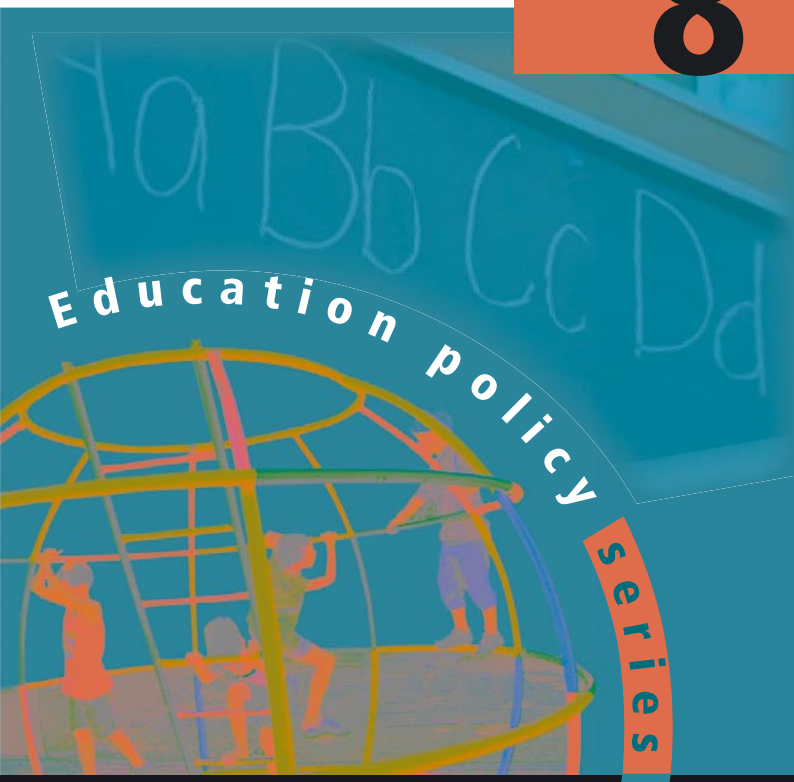
Program evaluation:
large-scale and small-scale studies

Lorin W. Anderson
and T. Neville Postlethwaite

8



Education policy series

International Academy of Education
International Institute for Educational Planning

The International Academy of Education

The International Academy of Education (**IAE**) is a not-for-profit scientific association that promotes educational research, its dissemination, and the implementation of its implications. Founded in 1986, the Academy is dedicated to strengthening the contributions of research, solving critical educational problems throughout the world, and providing better communication among policy makers, researchers, and practitioners. The seat of the Academy is at the Royal Academy of Science, Literature and Arts in Brussels, Belgium, and its co-ordinating centre is at Curtin University of Technology in Perth, Australia.

The general aim of the Academy is to foster scholarly excellence in all fields of education. Towards this end, the Academy provides timely syntheses of research-based evidence of international importance. The Academy also provides critiques of research, its evidentiary basis, and its application to policy.

http://www.smec.curtin.edu.au/iae/

The International Institute for Educational Planning

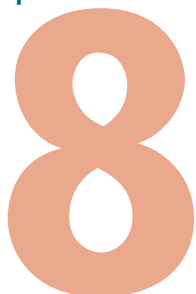
The International Institute for Educational Planning (**IIEP**) was established in Paris in 1963 by UNESCO, with initial financial help from the World Bank and the Ford Foundation. The French Government provided resources for the IIEP's building and equipment. In recent years the IIEP has been supported by UNESCO and a wide range of governments and agencies.

The IIEP is an integral part of UNESCO and undertakes research and training activities that address the main priorities within UNESCO's overall education programme. It enjoys intellectual and administrative autonomy, and operates according to its own special statutes. The IIEP has its own Governing Board, which decides the general orientation of the Institute's activities and approves its annual budget.

The IIEP's mission is capacity building in educational planning and management. To this end, the IIEP uses several strategies: training of educational planners and administrators; providing support to national training and research institutions; encouraging a favourable and supportive environment for educational change; and co-operating with countries in the design of their own educational policies and plans.

The Paris headquarters of the IIEP is headed by a Director, who is assisted by around 100 professional and supporting staff. However, this is only the nucleus of the Institute. Over the years, the IIEP has developed successful partnerships with regional and international networks of individuals and institutions – both in developed and developing countries. These networks support the Institute in its different training activities, and also provide opportunities for extending the reach of its research programmes.

http://www.unesco.org/iiep/



ISBN: 978-92-803-1299-7

Preface
Education Policy Series

The International Academy of Education and the International Institute for Educational Planning are jointly publishing the Education Policy Series. The purpose of the series is to summarize what is known, based on research, about selected policy issues in the field of education.

The series was designed for rapid consultation "on the run" by busy senior decision makers in Ministries of Education. These people rarely have time to read lengthy research reports, to attend conferences and seminars, or to become engaged in extended scholarly debates with educational policy research specialists.

The booklets have been (a) focused on policy topics that the Academy considers to be of high priority across many Ministries of Education – in both developed and developing countries, (b) structured for clarity – containing an introductory overview, a research-based discussion of around ten key issues considered to be critical to the topic of the booklet, and references that provide supporting evidence and further reading related to the discussion of issues, (c) restricted in length – requiring around 30-45 minutes of reading time; and (d) sized to fit easily into a jacket pocket – providing opportunities for readily accessible consultation inside or outside the office.

The authors of the series were selected by the International Academy of Education because of their expertise concerning the booklet topics, and also because of their recognised ability to communicate complex research findings in a manner that can be readily understood and used for policy purposes.

The booklets will appear first in English, and shortly afterwards in French and Spanish. Plans are being made for translations into other languages.

Four booklets will be published each year and made freely available for download from the websites of the International Institute for Educational Planning and the International Academy of Education. A limited printed edition will also be prepared shortly after electronic publication.

This booklet

It is increasingly incumbent upon ministries of education to build evaluation into new programs – especially those programs where substantial amounts of money are being spent. Each new program will usually be accompanied by questions about the impact and effectiveness of the program. For example, ministries may direct increased resources to classrooms and schools. Question: "Have these resources gone where desired and have the increased resources had an effect on student achievement?" Or, a new curriculum may have been introduced. Question: "How was the curriculum introduced and what problems occurred with its implementation?" Or, a new teacher in-service program may have been developed. Question: "Did the teachers learn what they were meant to learn? And, if so, did what the teachers learn have an effect on what students learned in terms of achievement, attitudes, and/or behaviour?" These are some examples for the case of general school education, but the same is true for new programs in pre-schools, in schools for the handicapped, in vocational education, and so on. All education programs need to include an evaluation component if their success is to be determined, and if weaknesses in the programs are to be identified and corrected.

When introducing new education programs it is not easy to assess whether they have had an effect on student learning. The kind of research design needed to get at the true cause of changes in student learning will vary according to the type of learning specified in the program goals. At the same time, however, there are some similarities in terms of sound and defensible evaluation designs. For example, it is always important that some measure of student learning be made at the beginning of a program. Education programs cannot be said to be effective if there are no measurable improvements in student learning over time. Similarly, some comparison group, or groups, of teachers and students should be included in the study. If there are measurable changes in student learning over time, but the magnitude of the changes is not

different from changes that occur in non-program students, then the program cannot be said to be effective.

This booklet is about small-scale and large-scale program evaluation studies. In many cases, both small- and large-scale studies are needed within any one evaluation project. In this booklet the authors argue that there is a need for both kinds of studies – provided that they are conducted according to scientific standards.

This booklet is the eighth in the Educational Policy Series developed by the International Academy of Education and the International Institute for Educational Planning. Each booklet seeks to bring research evidence to bear on important topics in educational policy.

Lorin Anderson

is a Distinguished Professor (Emeritus) at the University of South Carolina, United States, where he served on the faculty for over 30 years. His research interests have focused on the areas of effective teaching, affective assessment, and curriculum design. He is the senior editor of* A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy *(Allyn & Bacon, 2001) and he is a Fellow of the International Academy of Education.

Neville Postlethwaite

has been active in evaluation in education for the past 50 years and in particular in the evaluation of educational systems. He played a central role in the establishment of the International Association for the Evaluation of Educational Achievement (IEA). He is Professor of Education (Emeritus) at the University of Hamburg, Germany, and is still actively involved in a range of cross-national research projects concerned with monitoring the quality of education. He is a Fellow of the International Academy of Education.

Jointly published by:

The International Institute for Educational Planning (IIEP)
7-9 rue Eugène Delacroix
75116 Paris
France

and:

The International Academy of Education (IAE)
Palais des Académies
1, rue Ducale
1000 Brussels
Belgium

Design and layout by: Sabine Lebeau

ISBN: 978-92-803-1299-7

Table of contents

1 Introduction

Some basic definitions and terminology

Ministries of education have to evaluate the new programs they introduce. If they do not, they have no idea if a new program has been implemented as designed and, not what effect the program has had on student achievement, attitudes, and/or behaviours. Evaluation becomes especially important when large amounts of money are invested in a program. This usually means building in evaluation from the very onset of the program in order to see if the program has had the desired (or any) effect.

In educational evaluation there is a great deal of jargon – such as "formative and summative evaluation", "experimental design", "survey research", "quantitative and qualitative" studies, and numerous other technical terms. This jargon can be very confusing for the lay reader. In this booklet the concepts and terminology of large and small-scale evaluation studies have been examined within the framework of program evaluation. However, no distinction has been made between quantitative studies and qualitative studies because that will be the subject of another booklet.

Program evaluation may be defined as the systematic collection of information about the characteristics, activities, and outcomes of programs for the purpose of making judgements about program effectiveness, improving program effectiveness, and/or informing decisions about future program development (Centers for Disease Control and Prevention, 1999). As this definition suggests, the evaluation of a program can emphasize its characteristics, its activities, its outcomes, or some combination of these.

"Characteristics" include the context within which the program is operating, the resources used to support the program, and the staff used to implement the program.

"Activities" include those actions that must be taken in order to implement the program properly. Instructional strategies, teaching practices, teacher behaviours, student behaviours, teacher-student interactions, and student work are but a few of the activities associated with school programs.

"Outcomes" refer to the impact of the program on those for whom the program is intended. These outcomes may be specified in advance (such as in the form of program goals) or may be unintended outcomes of program design and/or implementation (which are often referred to as "consequences").

In addition to differences in characteristics, activities, and outcomes, program evaluation can differ in terms of its purpose(s). Program evaluation can focus on program effectiveness. Typically, this is referred to as "summative evaluation" (Scriven, 1967). Program evaluation can also focus on program improvement. Traditionally, this has been labelled "formative evaluation." In addition, program evaluation can focus on informing decisions about future program development. Typically, this involves both formative and summative evaluation and places a greater premium on the decision-maker than on the evaluator. For this purpose, clear and meaningful communication between evaluator and decision-maker is absolutely essential.

Finally, although generally termed "program evaluation," the evaluation concepts, principles, and procedures described in this booklet also pertain to the evaluation of educational systems. In many respects, an educational system can be viewed as a set of programs with complementary goals that are organized into various administrative units and are consistent with, or facilitative of, the overall mission of the educational system.

The purpose of this booklet is to describe the current state of knowledge of the field of program evaluation. Throughout the booklet, an emphasis has been placed on basic concepts and principles and generally accepted best practices. The booklet is organized around seven principles of sound program evaluation. A bibliography of pertinent publications has been included at the end of the booklet.

2 Differences between large-scale and small-scale studies

Large-scale and small-scale evaluation studies differ in their purposes and scopes as well as the generalizability of their results.

There are five main features that differentiate large-scale from small-scale evaluation studies. Although the most obvious difference is size, the studies also differ in terms of their purpose, cost, generalizability of results, and type and complexity of data analysis.

a. Size of sample

Small-scale studies usually have small samples. Small usually means not more than about 30 schools or 60 classes. Since these are not enough to cover the variation in a total grade group in a country, small-scale studies often cover only a region or district and not the nation as a whole. Large-scale studies usually encompass over 100 schools and some 20 to 30 students within each of the schools. In this case a sample of schools is drawn from the whole system of education.

b. Purpose of study

Small-scale studies tend to be used for exploratory purposes and focus on developing measures, developing curricula and textbooks, developing new teaching methods, and so on. Large-scale studies are used for describing the system as a whole and parts within it (for example, regions within the system). The descriptions usually involve achievement in different subject areas, attitudes and behaviours of students and teachers, and, quite often, the relationships among many of the factors operating in the system. Small-scale studies are usually used in formative evaluation whereas large-scale studies are usually used for summative evaluation purposes.

c. Cost

Small-scale studies cost much less than large-scale studies. To undertake a data collection involving 200 schools across a nation is quite costly – not only in terms of the actual collection of data, but also in terms of the data management (data entry, data cleaning, and data analysis). For the same cost a small-scale study could involve many observations and include many variables, thus producing a more fine-grained description of a limited number of schools.

d. Type and complexity of data analysis

In large-scale studies more complex statistical analyses are often required. For example, with large numbers of schools and teachers, the simultaneous effect of many variables on achievement can be examined. This examination is possible even when the variables represent different units of analysis (for example, student, classroom, and school). These kinds of analyses are important in order to identify interactions of variables and their effect on achievement. In small-scale studies it is not possible to use these kinds of analyses because of too few schools or students. Small–scale studies that include additional amounts of qualitative data can sometimes become quite expensive because the collection of such data is labour intensive.

e. Generalisability of results

If a study targets a single grade level and subject area in an educational system and samples schools and/or classrooms at that grade level and that subject area within that system, the results can be generalized to that grade level and subject area for the entire system within certain limits (the standard errors of sampling). To know these limits, the sampling errors must be reported for all estimates included in the study. When small-scale studies are used, such generalizations are difficult because the sampling errors are very large.

3 Linkages between small-scale studies and large-scale studies

Using small-scale studies to design data collection instruments and data management procedures for large-scale studies.

Some evaluators see small-scale and large-scale studies as incompatible. In fact, small-scale studies are an important part of the overall evaluation process because they may be used to make decisions about what to measure, how to measure, the choice of procedures for data collection and data entry, and the preliminary testing of constructed composite factors.

a. Identifying what to measure

A small-scale study may be used to observe classes in several different schools in order to suggest factors that are making a difference between high-scoring schools/classes/students and low-scoring schools/classes/students. This list of factors can then be used to generate hypotheses that can be tested in a large-scale study.

b. Determining how to measure

The important factors identified in a small-scale study need to be measured. Small-scale studies permit the low-cost exploration of alternative approaches to measurement. This provides an opportunity to test each measure's validity and reliability, and to decide the best approach to measurement (for example: surveys, interviews, and observations).

c. Trialing instruments and data collection procedures

All instruments used in large-scale studies need to be trial-tested on a smallscale and then revised based on the

analysis of the try-out data. This enables the data collection procedures to be changed so as to ensure an improved data collection in the main study. In addition, codebooks that are required for data entry can be trialed before they are used in the main study.

d. Constructing composite factors

Composite factors constructed from several variables are often used as key independent variables in large-scale studies. For example, the education level and employment status of parents are often combined into a new composite factor called "socio-economic status." Small-scale studies permit all steps in the construction process to be fully tested before the main study.

Using small-scale and large-scale studies to develop and test new curriculum materials

The development and testing of curriculum materials requires a thoughtful mixture of small-scale studies that are employed in the development phase, and large-scale studies that are employed in the testing phase.

There are six main steps in the development of curriculum materials.

a. Establishing educational goals and objectives

The goals of an educational system are usually established by governments. However, these goals are often rather abstract. In order to provide more specific guidance for the development of curriculum materials, further consultations are required with important stakeholder groups so as to generate more detailed (operationalized) goals and objectives. For example, the classification of goals and objectives might commence by consulting with small samples of employers, academics, teachers, and the public. These consultations will provide information about valued knowledge, skills, and attitudes. In addition, small-scale studies may be undertaken to identify how the situations and expectations of different categories of students (based on gender, socio-economic background, geographic locations, etc.) have changed and the implications these changes have for modifications in the curriculum.

b. Writing the units or modules

This step requires the establishment of an agreed set of descriptive parameters for the curriculum materials (such as structure, content, and format), and these may be informed by data collected from a small-scale study.

c. Small-scale trial testing of materials

In this step a small-scale study is usually undertaken in six to eight schools in a manner that permits a good deal of interaction between the curriculum developers and the teachers in the try out schools. This small coverage of schools enables the collection of in-depth information. For example, tests are developed for each curriculum module in order to identify which objectives are being poorly achieved, and to determine the reasons for this poor performance. Teachers are asked about problems they experience with each portion of the written materials and this information is linked to the actual test results.

d. Large-scale evaluation of the new materials

This step involves conducting a large-scale study in which a sample of about 100 schools is drawn to represent the total range of schools for which the curriculum is intended. The participating teachers are trained so they understand and are able to implement the new curriculum properly or faithfully. In general, the curriculum developers need to know if an objective is being well achieved (with about 80 percent of the students answering related items correctly), moderately achieved (with about 50 percent answering related items correctly), or poorly achieved (with fewer than 30 percent answering related items correctly). The weaker parts of the curriculum are identified and the necessary revisions are made. In addition, the teacher training materials and procedures are completed and documented.

e. Large-scale evaluation of teacher training for the new materials

This step involves an examination of the appropriateness and effectiveness of the teacher-training materials for a large sample of teachers. These materials must help teachers understand the curriculum and acquire the

skills needed to effectively implement the curriculum as designed. The effectiveness of the training in terms of its desired changes in teacher knowledge and skills should be evaluated at this time.

f. Large-scale global evaluation study

The final step in the process takes place after the curriculum has been in the schools for two or three years. At this point a large-scale evaluation is undertaken. This study involves a probability sample of schools being drawn (see below) such that the errors in estimating student achievement are not too large. In addition, comparison groups of teachers and students are identified and baseline measures of "program" and "non-program" students are taken.

5 Sample design for small-scale and large-scale studies

Different sampling procedures for schools and students are required for small-scale and large-scale studies. Large-scale studies are usually conducted on probability samples of schools and students, while small-scale studies are usually based on judgement samples.

a. Probability samples for large-scale studies

Large-scale studies must have probability samples such that each person in the target population has a non-zero chance of selection into the sample. Typically, educational evaluators include two-stage samples, with the schools being selected first with probability proportional to the size of the defined population within each school. Students are then selected at random within each selected school. Occasionally, three-stage samples are used, with regions or districts being selected at the first stage. Given that sampling frames and response rates are seldom perfect, sampling weights must be calculated in order to adjust for variations in probabilities and non-response. Finally, standard errors of sampling need to be calculated for every estimate given in the evaluation report.

b. Judgement samples for small-scale studies

In small-scale studies, judgement samples are selected. The sample schools are selected so as to provide a good coverage of the variety of schools in a school system. Before the samples are drawn, it is important that the evaluators obtain relevant information about the schools

that might be included in the study. If, for example, the aim of the study is to examine differences in mathematics achievement, then the schools are selected so as to span the known differences in mathematics achievement. If the study is intended to examine differences in behavioural problems in schools, then the sample must be based on, for example, information about the severity of behavioural problems. "Rectangular" samples are often drawn so that the performance of students may be tested at the extremes of the distribution.

6 Reporting small-scale and large-scale studies

The report of a small-scale or large-scale evaluation study must be sufficiently detailed to permit other evaluators to replicate the study.

One of the most important issues in conducting an evaluation study is the report of the research that has been undertaken. The evaluation study report should address the areas below:

a. Aims of the study

The aims of the study and the related research questions for the study must be stated very clearly and their relevance must be obvious. Because the emphasis in many program evaluations is on program effectiveness, this means designing a study that enables the evaluator to attribute changes in student learning (achievement, attitudes, and/or behaviours) to the program *per se*, and **not** to a whole host of extraneous factors (for example, differences in prior student learning, differences in teacher quality, differences in school organizational structure, etc.).

b. Questionnaires

The questions included in the questionnaire must be thoroughly trialled. Where scales are constructed from groups of questions, there must be clear evidence that the reliability of the scale is high and that the scale makes a valid assessment of the construct that is being considered.

c. Tests

The specific cognitive objectives must be clearly defined before item writing begins. Items must be written, checked, and then trialled, with poor items being discarded. The final

structure of the test (whole test or parts of the test) must be consistent with the research questions.

d. Reliability and validity

Tests, questionnaires, and any other instruments used in all evaluation studies should result in measures that have high reliability and validity. When observation is used, the observation schedule must be prepared in such a way that any trained observer can use it to produce accurate records or codings based on the observations. Furthermore, there should be high inter-observer reliability (that is, agreement among observers). Where the observer is expected to describe, rather than infer, what is observed, the inter-observer reliability should be very high (not less than 0.90). When inference on the part of observers is required, the inter-observer reliability is generally lower, but should not fall below 0.80.

e. Sampling

The target population of the study should be precisely described. If a judgement sample is required, it should be clear if it was a rectangular distribution. If a probability sample is required, it should be clear what level of sampling error will be tolerated (for example, typically not greater than 5 percent for a percentage). In addition, information should be provided on how the sample was drawn and how the sample weights (if any) were calculated. Decision rules for excluding students in the defined population from the study should be established – with not more than 5 percent of the students in the defined population allowed to be excluded.

f. Conduct of the study

The steps for producing the data collection instruments, contacting schools, tracking schools, teachers and students, selecting and training the data collectors, administering the instruments, and returning the instruments to a central place should be made explicit.

g. Data entry and cleaning

The preparation of the data entry program, the training of those entering the data, the steps involved in data entry and data cleaning, and the rules employed to make decisions about the data should be clear.

h. Data analysis

The different types of data analysis that will be performed should be clear and appropriate for addressing the research questions. In addition, the types of data analysis should be specified early in the planning process. Additional data analyses can be performed as the situation warrants.

i. Dissemination of results

It is very useful to have different reports produced. One is the research report itself which explains all. A second is a short report for senior policy-makers. A third, for the public, is not easy to write, because it must be written simply, but without distorting the results. Some of the most successful evaluation studies have had their results disseminated through the medium of television and, increasingly, on the Internet.

7 Interpreting the results for small-scale and large-scale studies

The results of evaluation studies do not speak for themselves; they must be interpreted for policy-makers and other interested parties.

Interpretation is the process of attaching meaning to the data that are collected, analyzed, and reported. Too often, data are collected and analyzed but the next step (putting the results in context and making sense of the results before reporting them) is not taken. For example, what does it mean that 45 percent of adolescents responding to a questionnaire believe that drinking alcohol and using drugs are harmful to their health? Is this good? Is it increasing or decreasing? How does this compare with other schools, local education agencies, or countries? What does it mean in terms of health and safety? What does it mean in terms of the effectiveness of drug and alcohol prevention programs? When concerns for interpretation are addressed, there are some general guidelines that should be followed.

a. Using paradigms

Interpretation always requires looking at the data through some lens – some framework – some model – some paradigm. The data collected from observations in classrooms can be interpreted from the perspective of the process-product model framework (where teacher behaviours are believed to influence student achievement directly) or the mediating process model (where the perceptions and behaviours of students are believed to mediate or intervene between the teacher behaviours and student achievement). Although the data are the same,

the interpretations of the data will differ depending on the model (lens) through which the data are viewed or examined.

b. Using comparisons

Interpretation almost always involves some type of comparison. Baseline data, control or comparison groups, predefined standards of expected performance, and standards indicating statistical significance are all examples of potentially relevant comparisons. The use of baseline data means that data are collected prior to the adoption and implementation of a new or revised program. Post-implementation data are then compared with the baseline data to assess change over time. Control or comparison groups are groups of students who are not participants in the program being evaluated but are similar to the students who are program participants in many respects (for example, gender, ethnic group, and prior achievement). Comparing, say, test scores of students in the program with those not in the program permits an interpretation of the impact of the program on student achievement. Predefined standards of expected performance are levels of performance that will indicate that the program has been successful or effective. For example, a vocational-technical program might be said to be "successful" if 80 percent of the students are employed in a program-related job within six months after graduation. Finally, standards of statistical significance are used in association with baseline data and control or comparison groups. Statistical significance indicates whether the difference between the baseline and post-implementation data, or between the program students and comparison students, are sufficiently large that they cannot be attributed to random events or chance.

c. Statistical versus substantive significance

When engaging in quantitative evaluations, it is important to remember that statistical significance does not always mean substantive significance. Winning the battle of statistical significance is not the same as winning the war of policy or practical relevance. A decrease of one or two percent in the rate of illiteracy will not be nearly as

persuasive to policy makers as a 10 percent shift, regardless of the level of statistical significance associated with the results.

d. Multiple data sources

Interpretation is enhanced when multiple sources of data are available and considered. If, for example, a program focuses on decreasing the number of students dropping out of school before school completion, data in addition to rates of dropping out will add to our understanding of the program effectiveness. Examples might include a student's academic performance, a student's sense of belonging, a student's level of engagement with the life of the school, and a student's friendship patterns within and outside school.

e. Generalizability

The range and limits of generalizability of the evaluation results should be discussed. A rigorous evaluation, unmarred by large errors and revealing important outcomes, can help demonstrate that a program has worked in a particular setting. Unfortunately, no single evaluation can demonstrate that the program will work equally well in another setting.

f. Different perspectives

Because interpretation depends on the lens or framework of the person making the interpretation, greater understanding usually results when multiple people are involved in examining and discussing the data. When the same interpretation is made by people operating from different perspectives, you really have something.

8 Writing evaluation reports for different audiences

In preparing evaluation reports, attention must be paid to the audience for whom the report is intended.

Although evaluation reports are written by evaluators, they are not typically written for evaluators. There are three main audiences for evaluation reports: educators, policy-makers, and the general public (including members of the news media). With few exceptions, members of these audiences do not possess an understanding of technical issues, nor are they familiar with the terminology used by those who conduct and report evaluations. When writing for these audiences, the guidelines presented below should be followed.

a. Target audiences

Write the report for a target audience of general readers who may be unfamiliar with the problems and issues under investigation. Include general readers in the group to whom you circulate drafts of the report – so that their reactions can be used to improve the report. In addition, report the results in concise and straightforward language which avoids excessive use of technical terminology and jargon.

b. Contexts

Be sure to include information about the context within which the evaluation was conducted. Contextual information is needed to help audiences interpret the evaluation. A study context should be described in sufficient detail that members of the various audiences can determine the likely impact of the context on project implementation and effectiveness.

c. Objectives and purposes

Link the results with the intended purpose and objectives for which the evaluation was conducted. Often a question-answer format is useful in this regard – with the questions derived from the intended purposes and objectives, and the answers derived from the data collected, analyzed, and interpreted.

d. Balance reporting formats

Avoid over-reliance either on either narrative or tables and charts. Strive for a balance between the two. When tables and charts are included, use the narrative to walk the reader through the data summarized in the tables and charts so as to enhance understanding on the reader's part.

e. Honesty

Most evaluations yield a mix of positive and negative findings, which should be presented in a balanced manner. When results are contradictory or inconsistent with the results of similar evaluations, provide reasonable explanations for the contradictions and inconsistencies.

f. Linking results and recommendations

Emphasize the link between the proposed recommendations and the results of the evaluation. Whenever possible, review the proposed recommendations with those responsible for implementing them before issuing a final report.

g. An executive summary

Always include an executive summary. Policy-makers, in particular, often do not have the time to read the entire report. The executive summary should include, at a minimum, the aims of the study and related research questions, the methods used to conduct the study, the major results of the study, and the recommendations for policy, practice, and/or research that follow from the results.

h. Conflicts in interpretation

If the report is prepared by a team of evaluators, be sure that any differences in interpretation or perspective are addressed and either resolved or included in the report in the form of comments in the text or as a special section. At

the same time, however, one and only one person should be responsible for the quality and completeness of the final written report.

i. Polishing and proofreading

Be sure to proofread and polish the final draft of the report prior to dissemination. Typically, multiple readers are needed to catch all of the errors and to improve the general readability of the report. Although tedious, proofreading contributes greatly to the appearance of a quality study and report.

9 Avoiding inherent dangers

Several dangers that may prevent an evaluation report from being used for the purposes for which it was intended.

Evaluation studies often challenge established procedures and practices – and therefore they can be viewed negatively or with a great deal of suspicion. The following are among the most frequently occurring dangers inherent in evaluation studies.

a. Suppression of results

Bureaucrats acting on behalf of governments may decide what aspects of the study should and should not be allowed to be made public. A single person has been known to read and cross out in red those parts of a draft report that should not be included in the final report. The reason for this practice is that the person fears that a particular result could embarrass the government. Governments may also suppress the publication of the results of an evaluation study by providing a variety of reasons to justify their doing so. Among the reasons given are (i) the study leaves important questions unanswered; (ii) much of the evidence is inconclusive; (iii) the figures are open to multiple interpretations; (iv) certain findings are contradictory; and (v) some of the main conclusions and recommendations have been questioned (Lynn and Jay, 1985). Yet another form of suppression is to say that the research needs to be replicated before the results of the present study can be published.

b. Ranking of countries and regions or schools

When evaluation studies have included all schools within a country it is sometimes the case that schools or regions have been ranked and the rankings made public. The rationale given is that "parents want to know." It is easy for the teachers in a school in a high socio-economic area to have their students learn a lot and do well. It is much more difficult for teachers in schools in very poor socio-economic areas to have their students achieve at this same level. As a consequence, some countries produce rankings only after taking into account (or controlling for) socio-economic differences. There is little point in ranking schools in studies where only a small sample of schools has been involved. However, when all regions have been included in a study and where the estimates of achievement are valid and accurate, it is reasonable to use the data to rank regions. It is, of course, a political matter as to whether it is wise to do this.

The rank a country has in an international study of achievement is purely a matter of which other countries are in the study. Much more important than merely ranking countries is understanding the reasons for the differences among countries. Armed with this understanding, improvement in the educational systems of the poorly performing countries becomes a possibility.

c. Basing generalizations on only a few schools

There may well be a case where a very good study is conducted on, say, 12 schools. However, there is no way that 12 schools can be a good sample of all schools except in a small country where all of the schools are very similar in terms of various demographic factors (for example, size, socio-economic status of students, etc.). Because this is usually not the case, great care must be taken not to generalize the results of a study containing very few schools to an entire region or country.

d. Evaluation studies that include high-stakes testing

High-stakes tests are those whose results are used to make life-changing decisions. The decisions may pertain to a school, a teacher, or a student. For example, in some school systems a school that has very poor achievement results

over a period of three or four years can be closed down or have the entire staff replaced. Similarly, an individual teacher who has consistently had very poor results with her students over a similar time period may well be singled out for special attention. When high-stakes testing is involved (i) all students in all schools at a certain grade level must be included in the study, and (ii) the results of the study must be consistent over some reasonable time period (for example, three to five years).

e. Confusion of correlation and causality

The purpose of most program evaluations is to link program implementation with some measure(s) of student learning. Under most circumstances, however, there are factors other than the program that could influence student learning. If program implementation is one of several factors associated with increased student learning, it can be said that program implementation is correlated with student learning. However, unless the other factors are eliminated, it cannot be said that program implementation **caused** student learning to increase. Poorly designed evaluation studies can result in the confusion of correlation with causality. Well designed evaluation studies are needed if we are to properly disentangle correlation from causality.

10 Conclusion

In this booklet it has been stressed that program evaluation studies, whether large-scale or small-scale, need to be planned from the beginning of a program and must be conducted with scientific rigour. Both kinds of study usually form part of the whole bundle of evaluations of a particular program. An example was given of the use of both kinds of studies in a program to develop curriculum materials.

Small-scale studies can be used for identifying important variables or factors that should be included in a large-scale study. They can also be used for trial testing test items and questionnaire questions, for deciding on the best way to form factors or constructs, and for having some detailed examples of general points that may emerge from a large-scale study. Small-scale studies can also be used to trial-test procedures that are to be used in large-scale studies. But it is impossible to generalize from small-scale studies. Generalizations can only be made from large-scale studies based on probability samples.

The scientific rigour that must be used in both kinds of studies was exemplified by the example of good sampling procedures – whether for probability samples or judgement samples. Suggestions were made for the reporting of evaluation studies and the care to be taken when interpreting results.

Whichever kind of study is conducted, the reports are often written for different types of audiences. Care must be taken in the writing of these reports. Hints were given based on the experience of the authors. Finally, there are inherent dangers when conducting evaluation studies. Examples of the kinds of dangers were given and ways of dealing with them suggested.

References

Centers for Disease Control and Prevention. (1999). *Framework for program evaluation in public health.* Atlanta, GA: Author.

Lewy, A. (Ed.). (1977). *Handbook of curriculum evaluation.* Paris: International Institute for Educational Planning, UNESCO.

Lewy, A. (1991). *National and school-based curriculum development.* Fundamentals of Educational Planning Series No. 40. Paris: International Institute for Educational Planning, UNESCO.

Lynn, J. & Jay, A. (1985). *Yes, Prime Minister.* London: BBC.

Kish, L. (1987). *Statistical design for research.* New York: John Wiley & Sons.

OECD (2005). *PISA 2003 Technical report.* Paris: Author.

OECD (2005). *PISA 2003 Data Analysis Manual.* Paris: Author.

Patton, M.Q. (1997). *Utilization-focused evaluation: The new century text.* Thousand Oaks, CA: Sage.

Payne, D.A. (1994). *Designing educational project and program evaluations: A practical overview based on research and experience.* Boston: Kluwer Academic Publishers.

Schubert, W.H. (1986). *Curriculum: perspective, paradigm, and possibility.* New York: Macmillan.

Scriven, M. (1967). The methodology of evaluation, in R. W. Tyler, R. Gagné & M. Scriven (Eds.). *Perspectives of curriculum evaluation* . AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally.

Wolf, R. M. (1990). *Evaluation in education, 3rd edition*. Westport, CT: Praeger Publishers.

Payne, D.A. (1994). *Designing educational project and program evaluations: A practical overview based on research and experience*. Boston: Kluwer Academic Publishers.