



## Evaluating Driver Education Programs



*Prepared by*

Larry Lonero  
Kathryn M. Clinton

*Northport Associates*  
182 Bagot Street  
Cobourg, Ontario  
Canada  
K9A 3G2  
905-377-8883  
[www.northportassociates.com](http://www.northportassociates.com)

*Prepared for*



607 14th Street, NW  
Suite 201  
Washington, DC 20005  
800-993-7222  
[www.aaafoundation.org](http://www.aaafoundation.org)

*October 2006*

Management Overview

---

This report was funded by the AAA Foundation for Traffic Safety in Washington, D.C. Founded in 1947, the AAA Foundation is a not-for-profit, publicly supported charitable research and education organization dedicated to saving lives by preventing traffic crashes and reducing injuries when crashes occur. Funding for this report was provided by voluntary contributions from AAA/CAA and their affiliated motor clubs, from individual members, from AAA-affiliated insurance companies, as well as from other organizations or sources.

This publication is distributed by the AAA Foundation for Traffic Safety at no charge, as a public service. It may not be resold or used for commercial purposes without the explicit permission of the Foundation. It may, however, be copied in whole or in part and distributed for free via any medium, provided the AAA Foundation is given appropriate credit as the source of the material.

The opinions, findings, conclusions, and recommendations expressed in this publication are those of the authors and are not necessarily those of the AAA Foundation for Traffic Safety, any individuals who peer-reviewed this report, or advisory group members. The AAA Foundation for Traffic Safety assumes no liability for the use or misuse of any information, opinions, findings, conclusions, or recommendations contained in this report.

If trade or manufacturer's names are mentioned, it is only because they are considered essential to the object of this report and their mention should not be construed as an endorsement. The AAA Foundation for Traffic Safety does not endorse products or manufacturers.

©2006, AAA Foundation for Traffic Safety



# Contents

<b>Glossary of Acronyms</b>	<b>➤ 5</b>
<b>Acknowledgements</b>	<b>➤ 7</b>
<b>Preface</b>	<b>➤ 9</b>
<b>Introduction</b>	<b>➤ 11</b>
Why Evaluate Driver Education Programs? . . . . .	13
Does Your Program Have What it Takes to be Evaluated? . . . . .	14
The Goals and Objectives of Driver Education . . . . .	15
<b>Overview of Program Evaluation</b>	<b>➤ 19</b>
What is Program Evaluation? . . . . .	19
Defining the Two Basic Functions of Evaluation . . . . .	22
Demystifying Evaluation Concepts and Terms . . . . .	23
Evaluation Standards . . . . .	26
<b>Driver Education Evaluation—Past and Present</b>	<b>➤ 27</b>
Historical Overview of Driver Education Evaluation . . . . .	27
Reviews of Driver Education Evaluations . . . . .	28
Individual Evaluations of Driver Education Programs . . . . .	32
Limitations of Past Driver Education Evaluations . . . . .	36
<b>Techniques of Evaluating Beginner Driver Education</b>	<b>➤ 39</b>
Types and Levels of Evaluation . . . . .	39
Overview of Evaluation Steps . . . . .	43
Five Steps to Evaluating Driver Education Programs . . . . .	44
<b>Selected Details in Planning Evaluation Projects</b>	<b>➤ 47</b>
The Logic Model—A Logical Place to Start . . . . .	47
Evaluation Questions and Targets . . . . .	49
Overcoming Barriers to Evaluation . . . . .	51
Evaluation Resources and Help . . . . .	51
<b>Thinking Ahead—Evaluation in the Future of Driver Education</b>	<b>➤ 53</b>
Conclusion . . . . .	54
<b>References</b>	<b>➤ 55</b>
<b>Appendix A: Glossary of Terms</b>	<b>➤ 59</b>
<b>Appendix B: Evaluation Resources</b>	<b>➤ 69</b>




## **GLOSSARY OF ACRONYMS**

<b>AAA</b>	Formerly American Automobile Association
<b>AAAFTS</b>	AAA Foundation for Traffic Safety
<b>CAA</b>	Canadian Automobile Association
<b>DWI</b>	Driving while intoxicated
<b>GDL</b>	Graduated Driver Licensing Program
<b>NHTSA</b>	National Highway Traffic Safety Administration
<b>RCT</b>	Randomized controlled trial
<b>SES</b>	Socioeconomic status



## ACKNOWLEDGEMENTS

The authors would like to acknowledge the considerable contribution of several people to the development of *Evaluating Driver Education Programs: Comprehensive Guidelines* and its two companion documents, of which this is one. The *Guidelines* were developed with input from a small group of experts who served as project team members. Included were Mr. Dan Mayhew and Dr. Douglas Beirness from the Traffic Injury Research Foundation, Mr. Val Pezoldt from the Transportation Research Center of Texas Transportation Institute, Dr. Douglas Black of Northport Associates, Dr. Erik Olsen from the National Institute of Child Health and Human Development, and Mr. John Brock of General Dynamics Information Technology. The *Guidelines* and companion documents greatly benefited from the astute insight and thoughtful comment provided on project direction and drafts of the *Guidelines* by these team members.

The authors would also like to acknowledge the contribution of Advisory Group members, who attended a two-day workshop and provided comments on a draft of the *Guidelines*. This input was invaluable in conceptualizing and organizing the *Guidelines* and the companion documents.

The Advisory Group membership included:

Mr. Walter J. Barta, Alberta Motor Association  
Dr. Raymond Bingham, University of Michigan Transportation Research Center  
Mr. Eric Chapman, California Department of Motor Vehicles  
Professor Mary Chipman, University of Toronto  
Dr. Richard Compton, National Highway Traffic Safety Administration  
Dr. Susan Ferguson, Insurance Institute for Highway Safety  
Mr. John Harvey, Association of State Supervisors of Safety and Driver Education,  
and Oregon Department of Transportation  
Mr. Andrew Krajewski, Maryland Motor Vehicle Administration  
Mr. Scott Masten, University of North Carolina, Highway Safety Research Center  
Dr. James McKnight, Transportation Research Associates  
Dr. James Nichols, Nichols and Associates  
Dr. Judith Ottoson, Georgia State University  
Mr. Ray Peck, R.C. Peck & Associates  
Ms. Debbie Prudhomme, Driving School Association of the Americas  
Dr. Allen Robinson, American Driver and Traffic Safety Education Association

Dr. Michael Scriven, Western Michigan University

Dr. Bruce Simons-Morton, National Institute of Child Health and Human  
Development

Dr. William Van Tassel, AAA National Headquarters

The authors would like to extend their appreciation to the AAFTS and BMW of North America who supported this project, and to the Foundation staff who provided insights and guidance during the development of the *Comprehensive Guidelines* and companion documents.



## PREFACE

This document, *Evaluating Driver Education Programs: Management Overview*, provides an introduction to evaluating driver education programs. It is intended for driving school owners, driver educators, program managers, administrators, and others with limited background in research methods. The *Management Overview* provides a general introduction to the art and science of program evaluation, with a specific focus on how program evaluation concepts and methods can be applied to driver education evaluation.

There are two companion documents:

*Evaluating Driver Education Programs: Comprehensive Guidelines*, a more extensive and detailed evaluation manual; and

*Evaluating Driver Education Programs: How-To Guide*, a practical, hands-on guide on how to do basic formative evaluations.

The *Comprehensive Guidelines* provide a detailed background for planning, conducting, and integrating effective evaluation into beginner driver education program development and policy. Covering a range of evaluations from simple to complex, it is written primarily for program evaluators, researchers, and other technical audiences. The *Comprehensive Guidelines* include actual tools, such as questionnaires, focus group guides, and logbooks that can be used or adapted for evaluating beginner driver education programs.

The *How-To Guide* provides hands-on, step-by-step guidance for actually conducting a basic formative evaluation. Developed especially for driving school operators and owners, program developers, and managers, it is intended to assist basic evaluation for improving a driver education program.

Taken together, the three documents are intended to meet the needs of different people in the driver education field and to support better, more focused evaluations. The documents provide a set of tools that can be used to carefully and rigorously examine beginner driver education programs. It is hoped that their use will result in a growing body of evaluation data that can be built upon, leading to better driver education programs and, ultimately, safer young drivers.

The three documents and related evaluation resources are also available on the website of the AAA Foundation for Traffic Safety, [www.aaafoundation.org](http://www.aaafoundation.org).

*. . . “the most common characteristic of driver education evaluation has been the lack thereof.”*



# Introduction

This *Management Overview* provides a general introduction to the art and science of program evaluation as applied to driver education. Driver education, as we use the term here, means the initial, pre-licensing instruction of beginner drivers. Education or training for experienced licensed drivers is not included. As we shall see later, beginner driver education presents special challenges to the evaluator.

A brief summary of the earlier reviews of the evaluation literature is included in this *Overview*. Readers interested in the technical details of the evaluation literature should read the *Comprehensive Guidelines*. As the reviews show, there have been both good and not-so-good evaluations in the driver education field. The best studies have been well designed and meet reasonable quality standards. While there have been some very poor evaluations, the most common characteristic of driver education evaluation has been the lack thereof.

Beginner driver education is not alone in this, as very few programs that try to improve performance of all types of drivers have been evaluated in any objective, systematic way. Little evaluation takes place beyond the unsystematic, bureaucratic approach that researcher Pat Waller called “feel-good evaluation.” She defined this type of evaluation by the following example from driver improvement. “Convicted drivers went through driver improvement clinics, afterward maintaining that they had learned a lot, and everyone felt good about the programs” (Waller 1992, 109). Most driver programs continue because of political and bureaucratic beliefs that “something must be done,” or perhaps more precisely, it must appear that “something has been done.”

To help the field move beyond the current status, the *Overview* is focused on understanding basic evaluation concepts and methods and what it takes to perform a useful, credible evaluation. It is intended for driver educators, managers, and administrators with limited technical background in evaluation research or statistics. It avoids unnecessary jargon and defines essential technical terms and concepts for the benefit of generally interested

stakeholders in driver education evaluation, who are more likely to be “consumers” rather than producers of evaluations. This category includes:

- Private driving school industry operators and associations
- Potential investors in the driver education industry
- AAA and CAA Clubs
- Local and regional high school program coordinators
- State driver education administrators and regulators
- Insurers

The *Overview* is intended to provide a working knowledge of the entire range of evaluation levels, from program operations to safety impacts. It will not make the reader an evaluator, but it will help readers understand evaluation, talk to evaluators, and make decisions about evaluations.

While this document is specifically addressed to evaluation of driver education, the reader may also be interested in *The Art of Appropriate Evaluation: A Guide for Highway Safety Program Managers*, by the National Highway Traffic Safety Administration (NHTSA). This book and other printed and web-based resources are listed in Appendix B. The NHTSA guide is an excellent introduction to evaluation. However, it differs in purpose from this document. The NHTSA guide is intended for local and state managers of highway safety interventions, such as increasing seatbelt use, and it does not cover evaluation of ultimate safety impacts of programs. It assumes that changes in behavioral measures, such as belt use or DWI rates, are enough evaluation, since these interventions have well-documented and widely acknowledged safety effects. Since the ultimate safety effects of driver education are not yet accepted as proven, we cannot ignore safety impacts in discussing evaluations.

The *Overview* is divided into six sections, including this introduction. The following section presents an overview of evaluation as a form of research. Section 3 presents the historical background in evaluations specifically addressed to beginner driver education. Section 4 provides an overview of relevant evaluation techniques, followed by Section 5 on planning and organizing evaluations. Section 6 takes a brief look at the future of driver education evaluation. A reference list and a glossary of evaluation research terms are also included.

## Why Evaluate Driver Education Programs?

Driver education programs seek to teach novice drivers the skills and knowledge necessary to perform as safe and responsible drivers. Formal, organized courses for beginners have long been a popular and convenient means of achieving independent mobility, which is important to both young people and their parents. As a safety measure, driver education “makes sense”—that is, most people think it is effective in preventing crashes. Insurance companies give discounts to driver education graduates. Parents think that it makes their children better, safer drivers.

Although conventional wisdom would support all forms of road safety education, including driver education, more objective evaluation is important. Scientific, systematic evaluation in driver education is needed to guide program development and policy. Evaluation can serve two basic goals in driver education. First, it can help to *improve* programs. Second, evaluation can demonstrate or *prove* (or *disprove*) a program’s impacts. Evaluating program impacts to prove their value is often seen as the first goal of evaluation. This is a very important goal—knowledge about the benefits of a program is needed to guide investment and policy decisions. However, for any program to become as effective as possible, and to continue to improve, one must evaluate what works and what doesn’t. We cannot really meet the second goal without meeting the first one—that is, until programs are developed and improved to be as good as possible, they will not have as great an effect as possible. For this reason, the *Overview* focuses on comprehensive evaluation and looks at a wide range of evaluation goals and methods. Many important questions need to be answered if evaluation is to achieve its potential in helping to improve driver education:

- Do driver education programs meet their learning objectives?
- Do driver education programs enhance or detract from safety?
- Do particular types of driver education programs lead to better results than others?
- Which components of driver education programs work best?
- How can driver education programs be improved?
- How can evaluation be improved to minimize its costs and maximize its ability to help improve driver education?

## Does Your Program Have What it Takes to be Evaluated?

It is important for program administrators to understand what participating in an evaluation requires. First, program information will be needed to enable the planning and operation of an evaluation project. This could include: details of the content and delivery of the program; marketing information; access to student test results; student identifiers; and help with obtaining parental consent for study participation. Some of the required information may be considered proprietary or as “trade secrets.” Evaluators should be sensitive to proprietary information and treat it as confidential. Program information is needed even if the evaluation project is just a basic formative evaluation to see how well the program is operating.

If one is thinking about an evaluation of the crash reduction effects of a program, there are a number of other key issues that a program manager needs to consider.

First, to whom will the program’s students be compared? Lack of an adequate comparison group is probably the most common error in past evaluations. Comparing students from a good program to the general public of the same age is not sufficient—the groups are too different in too many ways to be comparable.

Second, does the program have enough students to undertake an evaluation of crashes? Probably the second most common error has been to try to detect different crash rates with too few students. To reliably find a moderate crash improvement (e.g., 10%) requires a thousand or more students in each of the groups that are to be compared.


Third, are there good reasons to think that the program has a reasonable chance of producing a positive effect? Does the manager have solid evidence that the program is meeting its educational objectives?

Finally, can the manager live with the results? Evaluation research methods and findings should be made public, regardless of whether or not they turn out favorable to the program. Disclosure of results is a basic ethical principle of evaluation.

As with many things that are good for us in the long term, some short-term discomfort may go along with being evaluated. However, if the program manager is committed to building a more effective program, carefully planned, systematic evaluation is essential.

Despite a number of driver education evaluations over the years, most of these questions still lack clear answers. There is not yet much compelling evidence that young people who complete driver education programs drive more safely or have fewer crashes than those who receive less formal driver instruction. Some studies have shown safety benefits associated with driver education, but more studies have reported no such benefits, and many studies have been poorly designed, making demonstration of benefits unlikely or impossible. Some have suggested that driver education is associated with negative safety effects, usually because it allows students to start driving earlier.

Taken together, the safety benefits of beginner driver education are uncertain. Nevertheless, lines have been drawn between those who “believe in” driver education—mostly people involved in program delivery and administration—and those who believe that “driver education doesn’t work”—typically members of the driver research community. For more effective driver education in the future, key people in both camps must work together toward improving evaluation and using evaluation tools to help improve driver education programs.



*We have to consider more specifically what it is about young drivers and their crash risk that we hope education will change.*

## **The Goals and Objectives of Driver Education**

In most other fields of education, a program is considered successful if it meets learning objectives at the end of the course. Driver education is given a tougher mission, more like the case with a public health program. To be considered successful, it is expected to produce improved driving and measurable reductions in crashes. When discussing driver education and how it should be evaluated, we need to keep in mind the specific problems that driver education should help solve. Driver education is not a vaccine that can act directly against crashes. We have to consider more specifically what it is about young drivers and their crash risk that we hope education will change. This is important, because good program evaluation starts with a clear understanding of a program’s goals and objectives.

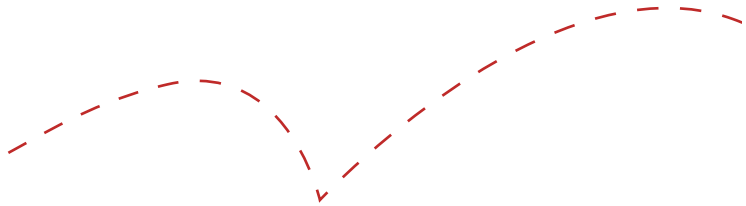
Young, inexperienced drivers have very high crash rates. It is worth considering

what contributes to their excessive risk and how their crash rates change over time. Per mile driven, 16-year-old drivers crash at ten times the rate of experienced adult drivers. This extremely elevated risk declines rapidly over the first few hundred miles of driving. By age 18, the crash rate is about one-third that of 16-year-old drivers. Furthermore, it takes a very long time for risk to level off at mature rates—as much as 10 years. Because crashes, especially serious ones, are the big problem, we have to consider what it is about young, novice drivers that causes such high risk, especially at the very beginning of their driving career.

Inexperience and limited skills certainly contribute to the high risk. Research shows that inexperienced drivers are less able to control attention, to scan the environment effectively, and to detect potential hazards early. The common sense idea that young drivers should just naturally be better drivers because of their quick reflexes is not really correct—they often take more time to make driving decisions than experienced drivers. Inexperienced drivers also perceive less risk in some high-risk situations and perceive more risk in certain lower-risk situations.

Undeveloped or under-developed skills are a big factor contributing to excess crash risk. However, novice drivers also tend to increase their risk through their own actions. For instance, they often tailgate, are overconfident in their abilities, and drive faster than reasonable given their skills, and conditions. Of course, at least some of their risky choices may result from lesser ability to anticipate and perceive risks. Recent research shows that most crashes result from simple mistakes consistent with skill deficiencies rather than from extravagant risk-taking behavior.

In considering crash prevention, the rapid decline of risk after early stages of driving suggests that inexperience might be the most important factor with beginning drivers. Since maturity occurs over a longer time frame than the first few months of driving, this may account for the longer time (5 to 10 years) it takes for risk to decline to levels com-



*While experts may argue over the relative importance of weak skills and risky choices, research results imply that both skill and risky behavior should be addressed in driver education and evaluation.*



parable to adults. While experts may argue over the relative importance of weak skills and risky choices, research results imply that both skill and risky behavior should be addressed in driver education and evaluation.

How driver education is evaluated depends on the objectives we choose for it in other ways as well. Crashes can be measured through drivers' self-report, government records, or insurance records—and these measures can provide quite different results that have important implications. For example, crashes per licensed driver can give quite different results than crashes per capita in the teenage population. Crash rate per miles driven gives different results from population-based measures. These measurement questions are not minor technical issues, since they raise fundamental evaluation questions. Is the proper success criterion for driver education safer mobility or a safer youth population? If concerned primarily with road safety, that is, safer mobility, we would use crash rate per miles driven. If concerned primarily with the overall safety of the youth population, we would use crash rate per teen.

*. . . “it is both possible and valuable to cut through the jargon  
and gain a working knowledge of program evaluation.”*



# Overview of Program Evaluation

This section takes a brief detour into the basic terms and concepts of evaluation research in general. It provides an overview of the basic concepts and terminology of program evaluation for the more detailed discussion of driver education evaluation that follows.

Evaluation is a basic part of everyday life. We are constantly assimilating many different kinds of information from past experiences, current conditions, personal preferences, and our education and goals in order to make judgments and choose actions. While informal, unscientific evaluation is part of everyday experience and can make us wiser and allow us to make better choices over time, professionals in education, the social sciences, and philosophy created a formal discipline called program evaluation or evaluation research. As a result, program evaluation has taken its place among interdisciplinary academic fields, with volumes of research and a variety of jargon, academics, professional consultants, scholarly journals, and conventions. The field is marked by lively disputes over theories, methods, findings, and terminology. Although the organized application of intellectual horsepower to evaluation is desirable, it does tend to make the field hard to understand, and possibly intimidating, for non-specialists. However, it is both possible and valuable to cut through the jargon and gain a working knowledge of program evaluation.

## What is Program Evaluation?

Program evaluation is more formal and organized than everyday evaluation judgments. Formal evaluation allows us to:

- Identify program strengths and weaknesses
- Reflect on and measure progress
- Identify ways to improve programs
- Make decisions about how to change programs

- Collect evidence on program effectiveness and impact
- Assess program efficiency
- Determine or strengthen program accountability
- Share what does and does not work with other program managers, researchers, and evaluators
- Influence policy makers, partners, sponsors, and funding agencies
- Establish a baseline for evaluation excellence

Program evaluation has various formal definitions provided by leading evaluation experts. The following are worth mentioning to help in understanding the wide scope of evaluation:

- The systematic determination of the quality or value of something (Scriven, in Davidson 2004).
- The systematic collection of information about the activities, characteristics and outcomes of programs to make judgments about the program, improve program effectiveness, and/or inform decisions about future programming (Patton 1997).
- An adaptation of social research methods to the task of studying social interventions so that sound judgments can be drawn about the social problems addressed, and the design, implementation, impact, and efficiency of programs that address those problems (Rossi, Lipsey, and Freeman 2004).

Note that the term *systematic* is prominent in these definitions. There is a lot of meaning packed into that term. It distinguishes program evaluation from vague impressions. It means evaluating in a clear, objective, and organized way. It means that evaluation is an essential part of the life cycle of a program. It means that it is just as important to evaluate a program as it is to carefully plan, develop, and deliver the program. And it means that it is essential to know what a program is accomplishing and how it is doing relative to its plan, not just once, but at many points during the program's life cycle.

Systematic also means that new studies build on the findings of earlier ones and answer new questions that are raised by earlier studies. Systematic replication of research is how science and technology progress over time. Isolated, one-shot research rarely leads anywhere. Unfortunately, as we will see later, evaluation in driver education has not been very systematic in the past, but that can change in the future.

Evaluation can take many different forms, and specialists have developed many different kinds of *evaluation models* to fit different organizations and needs. Three groups of models are defined below to illustrate the range of possible evaluation approaches (Trochim 2001). These three are chosen because they are most relevant to evaluating driver education.

- **Scientific-experimental evaluation models:** These models emphasize scientific rigor and objectivity. They include experimental and quasi-experimental evaluations as well as economic-based evaluations, such as cost/benefit analysis. These models give the most objective picture of program effects.
- **Management-oriented systems models:** These models are predominantly used in business and government but can be applied in many organizational settings. They emphasize comprehensive evaluation and organizational effectiveness. This approach is relevant because driver education is a business and cannot be effective unless it is economically viable and well managed.
- **Participant-oriented and social models:** These models emphasize the importance of subjectivity, observation and human interpretation in evaluation. Included are naturalistic and qualitative evaluation approaches, as well as client-centered, stakeholder and consumer-oriented approaches. These approaches are relevant because driver education is part of community life and cannot be effective unless we understand the needs and perceptions of students, parents, and other stakeholders in teen safety and mobility.

Each type of evaluation model brings unique and valuable perspectives to the evaluation process. Some perspectives are more “scientific”—objective, and quantitative. Others are more qualitative. While not all models are useful in every evaluation, an optimal evaluation framework should integrate the relevant aspects of all three categories. Combining different approaches can provide a broader and more detailed picture of a program and how well it works. Most important, having the big-picture approach can help us understand *why* it works, which is often needed to know how to make improvements.



## Defining the Two Basic Functions of Evaluation

Evaluation research can serve two basic purposes or functions in the development of any program, including driver education. Evaluations aimed at these different purposes are given different names. First are the more basic levels of evaluation, which are intended to improve program structure, content, and delivery. This is called *formative evaluation*. This type of evaluation research helps “form” the program. It does this by studying the *products and processes* inside the program, through which it is expected to meet its goals. It asks questions like, “Is the program operating consistently according to plan? Is it meeting its teaching objectives? How can it be improved?” It is the quiet, inward looking member of the evaluation family, but it is no less important. It should be an ongoing part of program development and quality management.

Evaluation can also demonstrate or prove the results, effects, or benefits of a program. It usually requires that the program has gone through the formative evaluation processes and improvements have been made. This more public and colorful member of the evaluation family relies on more demanding, highly technical evaluation research methods. Evaluation specialists call this *summative evaluation*. This term is easy to remember if we consider that it indicates the “sum” of the program’s effects, or that it summarizes the program’s effects or results. Summative evaluation looks at the outside effects of the program, that is, at the *outcomes and impacts* that it is supposed to produce. These evaluations ask questions like, “Does the program make people perform better?” and “Does it make them safer?”

This is why professional evaluators place a very high value on comprehensive evaluation, gained through multiple perspectives. They believe that an evaluation should assess all aspects of a program’s worth or value. Studies that address only selected questions and methods are considered *quasi-evaluations*. Comprehensive evaluations are most desirable whenever feasible, but they need not happen all at once. Even small-scale evaluations, limited by resource constraints, should be undertaken as a start, rather than opting for no evaluation. The key in such cases is to keep going and build up a bigger, more detailed picture over time.

## Demystifying Evaluation Concepts and Terms

It is important to approach the evaluation process with an understanding of some of the most frequently used evaluation research terms and concepts. The most important definitions are presented below. For a more extensive list, see the Glossary of Terms in Appendix A.

In program evaluation, a *program* is defined as planned activities that are intended to achieve specific outcomes for specific client groups. Programs are typically ongoing, as opposed to projects, which have a defined end point.

Program evaluation first requires understanding how the activities of the program are expected to achieve its goals and objectives. Specifying how the program is supposed to work in concrete terms is called *program theory*. Program theory is often displayed in a graphic form, called a *logic model*. Program theory and logic models describe the logical steps that are supposed to take place—how the program’s resources and activities are supposed to change specific things in order to achieve the program’s goals and objectives.

When first thinking about evaluation planning, it is important to identify the program’s *stakeholders*. These are the people who have an interest in the program and its effects. They are involved in or affected by the program, and may also be interested in or affected by an evaluation. Stakeholders can be individuals or groups including program staff, participants, community members and organizations, decision-makers, sponsors, and funders. A special subgroup of stakeholders are the program’s *target groups*, the people who are the clients, customers, users, or supposed beneficiaries of the program.

Many evaluations involve some sort of quantitative measurement and express the results as numerical data. It is important to keep a few critical concepts in mind when considering measurement, such as testing or survey questionnaires.

*Reliability* is the extent to which the measure is consistent. Questions such as, “Are we measuring consistently?” and “How stable is our measure?” involve reliability. Several types of reliability are important to evaluation. *Inter-rater reliability* is the degree to which a test gives similar results for more than one tester, such as different driver testers. *Internal consistency* is the degree to which different questions or other parts of a test consistently measure the same attribute. *Test-retest reliability* is the degree to which the measure produces

consistent results over several administrations assessing the same attribute of an individual. In other words, it answers the question “What happens if we give the same test to the same individual on two different days?” Reliability is usually expressed as a correlation number between the scores on different measures. These *reliability coefficients* can range between 0 and 1, with 1 being perfect reliability.

*Validity* refers to the soundness of the use and interpretation of a measure. It questions whether we are actually measuring what we are supposed to be measuring. Validity of a measure is also expressed as a *correlation coefficient*. In this case it shows the strength of the relationship between our test measure and some other kind of score that we accept as a true indicator of what we are trying to measure. An example of a validity check for a driving test would be the correlation between test scores and a measure of actual driving, such as driving record. There are also different types of validity. *Predictive validity* is how strongly a measure later predicts another score on some criterion. For example, if a driving test strongly correlated with drivers’ later driving records, the test would have good predictive validity.

A *variable* is an indicator assumed to represent some underlying concept. For instance, motor vehicle department crash records could be seen as a variable representing the concept of driver competence. *Independent variables* are the things that can be manipulated or selected in a research study, such as the kind of training that a driver receives or the number of hours of practice driving. *Dependent variables* are used to measure results, such as test scores or crash rates. *Confounding variables* are extraneous factors that could compromise the study by influencing the dependent variable. For example, teens who took driver education and those who did not take it may differ in motivation to begin driving as soon as possible and willingness to take risks. These differences could act as confounders and make it impossible to tell if taking or not taking driver education made a difference in acquisition of driving skills or involvement in crashes.

Evaluations can measure a program’s *intermediate outcomes*, *ultimate impacts*, or both. Intermediate outcomes are the knowledge, skills, attitudes, intentions, or values of the students that may (or may not) have been affected by the driver education program. Ultimate impacts are the measures such as crash and injury rates of students after they have become licensed.




Evaluations can also use research methods that are *qualitative*, *quantitative* or both. Qualitative measures involve words, and reflect in-depth meaning and understanding of beliefs, attitudes and reported behaviors from small groups of people. Interviews and focus groups are methods used to collect qualitative data. Quantitative methods involve numbers, and represent broader data from larger groups of people (called samples) that tell us “how many people think what,” often with the goal of being able to generalize to larger groups (populations). Questionnaires with multiple choice and true-false questions and systematic examinations of driving records are examples of quantitative methods.

When discussing quantitative results you will often see the term *statistical significance*, in the sense that some result is found to be statistically significant. This simply means that the result probably did not occur just by chance. For instance, a finding of a statistically significant relationship between gender and crash rates would mean that the finding is strong enough that there is only a small chance (usually 5%) that the observed result (for example, female teens have lower crash rates) could have occurred solely due to random chance. If the result is not significant, the possibility that it occurred just by chance cannot be rejected.

Statistical significance does not necessarily mean that the result is important. Sometimes a very small effect (e.g., a one percent difference in traffic violations) can be statistically significant, even if it is too small to be of any practical importance. The opposite can also occur—that an effect that looks fairly large in percentage terms is not statistically significant. This happens because statistical significance depends on *sample size*, that is, the number of observations being studied, as well as the observed results.

Many evaluations have had sample sizes too small to find statistical significance, even when there were fairly large differences in crash rates. For a moderate-sized difference in crash rates (say 10%) to be statistically significant, studies would need a few thousand participants, rather than the few hundred that have sometimes been studied. In addition to the size of difference that we want to be able to detect, sample size requirements also depend on the baseline crash rate. The larger the difference that you are willing to accept



**I**nterviews and focus groups are methods used to collect qualitative data. Quantitative methods involve numbers, and represent broader data from larger groups...

(or possibly are overlooking) and the higher the base crash rate, the smaller the sample size required. For example, if you are concerned about detecting a 5% decrease in crash rate, you need a much larger sample than if you are willing to concede that nothing smaller than a 15% decrease is important.

## Evaluation Standards

Based on experience in education and the other fields where evaluation has been used extensively, standards have been developed to help avoid evaluation errors. They are used throughout an evaluation as benchmarks against which to check its quality. The standards presented in *Evaluating Driver Education Programs: Comprehensive Guidelines*, were developed by the Joint Committee on Standards for Educational Evaluation (1994). They have been approved by the American National Standards Institute (ANSI) and have been widely adopted in many fields, including education, public health, injury prevention, and human services. The Joint Committee grouped the standards into four categories, each with a number of specific criteria. The four categories of standards are:

1. **Utility**—The evaluation serves the information needs of the intended users.
2. **Feasibility**—The evaluation is realistic, prudent, diplomatic, and frugal.
3. **Propriety**—The evaluation is conducted legally, ethically, and with regard for the welfare of those involved and affected by its results.
4. **Accuracy**—The evaluation reveals technically adequate information about the worth or merit of the program being evaluated.

The best evaluation studies in driver education have been well designed and they probably meet most of the quality standards, but there have also been some very poor evaluations. It is important to keep in mind that there are important technical and ethical matters that must be addressed, and the evaluation standards provide a good basis for addressing these issues.



# Driver Education Evaluation—Past and Present

This section provides an overview of the existing published evaluation research studies. The *Comprehensive Guidelines* contains a more complete discussion of the conclusions of recent reviews of evaluations and of some key individual evaluations. The aims of the review are to provide a basic understanding of the methods, findings, and limitations of past evaluations. This understanding is important for improving the future of driver education policy, program planning, and program management.

## Historical Overview of Driver Education Evaluation

Compared to the public health and education fields, beginner driver education has seen relatively few evaluations. In driver education, evaluation has usually meant attempting to assess short-term, direct safety impact. This safety impact has typically been defined as total reported crashes subsequent to taking the program. In most cases, graduates of driver education have been compared to new drivers who learned to drive in other ways. A number of evaluations have compared different forms of formal driver education. Some studies compared driver education students to a “control group” of new drivers who learned to drive through home-based instruction or some other form of instruction.

The largest and most influential driver education evaluation is known as the “DeKalb” study—named after DeKalb County in Georgia, where the study took place. The study involved randomly assigning 16,000 high school student volunteers to three groups—special intensive training, minimal training, or no formal driver education. The results did not show any dramatic, long-term benefit associated with the special course.

Reactions to the results of the DeKalb experiment had profound effects on driver education. In the U.S., driver education market penetration peaked in the early 1980s with about 80% of new drivers being formally trained. After that, however, many high school driver education programs were dropped. For instance, New Jersey schools offering driver education dropped from 96% to 40% between 1976 and 1986 (Simpson 1996). It has not

been clearly demonstrated whether the DeKalb results in effect caused the decline in U.S. high school driver education, or if it served as support for budget cutting. Regardless of how things went in the 1980s, difficult policy choices are presented when a hoped for benefit of a critical program does not appear as expected. Should resources be shifted to other kinds of programs? Or should additional resources be employed to improve driver education to make it effective in the ways expected?

Other studies besides DeKalb have also failed to detect a direct measurable change in crash rates of graduates compared to others. Although there have also been positive findings, many members of the safety research community have come to believe that “driver education does not work,” at least in reducing crashes.

This conclusion raises questions as to how such a counterintuitive situation might be possible. However, given the limited scope of beginner training, as well as its position at the very start of a long learning curve, it is possible that the driver education experience can be overshadowed by other experiences, overconfidence, earlier licensure and increased exposure to risk, and relaxed parental supervision. Since so much of drivers’ learning takes place after licensing, it may be that potentially beneficial effects of traditional driver education are offset by other influences. It may also be possible, as researchers have suggested, that driver education in the past has not provided the best possible content in the best ways (e.g., Mayhew and Simpson 1997).

Unfortunately, as will be discussed in more detail in the next section, driver education evaluation has also been rather unsystematic as well as limited in quantity. Even the randomized controlled trial (RCT) experiments that have been conducted have suffered from methodological problems that have made their results less than definitive.

## **Reviews of Driver Education Evaluations**

This section briefly discusses the conclusions of the most significant recent reviews of earlier driver education evaluations. The more recent individual quantitative evaluations, along with selected older evaluations, are discussed in the next section.

The companion report, *Evaluating Driver Education Programs: Comprehensive Guidelines*

presents a new review of published driver education evaluation research, including earlier reviews of evaluations. Its main focus is to identify the strengths and weaknesses of the highly diverse driver education evaluation literature. Unlike earlier reviews, the main purpose is not to determine whether driver education has “worked” in the past. Rather the new review is more focused on the evaluations themselves, with the intent of determining how evaluation research can be improved to help driver education work better in the future.

In the last decade or so, there have been a number of broad reviews of evaluation of driver education, usually in conjunction with other forms of driver instruction or graduated licensing (GDL) (Lonerio et al. 1994, 1995; Mayhew and Simpson 1997; Woolley 2000; Christie 2001; Mayhew and Simpson 2002; Engström et al. 2003; Siegrist 2003; Smiley, Lonerio, and Chipman 2004).

Early evaluations used totally uncontrolled comparisons between driver education graduates and others. These early, uncontrolled comparisons tended to show that driver education graduates crashed less than other new drivers. Nichols (2003) summarized the findings of the early evaluations as follows:

Although there were a few early studies which reported negative results, the majority of studies conducted at the time suggested that: (1) while the effects may be short-lived, driver education students had fewer accidents and violations than non-students; (2) complete courses involving both classroom and behind-the-wheel training were more effective than classroom-only courses; and (3) [High School Driver Education] was more effective in reducing accidents and violations than either parent training or commercial driver training. (p. 20)

The early studies made no effort to control for the ways in which driver education graduates were different from comparison groups other than the type of training that each group had received. That is, the studies failed to control for the effects of extraneous, confounding factors, so these quasi-experiments were not considered credible. Some subsequent evaluations were designed as full experiments and attempted to control for extraneous differences between driver education and comparison groups by random assignment to a training group or a control group (comprising informally trained novice drivers or those who took a different course). There is, of course, no such thing as an untrained new driver—everybody

has to learn some way to pass the licensing tests. As a result, there is no possible comparison of driver education versus no treatment, as might be the case in clinical trials, psychology lab experiments, or even in many other types of education program evaluation, such as driver improvement courses. Comparisons are always between different methods of learning to pass the same test.

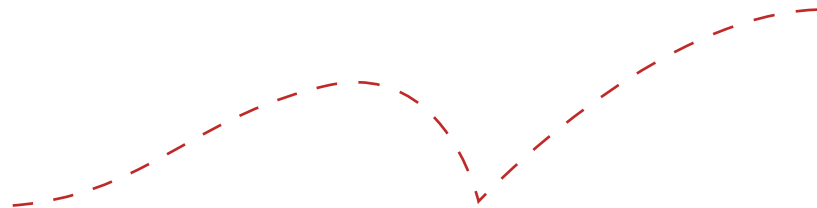
Graduated licensing, which typically requires new drivers to hold a learner's permit for a specific length of time, delays independent driving and restricts driving in some risky situations, such as late at night or with teen passengers. It has been the principal initiative to reduce young driver crashes in recent years. Evidence shows it is effective in reducing beginner driver crashes. At the Traffic Injury Research Foundation, Mayhew and colleagues (1997; 2002) performed a detailed review of evaluations of beginner driver education, in the wider context of graduated licensing. This review provided a sort of scorecard of evaluation findings, and the authors concluded that:

- Seven evaluation studies indicated a positive safety effect;
- Sixteen showed no effect; and
- Seven others suggested a negative safety effect.

Two additional studies that showed positive effects were not included. One was an econometric modeling study of driver education over 47 U.S. states (Levy 1988; 1990). None of the earlier reviews included a California experimental study (Dreyer and Janke 1979), which did find a positive effect on drivers' crash records.

Two recent systematic reviews covered only a small selection of evaluations. Vernick

*Graduated licensing, which typically requires new drivers to hold a learner's permit for a specific length of time, delays independent driving and restricts driving in some risky situations, such as late at night or with teen passengers. Evidence shows it is effective in reducing beginner driver crashes.*



and colleagues (1999) reviewed nine evaluations that met their methodological criteria. The intent of the review was broader than most, aimed at finding: 1) if driver education graduates were less likely to crash or more likely to become licensed to drive; and 2) whether driver education had broader public health effects in lowering community rates of crashes. All but one of the nine studies addressed U.S. high school programs. The reviewers concluded that no study that met their design criteria showed a “significant individual or community-level beneficial effect of driver education for high school-aged students” (p. 44). No explanation was offered for disregarding the findings of a significant beneficial effect on fatal crashes by Levy (1988; 1990), which was included among the studies reviewed.

Using an even narrower selection basis for a systematic review, Roberts and Kwan (2004) reviewed just three RCT experimental evaluations, all from the early 1980s. They also concluded that the evidence did not support safety impacts of driver education.

Christie (2001) published a detailed review of evaluations of various forms of driver instruction, including beginner driver education. He reviewed the same studies as Mayhew and Simpson (1997), as well as later ones. Apparently less impressed than Mayhew and Simpson with the limited positive impacts of driver education found in the literature, Christie concluded that no evidence shows beneficial effects of beginner driver education. He reiterated the view that driver education is harmful because it induces earlier licensing.

Another review by Woolley (2000) concluded that there is no conclusive link between skills-based training and crash involvement. Motivation and propensity to take risks are more important than any type of skills-based training, and driver education should be developed to address these critical factors.

Elvik and Vaa (2004) conducted a meta-analysis of 16 driver education evaluations from around the world. Meta-analysis is a technical review approach that statistically combines the findings of individual studies. The combined data of 16 studies found driver education graduates having 1.9% fewer crashes per driver. Results per kilometer driven found a 4% lower crash rate for graduates. When the combined results were limited to experimental studies, a different picture emerged. There was no difference in crashes per driver. Per kilometer, driver education graduates had 11% more crashes. The authors conclude that the combined evaluation results do not indicate that driver education reduces crashes.

Elvik and Vaa also briefly examined possible explanations for the generally disappointing findings. The first is that the evaluation research is too poor to detect the real effects of driver education. They refute this by indicating that the research overall is actually somewhat better than most road safety evaluation research. Also, combining the findings of individual studies overcomes the common problem of too-small sample sizes in some evaluations. The explanation favored by the authors is behavioral adaptation—less-skilled drivers taking more care and better-skilled drivers taking less care. Interesting questions arise if one believes behavioral adaptation is the reason behind the failure of driver education. Aside from building skills and knowledge, what would even the best driver education have to do to help overcome such motivational tendencies? Can evaluation more clearly demonstrate how driver education works or fails to work on crash rates?

## Individual Evaluations of Driver Education Programs

The great majority of driver education programs have never been formally evaluated, and most existing evaluations are severely limited in scope, power, and scientific rigor. In this section, individual evaluations of driver education programs are briefly described to give a flavor of the range of studies. A more complete discussion of the individual evaluations can be found in the *Comprehensive Guidelines* review.

Different research designs have been used in driver education evaluations. The evaluation studies reviewed are presented in brief tabular form, and represent fairly recent work in the field as well as some older studies of special importance. They are divided into three basic types, according to the basic research design that was used.

First are the *experimental studies*. These studies involve random assignment of drivers to various training conditions and comparison of subsequent crash rates and other measures. This type of study is similar to *randomized controlled trials* (RCTs) that are used for drug testing and other medical treatment research. They are considered by many researchers to be the “gold standard” of research designs. The random assignment to different education treatments means that all other factors that might cause a difference between the groups of drivers, except the training, are ruled out. The potential confounding variables are probably, but not certainly, “averaged out” by chance. Recall that possible *confounding variables*, which could compromise a driving records comparison between driver education groups, include such factors as area of residence, family income, and attitudes.



<b>Experimental Studies: Random assignment of drivers to training conditions</b>			
<b>Reference</b>	<b>Design</b>	<b>Results</b>	<b>Methodological Strengths/Limitations</b>
Dreyer and Janke 1979 <i>California</i>	<ul style="list-style-type: none"> <li>• 2057 students randomly assigned to two training conditions (off-road range training vs. on-road)</li> </ul>	<ul style="list-style-type: none"> <li>• Those receiving range practice had fewer recorded crashes, but tests scores were no different</li> </ul>	<ul style="list-style-type: none"> <li>• Randomized controlled trial</li> <li>• Intermediate measures</li> <li>• No follow-up survey for exposure and behavioral measures</li> </ul>
Ray et al. 1980 Stock et al. 1983 <i>DeKalb County, Georgia</i>	<ul style="list-style-type: none"> <li>• Intensive, minimal, and no driver education groups</li> <li>• About 6,000 students randomly assigned to each group</li> </ul>	<ul style="list-style-type: none"> <li>• Intensive training (SPC) drivers had better skills and fewer crashes during first 6 months of driving, but not beyond</li> </ul>	<ul style="list-style-type: none"> <li>• Comprehensive randomized controlled trial</li> <li>• Long follow-up—6 years</li> <li>• Formative evaluations and intermediate outcome measures</li> </ul>
Wynne-Jones and Hurst 1984 <i>New Zealand</i>	<ul style="list-style-type: none"> <li>• 788 students, 561 received course, 227 family or friend taught</li> <li>• Random assignment</li> </ul>	<ul style="list-style-type: none"> <li>• No reduction in collisions for driver education group</li> </ul>	<ul style="list-style-type: none"> <li>• Small control group</li> <li>• No formative evaluation or intermediate outcome measures</li> </ul>
Gregersen 1994 <i>Sweden</i>	<ul style="list-style-type: none"> <li>• 850 students received driver education course compared to controls</li> <li>• Random assignment</li> </ul>	<ul style="list-style-type: none"> <li>• Driver education group crashes significantly worse in first year, significantly better in second year</li> </ul>	<ul style="list-style-type: none"> <li>• Longer follow-up—2 years</li> <li>• Reasonable sample size</li> </ul>
Masten and Chapman 2003; 2004 <i>California</i>	<ul style="list-style-type: none"> <li>• 1,300 students randomly assigned to one of four instructional settings</li> </ul>	<ul style="list-style-type: none"> <li>• Home-based methods better for knowledge and attitude, except classroom better for knowledge test</li> </ul>	<ul style="list-style-type: none"> <li>• Sample size adequate</li> <li>• Well planned and controlled</li> <li>• Intermediate measures only</li> </ul>

*Quasi-experimental studies* compared differences between self-selected driver education students and those who learned to drive in other ways. The quality of these studies varied greatly. The earliest evaluations of this type made no effort to compensate for *confounding variables*. Most later studies used statistical methods to correct for some of the extraneous, confounding variables.

<b>Quasi-experimental Follow-Up Studies: Self-selected driver education students and those who learn to drive in other ways</b>			
<b>Reference</b>	<b>Design</b>	<b>Results</b>	<b>Methodological Strengths/Limitations</b>
Forsyth et al. 1995 <i>United Kingdom</i>	<ul style="list-style-type: none"> <li>Survey of 15,000 new drivers</li> </ul>	<ul style="list-style-type: none"> <li>Longer time learning to drive associated with fewer crashes for males</li> <li>More driving instruction was associated with more crashes</li> </ul>	<ul style="list-style-type: none"> <li>Several follow-ups over time</li> <li>Self-selection bias</li> <li>Self-reported data only</li> </ul>
Haworth et al. 2000 <i>Australia</i>	<ul style="list-style-type: none"> <li>Self-report crash effects for in-car training effects</li> </ul>	<ul style="list-style-type: none"> <li>Substantial crash differences in favor of in-car training condition, not statistically significant</li> </ul>	<ul style="list-style-type: none"> <li>Sample size too small</li> </ul>
McKenna et al. 2000 <i>Pennsylvania</i>	<ul style="list-style-type: none"> <li>Survey and crash records</li> <li>Random sampling for survey</li> </ul>	<ul style="list-style-type: none"> <li>Driver education not associated with lower crashes or convictions</li> </ul>	<ul style="list-style-type: none"> <li>Multi-variate statistical analysis used to control for confounding variables</li> <li>SES missing from control variables</li> </ul>
Lonero et al. 2005 <i>Manitoba</i>	<ul style="list-style-type: none"> <li>Survey and crash records</li> <li>Random sampling for survey</li> </ul>	<ul style="list-style-type: none"> <li>Driver education not associated with lower crashes or convictions</li> </ul>	<ul style="list-style-type: none"> <li>Multi-variate statistical analysis used to control for confounding variables</li> </ul>
Wiggins 2005 <i>British Columbia</i>	<ul style="list-style-type: none"> <li>Cohort record study</li> <li>Case control study with survey and records</li> </ul>	<ul style="list-style-type: none"> <li>New graduated license holders who took driver education had 26% more crashes</li> </ul>	<ul style="list-style-type: none"> <li>Multi-variate statistical analysis used to control for confounding variables</li> </ul>
Zhao et al. 2006 <i>Ontario</i>	<ul style="list-style-type: none"> <li>Self-report survey of high school students</li> </ul>	<ul style="list-style-type: none"> <li>Driver education associated with fewer crashes for learner license holders</li> </ul>	<ul style="list-style-type: none"> <li>Multi-variate statistical analysis used to control for confounding variables</li> </ul>

*Ecological studies* have considered impacts on crashes in large jurisdictions (e.g., whole states or countries) following changes in requirements or support for formal driver education. Ecological studies can also be used to compare results in states with different types of driver education. This usually involves complex statistical modeling. The statistical models are used to compensate for *confounding variables*, such as the other differences (besides driver education) that might exist among different states, or within a single state at one time compared to the same state at a different time.

<b>Ecological Studies:</b> Measure impacts on crashes following jurisdictional changes in requirements or support for formal driver education			
<b>Reference</b>	<b>Design</b>	<b>Results</b>	<b>Methodological Strengths/Limitations</b>
Robertson and Zador 1978 <i>27 States USA</i>	<ul style="list-style-type: none"> <li>Modeling study of driver education and fatal crash rates</li> </ul>	<ul style="list-style-type: none"> <li>No relation between proportion taking driver education and fatality rates</li> </ul>	<ul style="list-style-type: none"> <li>Not program specific</li> </ul>
Robertson 1980 <i>Connecticut</i>	<ul style="list-style-type: none"> <li>School boards with and without driver education</li> </ul>	<ul style="list-style-type: none"> <li>For school boards without driver education, total licensing and crashes of 16- and 17-year-olds decreased by 10 - 15%</li> </ul>	<ul style="list-style-type: none"> <li>Not enough data analysis presented</li> </ul>
Potvin et al. 1988 <i>Quebec</i>	<ul style="list-style-type: none"> <li>Mandatory driver education introduced in Quebec for all (formerly just 16- to 17-year-olds)</li> </ul>	<ul style="list-style-type: none"> <li>Increased number of young driver crashes due to increased number of licensed females aged 16 - 17</li> </ul>	<ul style="list-style-type: none"> <li>Large sample size</li> <li>Different time frames for treatment and control groups</li> </ul>
Levy 1988; 1990 <i>47 States USA</i>	<ul style="list-style-type: none"> <li>Large-scale modeling study of effects of mandatory driver education</li> </ul>	<ul style="list-style-type: none"> <li>Small but significant beneficial effect on fatal crashes</li> </ul>	<ul style="list-style-type: none"> <li>Not program specific</li> </ul>
Carstensen 1994; 2002 <i>Denmark</i>	<ul style="list-style-type: none"> <li>Mandatory driver education, new curriculum</li> </ul>	<ul style="list-style-type: none"> <li>Reduced crashes</li> </ul>	<ul style="list-style-type: none"> <li>Large sample size</li> <li>No control of confounding variables</li> </ul>

## Limitations of Past Driver Education Evaluations

Weaknesses in past evaluations are worth looking at to see what needs to be improved in the future. This is the badly needed part of the systematic advance in evaluation research in this field. Key areas for improvement that have been identified in driver education evaluation include:

*Weak program theory:* Theory, in the sense used here, means the logic that justifies thinking a program should meet its goals—that is, why we think it should work. There has been little evaluation or development of the theory underlying various driver education programs.

*Lack of formative evaluation:* Little formative evaluation of driver education programs probably means they are not as good as they could be. Courses vary greatly in quality. Furthermore, there has been limited evaluation of program differences. How well driver education students achieve, retain, and use desired skills and knowledge is still unclear.

*Methodological weakness:* Problems of scope, design, sampling, and confounding comparisons are common. They limit conclusions about the ultimate value of driver education at present and how its impact might be improved in the future.

*Lack of systematic follow-up:* Most of the past evaluations have been one-shot efforts that did not build from earlier work or try to answer questions raised by earlier studies.

Beginner driver education is challenging to evaluate in terms of safety impacts. Suitable comparison groups are hard to establish. Many earlier evaluations compared groups of young drivers who received different forms of driver education but also differed in other ways that might affect their driving record and other results. As explained earlier, these extraneous or confounding factors could include location of residence, income, or other important socioeconomic factors. Even when comparable treatment and control groups can be established, they are hard to maintain over time. People may not complete the course to which they were assigned, or they may move or decide that they do not want to participate in follow-up surveys or interviews. When different groups involved in a study have different dropout rates, this can seriously bias the results.

One of the most common errors in driver education evaluations is the failure to use large enough groups. In evaluations of crash rates, very large numbers of cases are needed. Many of the experimental studies used samples that were too small, making it unlikely that the statistics would detect a difference even if there had been an important difference in crash records (Engström et al. 2003). A recent Australian quasi-experiment observed substantial crash differences between training conditions, but, because the numbers of drivers were so small, one could not conclude with confidence that the differences were the result of anything other than chance (Haworth, Kowadlo, and Tingvall 2000).

*One of the most common errors in driver education evaluations is the failure to use large enough groups.*

Other study design problems have also undermined the usefulness of past evaluations of driver education. Most evaluations simply looked at crash rates, failing to look at intermediate measures of student outcomes. As a result, the ways to improve driver education programs have not been clearly identified. By intermediate measures, we mean students' knowledge, skills, and attitudes that have been affected by the program. If we do not know whether the program has had the desired direct effects on the students, it is hard to tell what parts of a program work or fail to work. If we do not know what and how well the students have learned, how can we tell why there was an effect (or no effect) on crashes? Intermediate outcome measures are necessary to see what is actually changing, and what is not.

If we fail to find out how and how much students drive after licensing, we again cannot understand the effects (or lack of effects) of a program. In driver education evaluation, "exposure to risk" involves the amount and type of driving. Differences in exposure are too often ignored in driver education evaluation. Types of exposure, such as time of day, presence of passengers, geographic areas, and different trip purposes also represent different levels of collision risk, especially for young drivers.

Since different methods of learning to drive affect when the beginner chooses to be first licensed, exposure information is an important part of any attempt to evaluate driver education programs. The possibility that exposure could be an important source of

confounding or bias in evaluation results has often been overlooked. Survey tracking of drivers' behavior during the follow-up period is necessary to see what is actually changing, and what is not.

In summary, methodological weaknesses have plagued many evaluations. Most evaluations have neglected to assess learning outcomes or have used sample sizes that are too small to reliably detect moderate program effects. Many have used comparisons between groups that were not really comparable. Such common inadequacies have led to different interpretations and controversy over the meaning of past evaluations. A more systematic future approach should correct these problems and build on the strengths of earlier evaluations. Science and other forms of knowledge build over time by correcting weaknesses and answering new questions raised by earlier research. There is no reason why this should not be applied to the science and knowledge of driver education as well.



# Techniques of Evaluating Beginner Driver Education

Driving and learning to drive are such basic parts of life that most people take them for granted. Driving, like evaluation, has also become the subject of many research studies, consultants, scholarly journals, conventions, and lively disputes. Of the large volume of research activity on drivers and driving, only a very small portion has been addressed to evaluating driver education.

Driver education can be seen as serving many educational purposes—life skills training, citizenship, economic opportunity, social facilitation, and safety. For understandable reasons, the relatively small volume of research and evaluation has been addressed primarily to safety. The tragic results of teen crashes make safety such a dominant concern that it has been hard to look carefully at other aspects and outcomes. As a result, past driver education program evaluations have left many basic questions partially or completely unanswered.

Driver education programs differ widely, and one of the first things we need to do is to recognize the potential importance of these differences. It is unlikely that all approaches to driver education can or will be equally effective. Evaluation objectives and approaches should be different depending on the size and quality of the program, as well as on previous research and development work. There are numerous research, technical, logistical, and operational questions and details that need to be refined for effective evaluation.

## **Types and Levels of Evaluation**

Most evaluations of beginner driver education suggest it does not meet expectations in preventing crashes. We know that many other kinds of health promotion and injury prevention programs also do not work well, or that they work well only after they have been refined through years of systematic research and development. We have some guiding principles for programs that try to alter behavior for health and safety reasons, based on experience in public health, health promotion, industrial safety, and, to some extent, road

safety. Ongoing evaluation and continuous improvement are most important among the principles of successful programs.

A broad, comprehensive evaluation approach for driver education is one that realistically incorporates:

- The full scope of driver education program logic, context, products, and processes
- The entire range of outcomes and impacts
- All relevant delivery organizations, including businesses and governments
- All concerned stakeholders, including consumers, providers, insurers and regulators
- All appropriate evaluation models, methods and measures

*Evaluating Driver Education Programs: Comprehensive Guidelines* provides a framework that addresses the evaluation needs of all the key components of driver education, including:

- Theory—theoretical and logical bases of the program
- Context—political, economic and social environments that influence a program
- Standards—principles and regulations that govern a program
- Products—content of instructional materials
- Processes—educational delivery methods and management operation
- Outcomes—direct educational effects of the program on students, such as increased knowledge and skills
- Impacts—the social consequences of the program, such as crash reductions or increases

A brief summary of this framework follows to help readers understand the importance and utility of an organized structure for driver education program evaluation.




The two main purposes or missions of evaluation, formative and summative evaluation, are included in the framework. Within both evaluation types, many specific evaluation targets or aspects of the program can be evaluated. And there are numerous related success indicators or measures of effectiveness. Targets range from user needs and stakeholder expectations, to management processes, instructional products and processes, learning and behavioral outcomes, and safety impacts.

The framework also provides a wide range of data-gathering and research methods that are related to the evaluation targets. The list of possible evaluation methods is lengthy. Examples include program content selection (e.g., pilot testing and content analysis), qualitative research such as focus groups and interviews, standardization (e.g., benchmarking, certification and auditing), instrumented vehicle observation, questionnaire surveys, record studies and modeling, ecological studies, longitudinal studies, quasi-experiments, and randomized controlled experimental trials.

A critical step is to determine the goals for evaluation and the level of resources, time and effort to be committed. Using information on the evaluation's purpose, goals, and objectives as well as the available financial and human resources, we can identify a level of evaluation that provides the best fit between these feasibility criteria and the evaluation to be implemented. Four levels of evaluation are used to accommodate different evaluation goals and the full range of driver education programs.

Level 1 consists of relatively simple, formative evaluation activities. These can include describing the program, setting program goals and objectives, and identifying evaluation objectives, questions and targets for improving the program. Benchmarking the program against industry standards or surveying customers to determine satisfaction levels can also be undertaken. Examining instructor qualifications, the uniformity of instructional delivery, and other operational matters can be included. A well-managed local program provider or school authority can perform this level of evaluation, consisting primarily of formative evaluation and qualitative study methods.



**A** *critical step is to determine the goals for evaluation and the level of resources, time and effort to be committed.*

Level 2 extends the evaluation methods used for improving a program. It includes the activities of Level 1 and adds limited quantitative assessment of materials, methods, and student knowledge and skill outcomes. Level 2 can be considered by teams who are prepared to undertake a more intensive evaluation and have the ability to manage basic quantitative data, such as tracking test scores over time.

Level 3 expands the focus to outcome and impact evaluations, as well as more complex quantitative methods and benchmarking of organization quality. Activities such as audit compliance, quality management certification, testing and observation to evaluate student skill and knowledge outcomes, and quasi-experiments to assess safety and mobility impacts can be considered. This level is more likely to be undertaken by organizations with substantial resources, such as major program providers, large materials suppliers, industry associations, and state or provincial governments. It requires basic technical skills in research and evaluation methods.

Level 4 involves comprehensive outcome and impact evaluations using relatively advanced and specialized statistical methods and measurement techniques. It can include special knowledge tests and observation methods, instrumented vehicles and simulators to evaluate student skill and outcomes, large-scale record studies for safety impact evaluation, and socioeconomic analyses. This is the broadest level of evaluation and requires substantial resources and technical expertise. These are most likely available to national governments and larger state or provincial governments, or large research organizations at universities or other major institutions.

The chart on the next page provides examples of activities that can be included in each level.

These levels assist potential evaluators to select and plan appropriate evaluation activities within their resource and evaluation capabilities. However, it is also important to look beyond immediate evaluation capabilities, and establish longer-term evaluation goals. Evaluation should become a progressive and integral part of program implementation and improvement. Remember, good evaluation is systematic research, and just doing it once is not enough.

Evaluation Activities	Evaluation Level			
	1	2	3	4
Take steps to build program evaluation and development capability	X	X		
Describe the program structure and environment	X	X		
Build a logic model for the program	X	X		
Benchmark the curriculum structure and materials to industry standards	X	X	X	
Evaluate customer satisfaction levels	X	X	X	X
Evaluate student reactions to materials and instruction methods	X	X	X	X
Evaluate student knowledge outcomes and skills through testing		X	X	X
Commit to continuous improvement through evaluation and development		X	X	X
Audit compliance with standards and regulations			X	X
Certify quality management			X	X
Evaluate student skill and knowledge outcomes through testing and observation			X	X
Evaluate safety and mobility impacts using quasi-experiments			X	X
Assess evaluation and development activities			X	X
Evaluate student skill and knowledge outcomes using instrumented vehicles, simulators				X
Evaluate driver education context and policy approaches				X
Evaluate safety impacts through ecological studies and experiments				X
Evaluate socioeconomic impacts through cost/benefit analyses				X

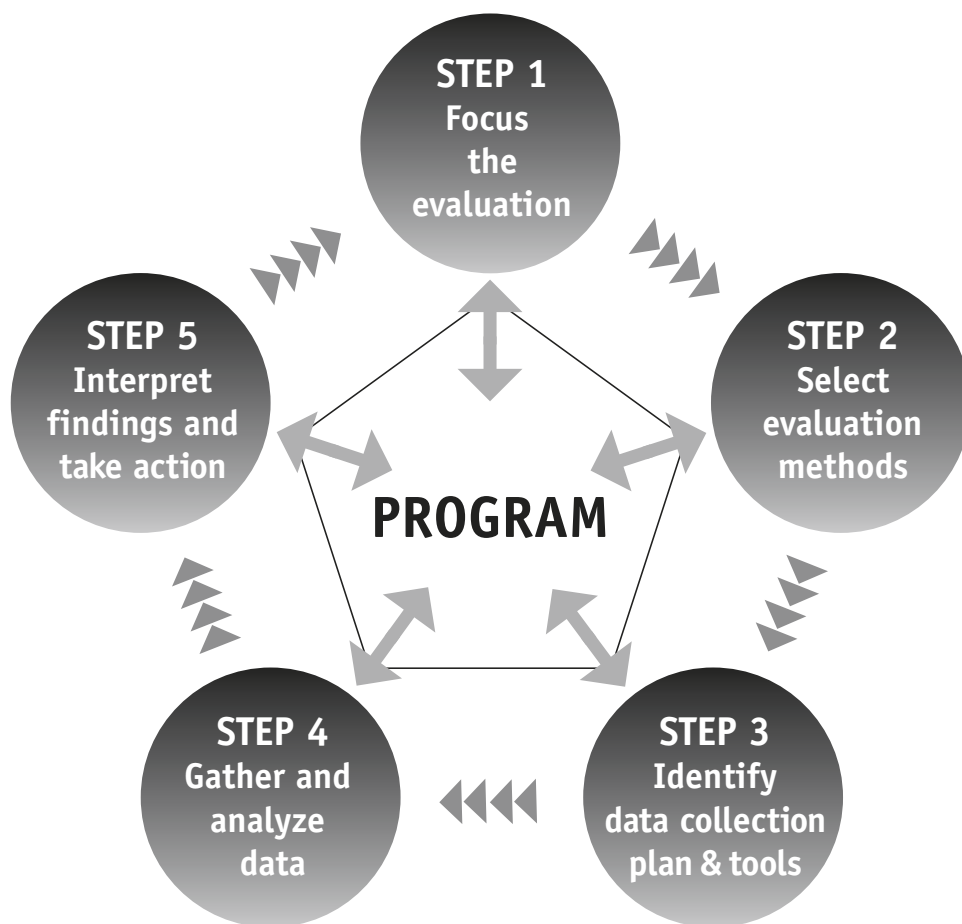
## Overview of Evaluation Steps

Planning an evaluation is a fairly complex undertaking, but it can be broken down into manageable steps. The *Comprehensive Guidelines* provide detailed evaluation direction using five steps to take the reader through the complete process of conducting an effective program evaluation. Research design guidance, including designing valid comparisons,

controlling potential biases or confounding variables, and determining sample sizes are key issues that have often been handled poorly in past evaluations. The *Comprehensive Guidelines* also provide guidance in the practical aspects of choosing the kinds of data to collect and how to collect it, and discuss appropriate data-handling and analysis procedures, interpretation, reporting, and use of evaluation results.

The evaluation process is shown in the following diagram, and then the five steps are briefly described to provide an overview of what is involved in a good program evaluation.

## Five Steps to Evaluating Driver Education Programs



The steps can be best understood by looking through the following overview of the activities involved in completing each one. The steps are discussed in more detail in the *Comprehensive Guidelines* for a comprehensive range of different types of driver education evaluations. Some of the steps are also discussed in *Evaluating Driver Education Programs: How-To Guide*, but they are limited to the more basic aspects of formative evaluation.

<b>STEP 1</b>		<b>Focus the Evaluation</b>	
1A. DESCRIBE THE PROGRAM		1B. PLAN THE EVALUATION	
<ul style="list-style-type: none"> <li>Identify stakeholders, and user and program needs</li> <li>Identify the program's vision, goals, and objectives</li> <li>Identify and document program activities, resources, and context</li> <li>Develop a program logic model</li> <li>Assess program readiness to be evaluated</li> </ul>		<ul style="list-style-type: none"> <li>Identify the purpose of the evaluation</li> <li>Identify knowledge from driver education evaluations</li> <li>Identify potential users and uses of the evaluation</li> <li>Identify key evaluation questions and targets</li> </ul>	
<b>STEP 2</b>		<b>Select the Evaluation Methods</b>	
2A. DETERMINE EVALUATION APPROACH		2B. DETERMINE EVALUATION DESIGN	
<ul style="list-style-type: none"> <li>Identify evaluation approach options</li> <li>Determine evaluation level</li> <li>Select research methods</li> </ul>		<ul style="list-style-type: none"> <li>Develop research design</li> <li>Determine sample sizes</li> <li>Develop ethics and rights of human subjects procedures</li> </ul>	
<b>STEP 3</b>		<b>Identify the Data Collection Plan and Tools</b>	
3A. DEVELOP DATA COLLECTION PLAN		3B. SELECT & ASSESS DATA COLLECTION TOOLS	
<ul style="list-style-type: none"> <li>Determine appropriate data types and data gathering methods</li> <li>Specify data and sources</li> <li>Identify indicators for program success</li> <li>Assess feasibility of data collection plan</li> </ul>		<ul style="list-style-type: none"> <li>Select, modify or develop tools</li> <li>Conduct quality assessment of tools and revise</li> </ul>	
<b>STEP 4</b>		<b>Gather, Analyze, and Summarize the Data</b>	
4A. DEVELOP LOGISTICS PLAN AND TRAINING PROCEDURES	4B. GATHER AND ENTER DATA	4C. ANALYZE AND SUMMARIZE DATA	
<ul style="list-style-type: none"> <li>Develop data collection logistics plan</li> <li>Develop procedures to train data collection personnel and conduct training</li> </ul>	<ul style="list-style-type: none"> <li>Ensure timely and consistent data collection</li> <li>Enter data and ensure accuracy</li> <li>Ensure confidentiality and security of data</li> </ul>	<ul style="list-style-type: none"> <li>Identify data analysis procedures and conduct analysis</li> <li>Assess, synthesize, and summarize data analysis results</li> </ul>	
<b>STEP 5</b>		<b>Interpret and Act Upon the Evaluation Findings</b>	
5A. INTERPRET AND DOCUMENT FINDINGS		5B. MAKE RECOMMENDATIONS AND TAKE ACTION	
<ul style="list-style-type: none"> <li>Interpret findings</li> <li>Prepare conclusions and make judgments</li> <li>Document evaluation process and findings in evaluation report</li> <li>Undertake peer review</li> </ul>		<ul style="list-style-type: none"> <li>Prepare recommendations</li> <li>Ensure feedback, follow-up, and dissemination of evaluation results</li> <li>Undertake actions to ensure use of evaluation and share lessons learned</li> <li>Determine changes to implement in next evaluation cycle and prepare action plan</li> </ul>	

*“One of the first tasks of an evaluation is to describe the program to ensure that everyone understands how its goals and objectives are linked to program activities.”*



# Selected Details in Planning Evaluation Projects

This section introduces a few of the most important evaluation planning tools. These will help you understand some of the evaluation strategies and activities that are essential to effective evaluations. They need to be discussed internally and with an external evaluator as well, if you are thinking of hiring someone to assist you with an evaluation.

## **The Logic Model—A Logical Place to Start**

One of the first tasks of an evaluation is to describe the program to ensure that everyone understands how its goals and objectives are linked to program activities. A logic model is a useful way to depict the critical parts of a program and how they are supposed to fit together. Logic models typically take the form of a flow chart, table or diagram to represent the relationships between program goals, assumptions, objectives, activities, target groups, stakeholders, and outcomes.

The point where you start to create a logic model depends on the program's stage of development. An existing program will use a top-down approach, starting with goals and objectives, and working down through activities to outcomes and impacts. For a program that is just being planned, a bottom-up approach will be a more likely choice. This means starting with the expected impacts and working up through the activities that are required to achieve the program's objectives.

Creating a logic model is a worthwhile exercise, even without a following evaluation, since it makes one think critically about vision statements, goals and objectives, strategic and operational plans, organizational structure, and budgets. It may also be helpful to consult with stakeholders to ensure that, from their perspectives, nothing critical has been omitted. Remember, however, that the logic model chart or table is intended to be a summary of the program and therefore should only be a page or two.

Logic models can take many forms. An example of a driver education program logic model in chart form is shown on the next page. This chart provides examples of the type of information that can be included in a logic model for a driver education program and for planning its evaluation.

<b>Example of a Driver Education Program Logic Model</b>			
<b>Program Goals and Objectives</b>	<b>Program Processes and Activities</b>	<b>Outcomes and Impacts</b>	<b>Target Groups</b>
<b>Goal: PROGRAM VIABILITY</b>			
Objective: Economic competitiveness	Marketing	Program sales	Management, students, parents
	Operations management	Efficiency	Management, students, parents
		Financial control	Management
	Quality control	Documented quality	Management, students, parents
	Government relations	Regulatory compliance	Management
	Customer service	Customer satisfaction	Management, students, parents
<b>Goal: DRIVER MOBILITY</b>			
Objective: Starting independent driving career	Classroom teaching	Basic knowledge	Students
	In-car practice	Basic skill	Students
		Student and parent confidence	Students, parents
<b>Goal: DRIVER SAFETY</b>			
Objectives: 1. Performance capability	Knowledge teaching	Rules	Students
		Expectations	
	Skills training	Vehicle handling	Students
		Attention control	
		Hazard perception	
	Risk appreciation		
2. Driving results	Insight training	On-road performance	Students
	Practice/habit formation	Crash reduction	Students



## Evaluation Questions and Targets

Once we have clarified the goals and objectives, and spelled out the logic of how the program is supposed to achieve them, we are in a position to identify the highest priority evaluation questions. These questions will be specific to each objective and to the activities that address each objective. Taking the objective *Performance Capability* from the sample logic model as an example, we might ask: Do our current *Skills Training* activities produce *Hazard Perception* skills in our students that are measurably better than those produced by another program?

Some evaluators use the “SMART” principle to ensure the feasibility and adequacy of evaluation questions: **S**pecific; **M**easurable; **A**ctionable; **R**elevant; **T**imely. These criteria also are a good way to check the priority assigned to the questions. If a question fails to meet any of these five criteria, the question can be revised or else eliminated as a high priority for this evaluation cycle.

Several factors to include in determining the highest priority questions are: the type of evaluation, the parts or areas of the program to be evaluated, and the specific targets in those areas. Guided by the program logic model, not all questions will be of equal importance, and only a limited number of questions can be answered in one evaluation cycle. The table on the next page illustrates a way of looking at program areas and specific evaluation targets for formative and summative evaluation purposes.

### THE SMART PRINCIPLE

Some evaluators use the “SMART” principle for the feasibility and adequacy of evaluation questions:

**S**pecific  
**M**easurable  
**A**ctionable  
**R**elevant  
**T**imely

Adapted from *A Program Evaluation Tool Kit*, Porteous, Sheldrick and Stewart 1997.

<b>Driver Education Evaluation Targets</b>		
<b>Evaluation Types</b>	<b>Program Areas</b>	<b>General Evaluation Targets</b>
Formative Evaluation	Program context	Regulatory environment
		Contractual environment
	Business processes	Quality management and control
		Marketing
		Customer service
	Program standards	Benchmarking and certification
		Transferability of the program
	Instructional products	Curriculum materials
		Tests and measurement
	Instructional processes	Instructor preparation
		Curriculum delivery; in-car practice
		Instructional facilities
Summative Evaluation	Student outcomes	Knowledge outcomes
		Skill outcomes
		Motivation outcomes
		Mobility outcomes
		Behavioral outcomes
	Social impacts	Crash reduction impacts
		Injury reduction impacts
		Socioeconomic impacts
Metaevaluation	Evaluation quality	Evaluation effectiveness

## Overcoming Barriers to Evaluation

Barriers to evaluation result from a number of factors. These include limited knowledge and resources, lack of requirements for evaluation, and certain common beliefs and misconceptions.

One seemingly common belief, which hopefully is a misconception, is that objective evaluation is risky to programs. A manager could say, "What if the scientific evaluation types find my program not to work very well? Better stick with feel-good evaluation." This belief is not altogether unfounded. Remember that driver education lost support after the DeKalb study in the 1970s.

Again the key is the *systematic* part of systematic program evaluation. If there has been an ongoing stepwise system of evaluation to improve a program, there should be little risk in moving ahead with further evaluations. Before doing a summative evaluation we should have good reason, based on earlier objective evaluations, to think that we will find the positive effects we are looking for. It is also important, of course, to plan for unexpected findings. In a systematic evaluation, enough should be known about the driver education program and how it is working that if a negative result is found, you should be able to make changes, improve the program, and evaluate again.

## Evaluation Resources and Help

An important aspect of an effective program evaluation is determining who will participate and when outside resources are needed. An in-house program of evaluation may be feasible, but it would not be unusual for evaluation team members to feel that they need some outside help. Gaps between identified evaluation needs and the internal resources and expertise are good indicators of whether outside expertise is needed.

To begin, consider what skills and interests staff might have in evaluation. Training can be found to build in-house capability that will then be available on an ongoing basis. An external evaluator may be brought in to conduct all the evaluation for an organization. However, most likely a combination of internal and external resources will be used.

Since staff buy-in is essential, an organization-wide discussion about the philosophy

and objectives of program evaluation is a good idea. Evaluation can be threatening to staff and managers unless it is managed openly. Staff may think their performance will be judged by evaluation findings. As a result, they could bias the results of some kinds of measures, such as in interpreting open-ended questions or focus group input. Supervision of procedures and data handling is key to ensuring the integrity of evaluation when program staff are involved. Nevertheless, program staff can play a crucial role in evaluation and these activities may be taken on without overburdening staff workload. Some of the evaluation tasks are probably already being performed. For example, instructors may already be collecting information on student test scores, customer satisfaction, or student preferences related to materials or delivery methods. Systematic examination of existing data is a useful part of evaluation and may yield recommendations for program improvement.

On the other hand, an outside evaluator can offer new perspectives and will have broader evaluation expertise and specialized resources available, for example, computer equipment, statistical software, support staff, libraries, and research databases. An internal evaluation team and the external evaluator together can determine the evaluation questions, design the evaluation, interpret the results, and apply the findings. The evaluation team should decide how the evaluator will be used—doing most of the evaluation tasks or providing guidance to staff who conduct most of the evaluation tasks.

Experience has shown that successful projects hire evaluators sooner rather than later. Search through professional associations, local colleges or universities, and on-line and print directories. Graduate students who are doing advanced work in driver research may be able to help. If you decide to hire an evaluator, it is still important to have staff involved in the evaluation design and implementation. External evaluators need guidance in understanding a driver education program's needs, operations, and goals.



# Thinking Ahead— Evaluation in the Future of Driver Education

Driver education is changing rapidly due to graduated licensing and other factors. Driver education traditionally meant instruction only before the new driver was licensed. Now, learning is prolonged. In a few jurisdictions, such as Finland and Michigan, new drivers are required to take a second stage of training after they have been driving as licensed drivers for a short period of time.

Instructional methods and program delivery are also undergoing change. Traditionally, all driver education activities involved face-to-face interaction between instructor and learner, although classroom instruction was often supported with film and video media, and, sometimes, with simulators. More recently, self-instruction, computer-based instruction, simulation, and web-based instruction have become prevalent. These have produced profound changes in the technological, business, and regulatory context in which driver education operates. Most changes are directed to delivery efficiency, and they are largely technology-driven and entrepreneurial rather than systematic and evidence-based. It is not yet clear whether these changes will improve the safety effectiveness of driver education, and this will be part of the challenge for new evaluations.

To a much greater extent than in the past, driver education is now highly diverse. Some high school driver education programs involve many thousands of students each year, while some jurisdictions' programs are small and may only teach a small minority of new drivers. Most commercial driving schools are relatively small, many having only a single location. A few driving schools have many locations and teach thousands of students each year. Web-based programs may also teach many thousands of students. The costs to a student of enrolling in a driver education program range from nothing to hundreds of dollars. Operating input costs also vary greatly, from well under \$100 per student to several hundred dollars. School operating standards range from none to strict centralized control and ISO certification, and instructor qualifications range from very low levels to highly qualified professional teachers.

The growing diversity of driver education programs reflects desirable and vigorous new development, and it may lead to greater effectiveness in the future. This diversity also increases the need for more and better evaluation. However, it complicates evaluation, since different programs have different evaluation needs. Driver education is becoming more complex and its effects may also become more complex.

The larger public and private programs have potential for a comprehensive range of evaluations. Evaluation methods can improve these programs, and it should be possible to demonstrate the safety impacts of larger programs where such impacts occur. Smaller programs need formative types of evaluations and evaluation of student learning outcomes. Use of appropriate approaches to evaluation can help make these programs as good as possible in terms of instructional and operational effectiveness.

## Conclusion

Renewed evaluation is needed to help driver education achieve continuous improvement and reach the goal of measurable safety impacts. The needed structure for evaluation builds on evaluation models, program theory and logic, and program evaluation standards. A suitable approach includes a composite evaluation model and framework. An ongoing series of stepped evaluation actions can be used to improve programs and raise the bar of their performance and outcomes.

Driver education researchers and practitioners need to determine the type and scale of evaluation that fit their specific circumstances. The selection of program evaluation will then be based on sound decisions about what the evaluation intends to achieve and how it will aid program improvement and impact.

The *Comprehensive Guidelines* are available to promote systematic, objective evaluation of driver education. Once adopted and implemented across North America, higher standards for evaluation should lead to more effective driver education evaluation and improved driver education programs.

While evaluation is important to improving the effectiveness and efficiency of driver education, it is also important to recognize its limitations. Failure to do this has led to unfortunate policy decisions. Evaluation of driver education, like driver education itself, is evolving and still has far to go.

## REFERENCES

- Carstensen, G. 1994. *Evaluation of a new driver education system in Denmark*. Linköping, Sweden: Swedish Road Transport Research Institute.
- Carstensen, G. 2002. The effect on accident risk of a change in driver education in Denmark. *Accident Analysis and Prevention* 34 (1):111-21.
- Christie, R. 2001. *The effectiveness of driver training as a road safety measure: A review of the literature*. Melbourne, Australia: Royal Automobile Club of Victoria (RACV) Ltd., Public Policy Group.
- Dreyer, D., and M. Janke. 1979. The effects of range versus nonrange driver training on the accident and conviction frequencies of young drivers. *Accident Analysis and Prevention* 11:179-98.
- Elvik, R., and T. Vaa, eds. 2004. *The handbook of road safety measures*. Amsterdam: Elsevier.
- Engström, I., N. P. Gregersen, K. Hernetkoski, E. Keskinen, and A. Nyberg. 2003. *Young novice drivers, driver education and training: Literature review*. Linköping, Sweden: Swedish National Road and Transport Research Institute.
- Forsyth, E., G. Maycock, and B. Sexton. 1995. *Cohort study of learner and novice drivers: Part 3, Accidents, offences and driving experience in the first three years of driving*. Crowthorne, Berkshire: Transport Research Laboratory.
- Gregersen, N. P. 1994. Systematic cooperation between driving schools and parents in driver education, An experiment. *Accident Analysis and Prevention* 26 (4):453-61.
- Haworth, N., N. Kowadlo, and C. Tingvall. 2000. *Evaluation of pre-driver education program*. Victoria, Australia: Monash University Accident Research Centre.
- Joint Committee on Standards for Educational Evaluation. 1994. *The Program Evaluation Standards. How to assess evaluations of educational programs*. 2nd ed. Thousand Oaks, CA: Sage Publications.
- Levy, D. T. 1988. The effects of driving age, driver education, and curfew laws on traffic fatalities of 15-17 year olds. *Risk Analysis* 8 (4):569-74.
- Levy, D. T. 1990. Youth and traffic safety: The effects of driving age, experience, and education. *Accident Analysis and Prevention* 22 (4):327-34.

- Lonero, L., K. Clinton, J. Brock, G. Wilde, I. Laurie, and D. Black. 1995. *Novice driver education model curriculum outline*. Washington, DC: AAA Foundation for Traffic Safety.
- Lonero, L. P., K. M. Clinton, B. N. Persaud, M. L. Chipman, and A. M. Smiley. 2005. *A longitudinal analysis of Manitoba Public Insurance Driver Education Program*. Winnipeg, Manitoba: Manitoba Public Insurance.
- Lonero, L. P., K. M. Clinton, G. J. S. Wilde, K. Roach, A. J. McKnight, H. Maclean, S. J. Guastello, and R. Lambie. 1994. *The roles of legislation, education, and reinforcement in changing road user behaviour*. Toronto, ON: Ontario Ministry of Transportation.
- Masten, S. V., and E. A. Chapman. 2003. *The effectiveness of home-study driver education compared to classroom instruction: The impact on student knowledge, skills, and attitudes*. Report to the Legislature of the State of California. Report no. CAL-DMV-RSS-03-203. Sacramento, CA: California Department of Motor Vehicles.
- Masten, S. V., and E. A. Chapman. 2004. The effectiveness of home-study driver education compared to classroom instruction: The impact on student knowledge and attitudes. *Traffic Injury Prevention* 5 (2):117-21.
- Mayhew, D. R., and H. M. Simpson. 1997. *Effectiveness and role of driver education and training in a graduated licensing system*. Ottawa, ON: Traffic Injury Research Foundation.
- Mayhew, D. R., and H. M. Simpson. 2002. The safety value of driver education and training. *Injury Prevention* 8 (Supplement II):3-8.
- McKenna, C. K., B. Yost, R. F. Munzenrider, and M. L. Young. 2000. *An evaluation of driver education in Pennsylvania*. Harrisburg, PA: Pennsylvania Department of Transportation.
- Nichols, J. L. 2003. *Driver education review*. Paper presented at the National Transportation Safety Board Forum on Driver Education and Training, October 2003. Washington, DC: National Transportation Safety Board.
- Patton, M. Q. 1997. *Utilization-focused evaluation: The new century text*. 3rd ed. Thousand Oaks, CA: Sage Publications.
- Porteous, N. L., B. J. Sheldrick, and P. J. Stewart. 1997. *A program evaluation tool kit—A blueprint for public health management*. Ottawa, ON: Public Health Research, Education and Development Program, Ottawa-Carleton Health Department.
- Potvin, L., F. Champagne, and C. Laberge Nadeau. 1988. Mandatory driver training and road safety: The Quebec experience. *American Journal of Public Health* 78:1206-9.
- Ray, H. W., M. Sadof, J. Weaver, J. R. Brink, and J. R. Stock. 1980. *Safe Performance Secondary School Driver Education Curriculum Demonstration Project*. Washington, DC: U.S. Department of Transportation, National Highway Traffic Safety Administration.



- Roberts, I., I. Kwan, and the Cochrane Injuries Group Driver Education Reviewers. 2004. School based driver education for the prevention of traffic crashes (Cochrane Review). In *The Cochrane Library* (1), Chichester, UK: John Wiley & Sons, Ltd.
- Robertson, L. S. 1980. Crash involvement of teenaged drivers when driver education is eliminated from high school. *American Journal of Public Health* 70 (6):599-603.
- Robertson, L. S., and P. L. Zador. 1978. Driver education and fatal crash involvement of teenage drivers. *American Journal of Public Health* 68 (10):959-65.
- Rossi, P. H., M. W. Lipsey, and H. E. Freeman. 2004. *Evaluation. A systematic approach*. 7th ed. Thousand Oaks, CA: Sage Publications.
- Scriven, M., quoted in Davidson, E. J. 2004. *Evaluation methodology basics*. Thousand Oaks, CA: Sage Publications.
- Siegrist, S. 2003. *Driver training and licensing—The European perspective*. Paper presented at the National Transportation Safety Board Forum on Driver Education and Training, October 2003. Washington, DC: National Transportation Safety Board.
- Simpson, H., ed. 1996. *New to the road: Reducing the risks for young motorists. First Annual International Symposium of the Youth Enhancement Service, June 8-11, 1995. Los Angeles, CA: Youth Enhancement Service*.
- Smiley, A., L. Lonerio, and M. Chipman. 2004. *Final Report. A review of the effectiveness of driver training and licensing programs in reducing road crashes*. Paris, France: MAIF Foundation.
- Stock, J. R., J. K. Weaver, H. W. Ray, J. R. Brink, and M. G. Sadof. 1983. *Evaluation of Safe Performance Secondary School Driver Education Curriculum Demonstration Project*. Washington, DC: U.S. Department of Transportation, National Highway Traffic Safety Administration.
- Trochim, W. 2001. *The research methods knowledge base*. Cincinnati, OH: Atomic Dog Publishing.
- Vernick, J. S., G. Li, S. Ogaitis, E. J. MacKenzie, S. P. Baker, and A. C. Gielen. 1999. Effects of high school driver education on motor vehicle crashes, violations, and licensure. *American Journal of Preventive Medicine* 16 (1S):40-46.
- Waller, P. F. 1992. New Evaluation Horizons: Transportation Issues for the 21st Century. *Evaluation Practice* 13 (2):103-15.
- Wiggins, S. 2005. *Graduated Licensing Program interim evaluation report—Year 3*. Vancouver, BC: Insurance Corporation of British Columbia.
- Williams, A. F. 2003. Teenage drivers: Patterns of risk. *Journal of Safety Research* 34 (1S):5-15.
- Woolley, J. 2000. *In-car driver training at high schools: A literature review*. Adelaide, Australia: Transport SA.

Wynne-Jones, J. D., and P. M. Hurst. 1984. *The AA driver training evaluation*. Traffic Research Report No. 33. Wellington, New Zealand: Ministry of Transport.

Zhao, J., R. E. Mann, M. Chipman, E. Adlaf, G. Stoduto, and R. G. Smart. 2006. The impact of driver education on self-reported collisions among young drivers with a graduated license. *Accident Analysis and Prevention* 38:35-42.

## APPENDIX A: Glossary of Terms

The following glossary of terms is a compilation of definitions from several evaluation sources. A more complete glossary of evaluation and research terms is contained in *Evaluating Driver Education Programs: Comprehensive Guidelines*.

### A

**Analysis:** The process of systematically applying statistical techniques or logic to interpret, compare, categorize, and summarize data collected in order to draw conclusions.

**Assumptions:** Hypotheses about conditions that are necessary to ensure that: (1) planned activities will produce expected results; (2) the cause and effect relationships between the different levels of program results will occur as expected.

**Auditing:** An independent, objective, systematic process that assesses the adequacy of the internal controls of an organization, and the effectiveness of its risk management and governance processes, in order to improve its efficiency and overall performance. It verifies compliance with established rules, regulations, policies and procedures and validates the accuracy of financial reports.

### B

**Benchmark:** Reference point or standard against which program effects can be assessed. A benchmark refers to the performance that has been achieved in the recent past by the same or other comparable organizations, or what can be reasonably inferred to have been achieved in similar circumstances. A referenced behavior for comparing observed performance at a given level.

**Bias:** A constant error; any systematic influence on measures, judgments, or statistical results, irrelevant to the purpose of the evaluation. Statistical bias is inaccurate representation that produces systematic error in a research finding. Bias may result in overestimating or underestimating certain characteristics of the population. It may result from incomplete information or invalid data collection methods and may be intentional or unintentional.



**Clinical Trial:** An experiment where the participants are patients, usually involving a comparison of a treatment group (who receive a treatment or intervention) and a control group who do not.

**Coding:** The process of transforming data, evidence, information, judgments, notes, and responses to numeric and/or alphabetic codes for data analysis.

**Comparability:** The similarity of phenomena such as attributes, performances, assessments, and data sources, being examined. The amount or degree of comparability is often used to determine the appropriateness of using one phenomenon in lieu of another and to help ensure fairness.

**Confidence Interval:** The probability, based on statistics, that a number will be between an upper and lower limit. The measure of the precision of an estimated value. The interval represents the range of values, consistent with the data that is believed to encompass the “true” value with high probability (usually 95%). The confidence interval is expressed in the same units as the estimate. Wider intervals indicate lower precision; narrow intervals indicate greater precision.

**Confidentiality:** The obligation not to disclose the identity of respondents, and the obligation of persons to whom private information has been given, not to use the information for any purpose other than that for which it was given.

**Content Analysis:** A set of procedures for collecting and organizing non-structured information into a standardized format that allows one to make inferences about the characteristics and meaning of written and otherwise recorded material.

**Control Group:** A group as closely as possible equivalent to an experimental treatment group (one that is exposed to a program, project, or instructional material), and exposed to all the conditions of the investigation except the program, project, or instructional material being studied.

**Cost/Benefit Analysis:** A type of analysis that compares the costs and benefits of programs in money terms. If the benefits as expressed in monetary terms are greater than the money spent on the program, then the program is considered to be of absolute benefit. Cost/benefit analysis can be used to compare interventions that have different outcomes, and comparisons are also possible across sectors.

## D

**Data:** The information produced by or used in an evaluation. Data are numbers, words, pictures, ideas, or any type of information used.

**Data Analysis:** The process of organizing, summarizing, and interpreting numerical, narrative, or artifact data, so that the results can be validly interpreted.

## E

**Effectiveness:** A measure of the extent to which a program achieves its planned results (outputs, outcomes and goals), or of how economically or optimally inputs (financial, human, technical and material resources) are used to produce outputs.

**Evaluation:** A time-bound exercise that attempts to assess systematically and objectively the relevance, performance and success of ongoing and completed programs. Evaluation is undertaken selectively to answer specific questions about what worked and what did not work, and why. Evaluation commonly aims to determine a program's relevance, validity of design, efficiency, effectiveness, impact, and sustainability.

**Evaluation Design:** A blueprint developed to answer questions about a program. It includes a clear statement about the purpose and plans for gathering, processing and interpreting the information needed to answer the questions. More specifically, it represents the set of decisions that determine how an evaluation is to be conducted, including identifying purposes and use of the information, developing or selecting of assessment methods, collecting assessment information, judging, scoring, summarizing and interpreting results, reporting evaluation findings, and following up on the evaluation results.

**Evaluation Methods:** Data collection options and strategies selected to match or fit the overall design and answer the evaluation questions. Methods depend on knowing who the information is for, how it will be used, what types of information are needed and when, and the resources available.

**Experimental Design:** The plan of an experiment, including selection of subjects, order of administration of the experimental treatment, the kind of treatment, the procedures by which it is administered, and the recording of the data (with special reference to the particular statistical analyses to be performed).

# F

**Feasibility:** The coherence and quality of a program strategy that makes successful implementation likely. The extent to which resources allow an evaluation to be conducted.

**Feedback:** Transmission of findings of monitoring and evaluation activities organized and presented in an appropriate form for dissemination to users in order to improve program management, decision-making and organizational learning. Feedback may include findings, conclusions, recommendations and lessons learned from experience.

**Focus Group:** A qualitative technique developed by social and market researchers in which 6-12 individuals are brought together and interactively give their views and impressions upon a specified topic. This can include sharing insights and observations, obtaining perceptions or opinions, suggesting ideas, or recommending actions on a topic of concern. Focus groups are often homogeneous with members being generally of the same age, gender and status to encourage participation. This method provides in-depth and insightful information from a relatively small number of people.

**Formative Evaluation:** A type of evaluation undertaken during program implementation to provide information that will guide program improvement. A formative evaluation focuses on collecting data on program operations so that needed changes or modifications can be made to the program in its early stages. Formative evaluations are used to provide feedback to program managers and other personnel about the aspects of the program that are working and those that need to be changed.

# G

**Goal:** A higher order objective to which a program or intervention is intended to contribute.

# I

**Impact:** Positive and negative long-term effects on identifiable population groups produced by a program intervention, directly or indirectly, intended or unintended.

**Indicator:** A specific, measurable item of information that specifies progress toward achieving a result. More specifically, a quantitative or qualitative measure of program performance that is used to demonstrate change and which details the extent to which program results are being or have been achieved.

**Inputs:** The resources used to conduct a program.

**Instrument:** A tool used to measure or study a person, event, or other object of interest. Examples are topic guides for focus groups (qualitative instrument) and questionnaires for surveys (quantitative instrument).

**Intermediate Measures:** Tests or instruments used to assess program outcomes, that is, measurements of things that are intermediate between the program and its impacts.

**Internal Evaluation:** Evaluation conducted by a staff member or unit from within the organization being studied.

**Interview:** A series of orally delivered questions designed to elicit responses concerning attitudes, information, interests, knowledge, and opinions. Interviews may be conducted in person or by telephone, and with an individual or a group. The three major types of interviews are: (1) structured, where all questions to be asked by the interviewer are specified in advance; (2) semi-structured, where the interviewer can ask other questions and prompts in addition to the specified questions; and (3) unstructured or open-ended, where the interviewer has a list of topics (topic guide), but no or few specified questions.

## L

**Learning Outcomes:** Products of instruction or exposure to new knowledge or skills. Examples include the mastery of a new skill or successful completion of a training program.

**Logic Model:** A systematic and visual way to present the perceived relationships among the resources available to operate the program, planned activities, and the changes or results that are to be achieved. This planning and evaluation tool most often takes the form of a graphic representation (flow chart, diagram or table) that depicts the linkages among program assumptions, goals, objectives, activities, target and stakeholder groups, and outcomes.

**Longitudinal Study:** A quasi-experimental study in which repeated measurements are obtained prior to, during, and following the introduction of an intervention or treatment in order to reach conclusions about the effect of the intervention. Can be either repeated measures or time series study.

# M

**Measure:** An instrument or device that provides data on the quantity or quality of that aspect of performance being evaluated.

**Metaevaluation:** Assessing an evaluation to judge its quality and/or assess the performance of the evaluators.

**Methodology:** A description of how something will be done. A set of analytical methods, procedures and techniques used to collect and analyze information appropriate for evaluation of the particular program, component or activity.

**Modeling:** Creating a numerical representation of a program or process for purposes of statistical analysis.

**Monitoring:** A continuous management function that aims primarily at providing program managers and key stakeholders with regular feedback and early indications of progress, or lack thereof, in the achievement of intended results.



**Objectives:** Specific desired program outcomes.

**Observation:** A research method, in which the investigator systematically watches, listens to and records the phenomenon of interest.

**Outcome:** The intended or achieved short- and medium-term effects of an intervention's outputs. Outcomes represent changes in conditions that occur between the completion of program outputs and the achievement of the program's impact.

**Outcome Evaluation:** An examination of a related set of programs, components and strategies intended to achieve a specific outcome. An outcome evaluation gauges the extent of success in achieving the outcome; assesses the underlying reasons for achievement or non achievement; validates the contributions of a specific organization to the outcome; and identifies key lessons learned and recommendations to improve performance.

**Outputs:** Products and services that result from the completion of activities within a program or intervention.



## P

**Pilot Study/Testing:** A small, preliminary test, dress rehearsal or trial run.

**Population:** The whole group about which the evaluator wants to draw conclusions. A sample is a subgroup taken from the population that is often meant to be representative of the population.

**Program Theory:** An approach for planning and evaluating programs or interventions. It entails systematic and cumulative study of the links between inputs, activities, outputs, outcomes, impacts and contexts of interventions. It specifies how activities will lead to outputs, outcomes and longer-term impact and identifies the contextual conditions that may affect the achievement of results.

## Q

**Qualitative Data:** Information gathered from evaluation methods such as personal interviews, focus groups, observations and documents such as case histories, correspondence, and records.

**Qualitative Evaluation:** A type of evaluation that is primarily descriptive and interpretative, and may or may not lend itself to quantification.

**Qualitative Research:** Research that produces findings not arrived at through statistical procedures or other means of quantification, and includes in-depth interviews, observations, and participant observation.

**Quantitative Data:** Information presented and/or summarized in numerical form.

**Quantitative Evaluation:** A type of evaluation involving the use of numerical measurement and data analysis based on statistical methods.

**Quantitative Research:** A research approach that measures social phenomena and obtains numerical values that can be analyzed statistically.

**Quasi-experiment:** A research method that compares naturally occurring or other groups that are not randomly assigned. Careful matching of treatment and control groups greatly reduces or may eliminate the likelihood that the groups were different in important ways at the outset.

**Questionnaire:** An instrument consisting of a series of questions and statements that is used to collect data and information.

# R

**Randomized Controlled Trial (RCT):** A research method in which comparisons are made between treatment and control groups that are established by random assignment of individuals from the same population.

**Random Sampling:** Selecting a number of individuals from a larger group or population, so that all individuals in the population have the same chance of being selected.

**Reliability:** The extent to which the measure is consistent and accurate, or the degree to which an instrument consistently measures an attribute. The questions “Are we measuring consistently?” and “How stable is our measure?” reflect concerns with reliability.

**Research:** The general field of disciplined investigation.

# S

**Sample:** A subset of people, documents, or things that is similar in characteristics to the larger population from which it is selected.

**Sample Size:** The number of individuals selected or drawn from a population for research purposes.

**Sampling:** Techniques used to obtain a subset of a population. This includes “probability sampling” where each subject has a known statistical chance of selection and “non-probability” sampling where subjects do not have a known statistical chance of selection.

**Self-Selection Bias:** The ways in which individuals who choose to expose themselves to a program differ from those who do not.

**Stakeholders:** People, groups or entities that have a role and interest in the aims and implementation of a program.

**Statistical Significance:** Results that are determined to have no more than a small, known probability of occurring by chance, according to appropriate inferential statistical methods.

**Successful Outcome:** A favorable program result that is assessed in terms of effectiveness, impact, and sustainability.

**Summative Evaluation:** Evaluation designed to present conclusions about the merit or worth of an object and recommendations about whether it should be retained, altered, or eliminated. It includes outcome and impact evaluation that assesses the program’s overall effectiveness.

**Survey:** A method of collecting information from a sample of the population of interest, often using questionnaires or interview protocols. This is usually a quantitative method, which allows statistical inferences to be drawn from the sample about the population.

## T

**Target Group:** The stakeholders of a program who are expected to gain from the results of that program. Sectors of the population that a program aims to reach in order to address their needs.

## U

**Utility:** The value of something to someone or to an institution. The extent to which evaluations meet the information needs of their users.

## V

**Validity:** The extent to which a measure captures the dimension of interest. It is the soundness of the use and interpretation of a measure. The question “Are we actually measuring what we’re supposed to be measuring?” reflects validity.

**Variable:** An indicator assumed to represent the underlying construct or concept.



## APPENDIX B: Evaluation Resources\*

*An Evaluation Framework for Community Health Programs*

The Center for the Advancement of Community Based Public Health. 2003

<http://www.cdc.gov/eval/evalcbph.pdf>

*A Program Evaluation Tool Kit—A Blueprint for Public Health Management*

Nancy L. Porteous, Barbara J. Sheldrick, and Paula J. Stewart

Public Health Research, Education and Development Program, Ottawa-Carleton Health  
Department, Ontario, Canada. 1997

[http://ottawa.ca/city\\_services/grants/toolkit/index\\_en.shtml](http://ottawa.ca/city_services/grants/toolkit/index_en.shtml)

*Basic Guide to Program Evaluation*

Carter McNamara. 2000

[http://www.mapnp.org/library/evaluatn/fnl\\_eval.htm](http://www.mapnp.org/library/evaluatn/fnl_eval.htm)

*Evaluating Health Promotion Programs Workbook*

Centre for Health Promotion, University of Toronto

<http://www.thcu.ca/infoandresources/publications/EVALMasterWorkbookv3.6.03.06.06.pdf>

*Evaluation in Health Promotion: Principles and Perspectives*

WHO, CDC and Health Canada

[http://www.euro.who.int/eprise/main/WHO/InformationSources/Publications/Catalogue/20040130\\_1](http://www.euro.who.int/eprise/main/WHO/InformationSources/Publications/Catalogue/20040130_1)

*Key Evaluation Checklist*

Michael Scriven

<http://www.wmich.edu/evalctr/checklists/kec.htm>

*Knowledge Required to Perform the Duties of an Evaluator*

D. J. Caron

The Canadian Journal of Program Evaluation, 1993. V. 8, No. 1

---

\*The websites listed in this section were correct at time of printing and are for informational purposes only. AAA Foundation does not endorse any particular organization or website.

*Online Evaluation Resource Library*

<http://oerl.sri.com/>

*Program Evaluation Kit*

First 5 LA

Los Angeles County Children and Families First Proposition 10 Commission, Research and Evaluation Department

[http://www.first5.org/docs/Community/CommRsrc\\_EvalKit\\_0603.pdf](http://www.first5.org/docs/Community/CommRsrc_EvalKit_0603.pdf)

*Programme Manager's Planning, Monitoring and Evaluation Toolkit*

The United Nations Population Fund (UNFPA)

<http://www.unfpa.org/monitoring/toolkit.htm>

*Project STAR—Support and Training for Assessing Results. Steps in Performance Measurement*

AmeriCorps. Corporation for National and Community Service

[http://www.projectstar.org/star/AmeriCorps/ea\\_home.htm](http://www.projectstar.org/star/AmeriCorps/ea_home.htm)

*Resources* (Information about evaluation or assistance in conducting an evaluation project)

CDC Evaluation Working Group

<http://www.cdc.gov/eval/resources.htm>

*Taking Stock—A Practical Guide to Evaluating Your Own Programs*

Sally L. Bond, Sally E. Boyd, and Kathleen A. Rapp

Horizon Research, Inc. 1997

<http://www.horizon-research.com/publications/stock.pdf>

*The Art of Appropriate Evaluation: A Guide for Highway Safety Program Managers*

National Highway Traffic Safety Administration. 1999

DOT HS 808 894

*The Framework for Program Evaluation in Public Health*

CDC Evaluation Working Group, Centers for Disease Control and Prevention. 1999

<ftp://ftp.cdc.gov/pub/Publications/mmwr/rr/rr4811.pdf>

*Utilization-Focused Evaluation Checklist*

Michael Quinn Patton

<http://www.wmich.edu/evalctr/checklists/ufechecklist.htm>

*W.K. Kellogg Foundation Evaluation Handbook*

W.K. Kellogg Foundation

<http://www.wkkf.org/Pubs/Tools/Evaluation/Pub770.pdf>

*W.K. Kellogg Foundation Logic Model Development Guide*

W.K. Kellogg Foundation

<http://www.wkkf.org/Pubs/Tools/Evaluation/Pub3669.pdf>