMC consists of the vice ministers of MOSF and other ministries in the administration.

The next step is to establish the Evaluation Task Force for each program selected for IEBP. Each Task Force consists of an evaluation team, representatives from the government offices carrying out the program, and officials from MOSF's Fiscal Institutions Division in charge of the operation of IEBP. MOSF delegates to KDI the authority to support the administrative works for IEBP and to recruit the evaluation teams. More specifically, the Public & Private Infrastructure Investment Management Center (PIMAC) at KDI serves as the agency of MOSF in IEBP implementation and provides various services including support and supervision of the evaluation task force, selection of the evaluation team, and general research on program evaluation. Recruitment of evaluation teams can be carried out through a public tender or direct solicitation by KDI. The evaluation team carries out the IEBP project and prepares the evaluation report. The team typically consists of one project manager, one to two researchers, and several research assistants. It is recommended to form an evaluation team as small as possible, preferably no more than three research members, to prevent communication glitches among team members. The project manager and researchers should possess adequate expertise in the evaluation works and preferably higher degrees in the related fields. Government officials participating in the task force play an important role in supporting data gathered by the evaluation team and facilitating the communication between the evaluation team and the program offices. The task force is co-chaired by the project manager and the chief of Fiscal Institutions Division in MOSF. As the chairman of IEBP Committee, the Director of the Fiscal Management Bureau has the overall responsibility of conducting a smooth and efficient evaluation.

To check the progress of the evaluation project, three working group meetings are held. In the first meeting, the evaluation team and KDI staffs discuss the evaluation schedule and the important issues that should be addressed. The other two meetings are preparatory sessions for the mid-term and the final reports. At those meetings, the evaluation team reports the progress of the evaluation project to KDI staffs. The evaluation team also reports the results of the evaluation to two conference meetings, mid-term and final, where all members of the task force participate.

Figure 4-12 | Structure of IEBP Committee



Completing the evaluation, the evaluation team should prepare the final report. It should be written in a clear and logical way. To provide a guide, KDI suggests an exemplary "table of contents" evaluation teams can take as a benchmark, which is shown in <Table 4-17>. The evaluation team can modify the format of the final report suggested in the table after consultation with KDI, if necessary. Plain words and short sentences are preferred in writing the report. Moreover, the report should be self-contained so that ordinary people with an average intelligence and general knowledge should be able to understand it. To do so, it is necessary to provide detailed explanations on the program structure and related information and the use of professional jargons should be avoided as much as possible. It is a good idea to include the glossary at the end of the report to help the readers understand.

The most important part in the final report is that the evaluation results and policy recommendations should be separated. By clearly distinguishing the objective analysis from subjective suggestions, we can prevent program offices or related stakeholders from denying the evaluation results even though they do not agree on the policy recommendations the evaluation team made. The final version of the report is confirmed by the Evaluation Task Force and submitted to the Fiscal Management Committee shown in [Figure 4-12]. MOSF later uses the information from IEBP, especially evaluation results and policy recommendations, for various budgetary purposes. The information may be utilized in setting the ceiling on expenditures of government offices, making the National Fiscal Management Plan, and compiling the annual budget.

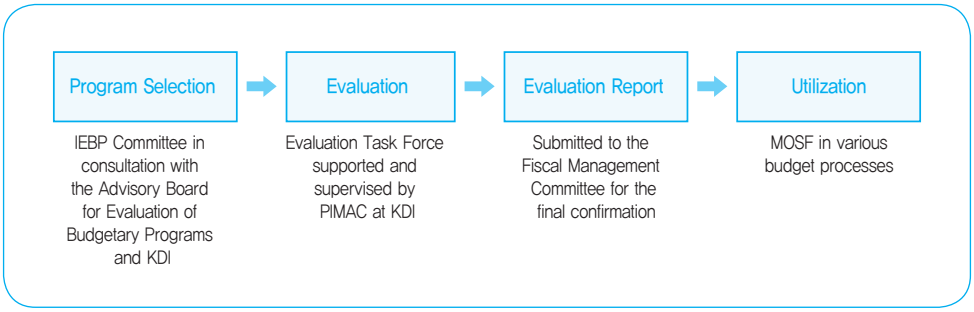
The flow of tasks under IEBP discussed above is summarized in [Figure 4-13].

Table 4-17 | The Structure of the Final Report of IEBP

Content	Note
Summary	
Introduction	• Brief program description, reason for evaluation, important issues addressed
Program Features	• Brief program description: program goals, background, history, and legal foundation • Delivery system: program agency, customers, stakeholders • Program budget: budget by year and sub-program, rate of budget run-out, medium-term investment plan under the National Fiscal Management Plan • Program performance: performance objectives and indicators in the performance plan, results of SABP and other evaluations • Similar programs
Important Issues and Evaluation Method	• Program objectives and intervention logic • Important issues that should be addressed in IEBP • Objective and scope of the evaluation
Experiences of Other Countries	• Explanation and comments on similar programs in other countries
Relevance	• Relevance of government intervention • Relevance of the division of labor between the central and local governments • Relevance of intervention methods
Effectiveness	• Performance indicators and reference for comparison • Evaluation model • Results of data analysis
Efficiency, Utility, and Sustainability	• Efficiency evaluation • Utility and sustainability evaluation
Summary & Policy Recommendation	• Summary of evaluation results and policy recommendation for improving program performance
References	
Appendix	• Detailed explanation on data and glossary

Source: Koh and Kim (2007)

Figure 4-13 | The Procedure of IEBP



Korea Development Institute (KDI) and Korea Public Finance Institute (KIPF)

KDI was founded by the Korean government in 1971 as a think-tank that carried out rigorous research on economic policy issues and assisted the government in formulating the long-term economic development strategy, known as the "Five-year Economic Development Plans". Researchers at KDI also conducted short-term research projects to evaluate current economic policy issues. Ever since its establishment in 1971, KDI has continuously carried out extensive policy-oriented research on almost all issues on Korean economy including macroeconomics, public finance, monetary and financial economics, industrial organization, labor market, social welfare and international trade. KDI has established itself as the leading economic research institute in Korea. Currently, KDI has three research departments, the Department of Macroeconomic and Financial Policy, the Department of Industry and Competition Policy, and the Department of Public Finance and Social Policy. Three research centers, the Economic Information & Education Center, Public & Private Infrastructure Investment Center (PIMAC), and Center for International Development and a graduate school, the KDI School of Public Policy and Management, are also affiliated to KDI. PIMAC was established as an affiliate of KDI through the merger of the Public Investment Management Center (PIMA) at KDI and the Private Infrastructure Investment Center of Korea (PICKO) at Korea Research Institute for Human Settlement (KRIHS) with the amendment of the Act on Public-Private Partnerships in Infrastructure in January 2005.Main functions of PIMAC include the execution of Preliminary Feasibility Study and Reassessment Study of Feasibility on large-scale publicly-financed projects, for which comprehensive research is conducted on the basis of economic and policy analysis. As a think-tank, PIMAC produces various reports and policy recommendations on improving the public investment system in Korea. On the other hand, PIMAC supports the government in developing policies and plans on Public-Private Partnership (PPP) and in implementing PPP projects.

PIMAC conducts Value for Money Tests and lends assistance in the designation of concessionaires. This is done through support in formulating Request for Proposals, evaluation of project proposals, and negotiation with potential concessionaires. PIMAC is also in charge of capacity building of public officials and provides managerial services for the PPP database. PIMAC, at the same time, develops guidelines for and implements the Preliminary Feasibility Study for projects procured from public institutions. It is also engaged in ex-post in-depth evaluation of government programs. In 2005, the Korean government designated PIMA, the predecessor of PIMAC, as the agency of MOSF in IEBP implementation. PIMAC provides administrative supports to IEBP evaluation projects and conducts research on program evaluations.

The Korea Institute of Public Finance (KIPF) was established in July 1992 for the purpose of policy-oriented research and analysis in all aspects of taxation and public finance, assisting the government in formulating national tax policies thus consequently contributing to the nation's economy. Since its foundation, KIPF has played an important role in the development of tax and budget policies and the improvement of tax administration. KIPF currently has three research centers – Research Center for Taxation, Research Center for Government Expenditure, and the Research Center for Public Institutions. In 2004, the Center for Performance Evaluation and Management (CPEM) was established under the umbrella of the Research Center for Government Expenditure to contribute to managing and evaluating the performance of government budgetary programs. CPEM performs research and education about the systems and techniques needed to manage the performance of government budget projects, and assessment of budget projects. In 2004, the Korean government designated CPME as the agency to support the administration of the Performance Goal Management System and the Self-Assessment of Budgetary Programs.

4.3. The Core Structure of IEBP: Five Evaluation Criteria

IEBP requires every evaluation team to assess program performance from five different perspectives – relevance, effectiveness, efficiency, utility, and sustainability. All evaluation reports are required to include the assessment from the first two perspectives and can decide whether or not to carry out assessments from the other three perspectives. That is, evaluation of a program in terms of efficiency, utility, and sustainability is optional.

Relevance refers to the justifiability or necessity of the government intervention in the market mechanism. Examination on the relevance of a program can be carried out by asking a series of questions. The evaluator starts with the question whether the program belongs to the roles of the government. If the answer is yes, then the next question would be whether the central government or the local government should take charge of the program. These

questions naturally lead to another question on whether the program should be allowed to continue in its current state, altered significantly, or merely allowed to elapse without being renewed. In addition, the evaluator may raise an issue concerning the relevance of the program in the future. The discussion of future relevance typically entails an examination of alternatives to the current program. Several reasons have been cited as the theoretical background for government intervention. Inefficient resource allocation due to market failure, unequal income distribution, and macroeconomic fluctuation are the most popular examples.

The next issue that should be addressed in IEBP is effectiveness of the program. In many cases, too much attention is paid to efficiency of the program. However, the problem is that efficient implementation of a program does not necessarily guarantee the resolution of the social or economic problem the program is supposed to deal with. Efficiency is one thing and effectiveness is another. A bad program design is more likely to result in ineffective use of government resources. It is especially true when program objectives were defined without clarity. If it turns out that the objectives of the program is not defined clearly, the evaluator should modify them and continue the evaluation project. In evaluating the effectiveness of a program, the evaluators should try their best to make a balanced approach by taking into consideration negative and unexpected impacts as well as positive and expected ones. Examining the effectiveness of the program constitutes the core task of IEBP. In other words, it is most important for the evaluators to check the overall effectiveness of a program and identify the causal relationship between the program and the outcomes.

When it comes to efficiency, the evaluator asks whether the same amount of output can be obtained from less inputs or more outputs from the same amount of inputs. To answer the questions, the evaluator should compare the current program with various alternatives. Therefore, it is very important to find appropriate alternatives in evaluating efficiency of a program. Examining alternative program designs, the evaluator should set the benchmark against which the performance of the program under evaluation is assessed. Difficulties may arise when there are no comparable programs to use as benchmarks or the evaluator has no previous experience with similar programs. Assessment of program performance in terms of efficiency provides very useful information that can be used in searching for a service delivery system with better performance.

Utility is assessed by comparing the outcome of a program with the level of needs the program customers originally expressed. Programs are said to have utility if they manage to bring about socially beneficial changes by satisfying needs of the target population. There exist intricate differences between effectiveness or efficiency and utility. Even if a program turns out to be effective or efficient, it may have failed to bring the expected level of utility to the program customers. For instance, a very effective job training program for the long-

term unemployed may not be so useful if only a small portion of the potential program customers actually participated in the training program. The evaluator should be able to find out why the participation rate is so low so that the program agency would reshuffle the program to induce more unemployed people to participate.

Sustainability requires the effect or utility of a program to last for a reasonable period after the completion of the program. Even if a program generates benefits, it may be of little value unless the benefits can be enjoyed for a very short period of time.

4.4. Assessment on Performance of IEBP

Since its introduction in 2006, IEBP had been completed for 55 programs by the end of 2011. <Table 4-18> reports a partial list of programs evaluated by IEBP. One noticeable point from the table is that the number of programs evaluated through IEBP decreased to 4 in 2010 from 10 in 2009. That is mainly because MOSF changed the basic unit of IEBP from individual programs to program group in 2010. A program group is the collection of individual programs that share the common objectives. Since a program group is defined in terms of program objectives rather than administrative purpose, a typical program group consists of multiple programs possibly belonging to different government offices. MOSF changed the unit of evaluation for several reasons; identification and integrated management of similar or duplicate programs across different ministries or program agencies, prioritization in resource allocation among competing programs, and establishment of a reasonable scheme of division of labor across ministries or program agencies.

Table 4-18 | Examples of Programs under IEBP

Year	IEBP Programs	Examples of Programs
2006	11	Supports for the unemployed youth, Promotion of cultural contents, Promotion of overseas employment, Workfare for the poor, Construction of national fishery harbor
2007	9	Fishermen insurance, Job training for the unemployed, Promotion of traditional marketplace, Promotion of innovative ability of the universities outside the Seoul metropolitan area
2008	12	Support for environment friendly agricultural infrastructure, Development of overseas natural resources, Support for SME innovation, Enlargement of farmland size
2009	10	Natural disaster insurance, Support for construction of housing projects, Early re-employment benefit, Support for pre-natal, post-natal, and maternity leave

Year	IEBP Programs	Examples of Programs
2010	4	Promotion of in-bound FDI, Subsidy to investment of start-up businesses, Employment subsidy to employers
2011	9	SME support, Support for multi-cultural families, Support for road safety, Promotion of social services, Support for energy efficiency enhancement

The results of IEBP are incorporated into program management as well as the budget process. Majority of the programs that showed unsatisfactory performance in IEBP underwent significant modifications in expenditure structure. For instance, the subsidy program to support investment of start-up businesses was terminated based on the ineffectiveness of the program identified by IEBP carried out in 2009. The Ministry of Labor used to operate an expenditure program that paid subsidies to employers who would retain workers irrespective of a worsening market environment. The program was selected as a subject of IEBP in 2010. Based on the evaluation results, the evaluator recommended that the size of subsidy should be reduced to the level before the global financial crisis of 2007/08 considering the fact that the Korean economy had significantly recovered from the damage done by the crisis. The recommendation was accepted by MOSF and the Ministry of Labor. The total size of the subsidy paid by the program was reduced to 27 billion Korean won in 2011 from 446 billion Korean won in 2009.

Programs are also re-designed or modified to incorporate the results of IEBP. Modification of the program focuses on improving program performance through accountability and efficiency. The Ministry of Small and Medium Enterprise operates a program that makes equity investment in the Korea Venture Capital Investment Fund, which is a fund of funds that provides capital to venture capital funds. The program was selected as a subject of IEBP in 2008 and the evaluation team suggested that the program should be modified in several ways to improve the program performance. In response to the recommendation of the evaluation team and the request from MOSF, the Ministry of SME modified the program structure to induce more private or foreign investment and to strengthen the cooperation among the government offices related to investment on venture capital funds. At the same time, further investment in the fund of funds by the program was restrained until it became obvious that the unsatisfied demand for venture capital investment increased to justify further injection of capital by the public sector.

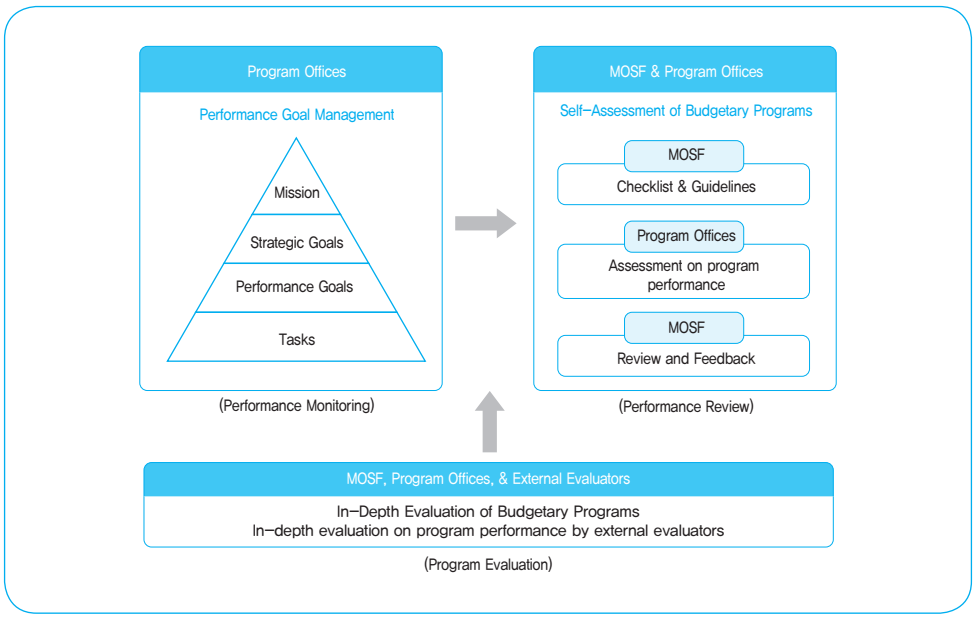
It is hard to find a systematic assessment of the performance of IEBP. Yet, reflecting on the experiences for almost a decade, scholars and practitioners in the area offer comments on some aspects of IEBP. First, the limitation of available data seriously hampers rigorous evaluation on program performance in many evaluation projects. Construction of the control group is particularly a daunting task since no program was designed with due regard

to program evaluation. No randomly assigned control group is available that an alternative is to construct a control group under quasi-experimental environment. It is however a very difficult task since few program offices track down people who do not participate in the program but possess similar characteristics to program participants. Few expect the problem to be resolved in the near future. In the meantime, evaluators have little choice but to rely on available data sources that were constructed for other purposes. Second, central government offices in charge of the programs under evaluation have shown a passive attitude toward IEBP. Successful evaluation is next to impossible without the cooperation of program offices since they are the primary source of information on program performance. We can understand the reluctance of government program offices on IEBP in the context that the way the subjects of IEBP are selected already implies the possibility of unfavorable results to the program as well as the program offices. According to the enforcement decree of the National Finance Act, MOSF selects the subjects of IEBP programs with some issues such as unsatisfactory performance results in SABP, suspicion of wasteful program implementation, and the need for objective examination on program performance. Therefore, it is more likely that the evaluator casts a negative verdict on the program performance, which again may lead to an adverse atmosphere in the competition for more budgetary resources. Acknowledging the possibility of a passive attitude from the program offices, MOSF invites them to express their opinion by participating in an evaluation task force, as well as the Fiscal Management Committee. However, it does not seem that the efforts are successful enough to induce a proactive attitude from the program offices. Further measures are needed to persuade program offices to cooperate with MOSF and the evaluators during the evaluation process. Third, some argue that MOSF took too ambitious measures when it switched the basic unit of IEBP from individual programs to program groups. The primary advantage of the switch is that we can obtain information on the relative effectiveness or importance of individual programs pursuing the same or similar performance goal. The information may play a pivotal role in prioritizing competing programs and allocating budgetary resources among them according to the priority. However, limited time and financial resources allowed for a typical ISBP project makes it very difficult to produce accurate information on the performance of many programs in a program group. We cannot exclude the possibility that the evaluation of a program group rather than individual programs may result in inaccurate information due to limited resources or inadequate expertise of the evaluators. Some experts strongly argue that considering the current state of expertise and resource allowed for program evaluation in Korea, it is premature to extend the realm of program evaluation too much and we might as well concentrate on individual programs as the unit of program evaluation.

5. Lessons from Korean Experiences

Despite its short history, the performance management system of budgetary programs has established itself as an important institutional element in the budget process in Korea. Needless to say, it is still a work-in-progress and nobody knows how it would evolve in the future. Still, various evidences suggest that that effort to introduce a strong and efficient performance management system by Korean government is somewhat successful. Many lessons can be learned from the Korean experience and we will discuss four of them that are thought to be of particular importance to other developing countries intending to introduce a performance management system for budgetary expenditure programs. First, a systemic approach should be pursued in establishing a performance management system. As discussed in Chapter 2, there are several evaluation techniques available and each country may choose different combinations of evaluation methods. The Korean government chose to build the performance management system based on three evaluation methods; performance monitoring, performance review, and program evaluation. They are named as the performance goal management of budgetary programs, the self-assessment of budgetary programs and the in-depth evaluation of budgetary programs, respectively. As shown in [Figure 4-14], the three evaluation techniques constitute a system with a hierarchical structure in terms of complexity and difficulty of the techniques. PGM is relatively easy to implement and does not require a lot of expertise while IEBP is the most complex and expensive evaluation method among the three techniques. In return for larger cost, IEBP provides better information in terms of accuracy and objectivity than the other evaluation methods. The three components complement one another to produce performance information in high quality that they may as well be regarded as an indivisible unit. One thing to note is the fact that the Korean government took a cautious approach in building the system by introducing the three components in a sequential manner. PGM was introduced in 2003, SABP in 2005, and IEBP in 2006, respectively. By doing so, program offices were given enough time to adjust themselves to a new policy environment that a smooth transition was fulfilled without major resistance.

Figure 4-14 | The Structure of the Performance Management System of Budgetary Programs in Korea



Second, for successful establishment of the performance management system, it is necessary to build a system in which the program performance is explicitly tied to the budget. Researches on the bureaucratic behavior[31] indicate that the objective of bureaucrats is not to maximize welfare of the public but to maximize the benefit of the organization they belong to. Indeed, bureaucrats have the tendency to strive to make the amount of resources they can control as large as possible while public interest is better served by maximizing program performance. Therefore, from the perspectives of the public, it is very important to make the bureaucrats behave in the interest of the public. An obvious option is to monitor bureaucrats very tightly but it costs a lot of resources. In addition, it is physically impossible to completely eliminate bureaucrats' self-interested behavior through monitoring. An alternative is to provide an incentive scheme that can induce bureaucrats to choose to behave for the interest of the public without outside intervention or monitoring. We can accomplish it by linking budgetary decision making to evaluation results. Since the objective function of bureaucrats consists of the amount of resources they can control, bureaucrats are led to serve the interest of the public by providing an incentive structure in which the interest of bureaucrats is aligned to be compatible with that of the public. An incentive scheme that makes the interest of agents coincide with that of the principal

31. See Tullock (1965), Buchanan *et. al.* (1980).

is called incentive compatible and explicit linkage between budget size and performance found in result oriented performance management system is an example of incentive compatible compensation scheme. The promise or announcement of the budget office to tie evaluation results to budgetary decision making is effective only if program offices believe in it. If program offices think that the budget office's announcement is incredible, effectiveness of incentive compatible compensation scheme is significantly reduced. One obvious way to add credibility is to enact a law that declares the tie between evaluation results and the budget is now supported by legal force rather than promise or policy of the budget office. In Korea, the credibility of the system is supported by the enactment of two laws that make information from the performance management system as indispensible components of official budgetary documents and require the results of performance evaluations to be incorporated into the budgetary process. The legal foundation for the performance management system is particularly important in developing countries in which the accumulation of social capital is insufficient that credibility or trust on promise or policy without legal power is generally low.

Third, it is essential to successful establishment of the performance management system to gain support and cooperation from program offices. Program offices generally show a passive attitude toward assessment or evaluation on the programs they manage. The reluctance of program offices becomes more conspicuous especially when the evaluation project is conducted by outsiders or unfavorable evaluation results are expected. However, performance evaluation may help program offices in the long run by providing them with the opportunity to improve program performance. The regular evaluations may produce information useful in modifying program design for better performance. Under a result oriented performance management system, better program performance leads to bigger discretionary power over budgetary resources for the government office managing the program. It is important to make the program offices acknowledge that performance evaluation ultimately helps them. They would accept the performance management system more proactively when it is thought to be consistent with their own interest. Specifically, it is necessary to educate all stakeholders including government employees on the importance of the performance management system and continuous evaluation. More importantly, it is also necessary to manage evaluation projects transparently by allowing representatives from program offices to participate and express opinions. The Korean system provides program offices with opportunities to express opinions on the evaluation projects as well as evaluation results. Program offices take the primary initiative and lead the whole process in PGM. As for SABP, program offices are primarily responsible for the operation of the evaluation process and MOSF acts as the facilitator by providing technical support and objective feedback. Program offices cannot exercise much control over the evaluation process since it is an evaluation by independent external evaluators. Yet, MOSF takes much

effort to induce cooperation from program offices in IEBP by involving them in the process as a regular member of the evaluation task force.

Fourth, it is very important to construct a supporting scheme to assist technical and administrative aspects of the performance management system since the whole process of the performance evaluation requires technicality and expertise. MOSF of Korea distributes manuals for PGM and SABP to make evaluation tasks as easy and simple as possible. MOSF also provides the checklist for SABP to help program offices. Significant benefits are expected from the effort to make the evaluation process simple and standardized. It helps lessen the burden of program offices and reduce the possibility of committing serious mistakes by program offices due to the lack of expertise. It would also be very helpful to involve external institutions with expertise in the evaluation process. The Center for Performance Evaluation and Management at Korea Institute of Public Finance is actively involved in both PGM and SABP. The Center assists MOSF by providing technical advice and administrative support. In addition, Public &Private Infrastructure Investment Management Center at Korea Development Institute is actively involved in the management of IEBP. These research institutes play an important role by conducting academic and practical researches on the performance management of the government sector and offering policy recommendations.

Angrist, J., and J.-S., Pischke, *Mostly Harmless Econometrics*, Princeton University Press, 2008.

Buchanan, J., Tollison, R., and G. Tullock, *Toward a Theory of the Rent-Seeking Society*, Texas A&M University Press, 1980.

Donaldson, S., "Mediator and Moderator Analysis in Program Development", In S. Sussman (ed.), *Handbook of Program Development for Health Behavior Research,* Sage Publications, 2001.

European Commission, *Evaluating EU expenditure Programmes*, January, 1997.

----------, "Evaluation of EU Activities; An Introduction," January, 2005.

Guo, S., and M. Fraiser, *Propensity Score Analysis: Statistical Methods and Applications*, Sage Publications, 2009.

Hair, J., Black, W., Babin. A., and R. Anderson, *Multivariate Data Analysis*, 7th Eds., Prentice Hall, 2009.

Harman, H., *Modern Factor Analysis*, University of Chicago Press, 3rd Eds., 1976.

Hatry, H., *Performance Measurement: Getting Results*, Urban Institute Press, 1999.

HM Treasury, Cabinet Office, National Audit Office, Audit Commission, and Office for NationalStatistics, "Choosing the Right FABRIC: A Framework for Performance Information," March, 2001.

Kim, J., and N. Park, "Performance Budgeting in Korea", *OECD Journal on Budgeting*, Vol. 7, No. 4, 2007, pp.1-11.

Koh, Y., and J. Kim, *Mannual: In-depth Evaluation of Budgetary Programs*, 2nd Eds., Korea Development Institute, 2007, in Korean.

Koh, Y., Yoon, H., and J. Lee, *Performace Management of Public Sector*, Korea Development Institute, 2004, in Korean.

Lee, W., "Anaysis on the Reliability of Self-Assessment of Budgetary Progrmas", 2010, *in Korean*.

Ministry of Strategy and Finance, *Manual: Self-Assessment of Budgetary Programs*, various years, in Korean.

National Assembly Budget Office, *Evaluation of Performance Report for FY 2010*, 2011, *in Korean*.

----------, *Evaluation of Performance Plan for FY 2013*, 2012, *in Korean*.

OECD, "Improving Evaluation Practices: Best Practice Guidelines for Evaluation and Background Paper", PUMA/PAC(99)1, 1999.

Office of Management and Budget, *Guide to the Program Assessment Rating Tool*, January, 2007.

Office of Management and Budget, "Performance Measurement Challenges and Strategies", June, 2003.

Park, H., *Current Issues and Policy Task in Self-Assessment of Budgetary Programs*, National Assembly Budget Office, 2009, *in Korean*.

Park, H., Ryu, D., Park, N., Baik, W., and S. Hong, *Improving Fiscal Institution and Fiscal Management System*, Korea institute of Public Finance, 2012, *in Korean*.

Park, M., and J. Park, "Research on Performance Management System to Improve Efficiency of Budget Management: SABP", Korea Institute of Public Administration, 2008, *in Korean*.

Park, N., and J. Won, *Performance Analysis of Budgetary Programs and Policy Implications*, Korea institute of Public Finance, 2012, *in Korean*.

Performance-Based Management Special Interest Group, *Performance-Based Management Handbook*, 2001.

Perrin, B., "Implementing the Vision: Addressing Challenges to Results-Focused Management and budgeting", OECD, 2002.

Politt, C., "Integrating Financial Management and Performance Management", *OECD Journal on Budgeting*, Vol.1, No.2, 2002, pp.7-37.

Treasury Board of Canada, *Program Evaluation Methods: Measurement and Attribution of Program Result*, 3rd Eds., March, 1998.

Tullock, G., *The Politics of Bureaucracy*, University Press of America, 1987.

www.ksp.go.kr

Ministry of Strategy and Finance, Republic of Korea

339-012, Sejong Government Complex, 477, Galmae-ro, Sejong Special Self-Governing City, Korea Tel. 82-44-215-2114 www.mosf.go.kr

KDI School of Public Policy and Management

130-722, 85 Hoegiro Dongdaemun Gu, Seoul, Korea Tel. 82-2-3299-1114 www.kdischool.ac.kr

9 791155 450352 94320

ISBN 979-11-5545-035-2

Knowledge Sharing Program
Development Research and Learning Network

- 130-722, 85 Hoegiro Dongdaemun Gu, Seoul, Korea
- Tel. 82-2-3299-1071
- www.kdischool.ac.kr