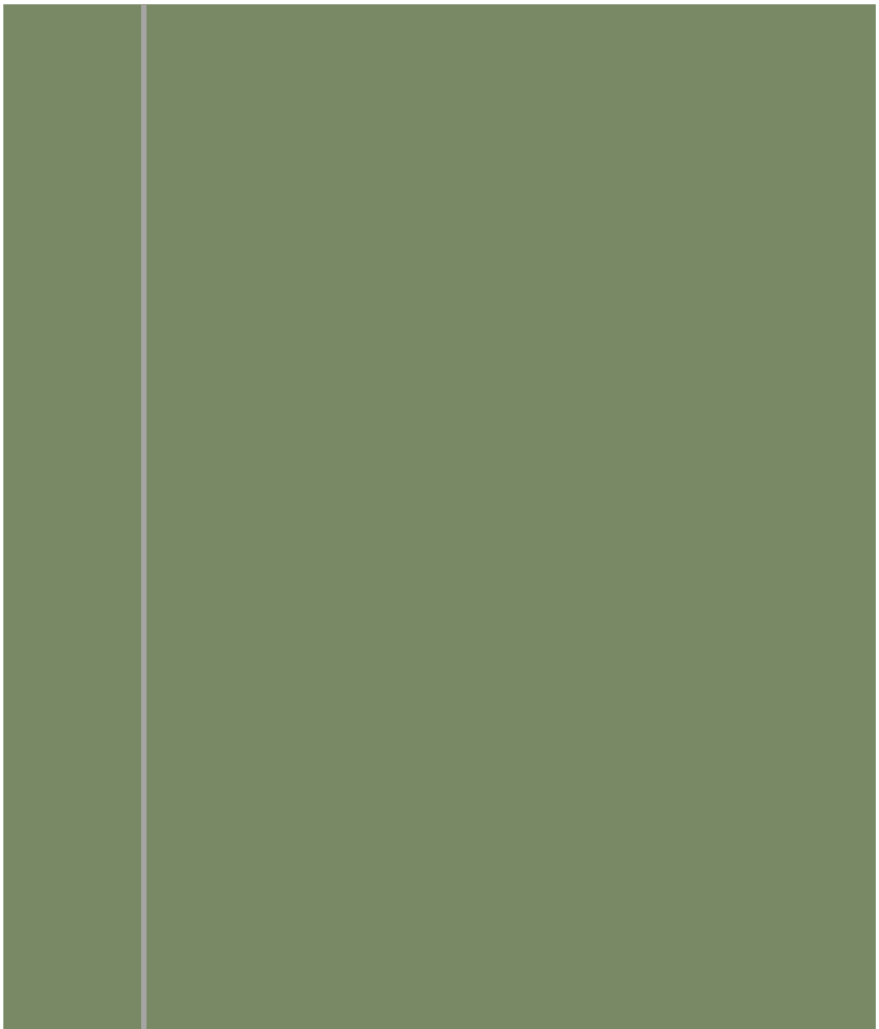
Guidance for Evaluating Traffic Safety Culture Strategies

The purpose of this document is to provide guidance to traffic safety practitioners about evaluating traffic safety culture strategies.

Center for Health and Safety Culture
Montana State University
P.O. Box 170548
Bozeman, MT 59717-0548
Phone: 406-994-7873
Fax: 406-994-7285
www.CHSCulture.org

Contents



BACKGROUND RESEARCH ON EVALUATING TRAFFIC SAFETY CULTURE STRATEGY

This guidance document was developed as a component of a project funded by Partnership for the Transformation of Traffic Safety Culture Transportation Pooled Fund Program lead by the Montana Department of Transportation. An article entitled "Assessing the Impact of Culture: A Systematic Analysis of Culture Interventions and Evaluations in Different Organizational Settings" established the context for this guidance. The following is the abstract of that article.

Over the last twenty years, transportation agencies have increasingly added culture-based approaches to the existing education, engineering, and enforcement strategies being used as a means of reducing traffic related injuries and fatalities. Despite this increased interest, there have been comparatively few evaluations of interventions designed to enhance traffic safety culture. At the same time, many other organization types have adopted culture-based strategies either to improve safety or to enhance other elements of organizational performance. In aggregate, the evaluations of culture-focused interventions across a range of settings offer an untapped body of information about the models of culture being leveraged to affect change, the intervention strategies used to impact culture, the impacts of these strategies, and more. This article presents the results of a systematic analysis of evaluations of culture-focused interventions across a variety of settings and seeks to identify patterns that could be useful to both researchers and practitioners. The findings of the study suggest that there are areas of substantial consensus regarding the nature and features of culture and the potential effectiveness of culture-based programs. At the same time, the findings also suggest that more conceptual and empirical work is warranted to further refine our understanding of culture and its functions and to build deeper understanding of how to leverage culture effectively to support health and safety efforts.
[Full article reference]

Glossary

Adapted from an *Introduction to Program Evaluation for Public Health Programs: A Self-Study Guide*[4]

Accountability: The responsibility of managers and staff to provide evidence to stakeholders and funding agencies that a strategy is effective and in conformance with its coverage, service, legal, and fiscal requirements.

Accuracy: The extent to which an evaluation is truthful or valid in what it says about a strategy, project, or material.

Case study: A data collection method that involves in-depth studies of specific cases or projects within a strategy. The method itself is made up of one or more data collection methods (such as interviews and file review).

Causal inference: The logical process used to draw conclusions from evidence concerning what has been produced or "caused" by a strategy. To say that a strategy produced or caused a certain result means that, if the strategy had not been there (or if it had been there in a different form or degree), then the observed result (or level of result) would not have occurred.

Comparison group: A group not exposed to a strategy or treatment. Also referred to as a control group.

Conclusion validity: The ability to generalize the conclusions about an existing strategy to other places, times, or situations. Both internal and external validity issues must be addressed if such conclusions are to be reached.

Confidence level: A statement that the true value of a parameter for a population lays within a specified range of values with a certain level of probability.

Control group: In quasi-experimental designs, a group of subjects who receive all influences except the strategy in exactly the same fashion as the treatment group (the latter called, in some circumstances, the experimental or strategy group). Also referred to as a non- strategy group.

Cost-benefit analysis: An analysis that combines the benefits of a strategy with the costs of the strategy. The benefits and costs are transformed into monetary terms.

Cost-effectiveness analysis: An analysis that combines strategy costs and effects (impacts). However, the impacts do not have to be transformed into monetary benefits or costs.

Cross-sectional data: Data collected at one point in time from various entities.

Data collection method: The way facts about a strategy and its outcomes are amassed. Data collection methods often used in strategy evaluations include literature search, file review, natural observations, surveys, expert opinion, and case studies.

Descriptive statistical analysis: Numbers and tabulations used to summarize and present quantitative information concisely.

Evaluation design: The logical model or conceptual framework used to arrive at conclusions about outcomes.

Evaluation plan: A written document describing the overall approach or design that will be used to guide an evaluation. It includes what will be done, how it will be done, who will do it, when it will be done, why the evaluation is being conducted, and how the findings will likely be used.

Evaluation strategy: The method used to gather evidence about one or more outcomes of a strategy. An evaluation strategy is made up of an evaluation design, a data collection method, and an analysis technique.

Experimental (or randomized) designs: Designs that try to ensure the initial equivalence of one or more control groups to a treatment group by administratively creating the groups through random assignment, thereby ensuring their mathematical equivalence. Examples of experimental or randomized designs are randomized block designs, Latin square designs, fractional designs, and the Solomon four-group.

External validity: The ability to generalize conclusions about a strategy to future or different conditions. Threats to external validity include selection and strategy interaction, setting and strategy interaction, and history and strategy interaction.

Focus group: A group of people selected for their relevance to an evaluation that is engaged by a trained facilitator in a series of discussions designed for sharing insights, ideas, and observations on a topic of concern.

Ideal evaluation design: The conceptual comparison of two or more situations that are identical except that in one case the strategy is operational. Only one group (the treatment group) receives the strategy; the other groups (the control groups) are subject to all pertinent influences except for the operation of the strategy, in exactly the same fashion as the treatment group. Outcomes are measured in exactly the same way for both groups and any differences can be attributed to the strategy.

Implicit design: A design with no formal control group and where measurement is made after exposure to the strategy.

Indicator: A specific, observable, and measurable characteristic or change that shows the progress a strategy is making toward achieving a specified outcome.

Inferential statistical analysis: Statistical analysis using models to confirm relationships among variables of interest or to generalize findings to an overall population.

Interaction effect: The joint net effect of two (or more) variables affecting the outcome of a quasi-experiment.

Internal validity: The ability to assert that a strategy has caused measured results (to a certain degree), in the face of plausible potential alternative explanations. The most common threats to internal validity are history, maturation, mortality, selection bias, regression artifacts, diffusion, and imitation of treatment and testing.

Interviewer bias: The influence of the interviewer on the interviewee. This may result from several factors, including the physical and psychological characteristics of the interviewer, which may affect the interviewees and cause differential responses among them.

Literature search: A data collection method that involves an identification and examination of research reports, published papers, and books.

Logic model: A systematic and visual way to present the perceived relationships among the resources you have to operate the strategy, the activities you plan to do, and the changes or results you hope to achieve.

Longitudinal data: Data collected over a period of time, sometimes involving a stream of data for particular persons or entities over time.

Matching: Dividing the population into "blocks" in terms of one or more variables (other than the strategy) that are expected to have an influence on the impact of the strategy.

Maturation: Changes in the outcomes that are a consequence of time rather than of the strategy, such as participant aging. This is a threat to internal validity.

Measurement validity: A measurement is valid to the extent that it represents what it is intended and presumed to represent. Valid measures have no systematic bias.

Multiple lines of evidence: The use of several independent evaluation strategies to address the same evaluation issue, relying on different data sources, on different analytical methods, or on both.

Natural observation: A data collection method that involves on-site visits to locations where a strategy is operating. It directly assesses the setting of a strategy, its activities, and individuals who participate in the activities.

Non-response bias: Potential skewing because of non-response. The answers from sampling units that do produce information may differ on items of interest from the answers from the sampling units that do not reply.

Non-sampling error: The errors, other than those attributable to sampling, that arise during the course of almost all survey activities (even a complete census), such as respondents' different interpretation of questions, mistakes in processing results, or errors in the sampling frame.

Objective data: Observations that do not involve personal feelings and are based on observable facts. Objective data can be measured quantitatively or qualitatively.

Objectivity: Evidence and conclusions that can be verified by someone other than the original authors.

Outcome evaluation: The systematic collection of information to assess the impact of a strategy, present conclusions about the merit or worth of a strategy and make recommendations about future strategy direction or improvement.

Outcomes: The results of strategy operations or activities; the effects triggered by the strategy. (For example, increased knowledge, changed attitudes or beliefs, reduced risky behaviors, reduced morbidity and mortality.)

Population: The set of units to which the results of a survey apply.

Primary data: Data collected by an evaluation team specifically for the evaluation study.

Probability sampling: The selection of units from a population based on the principle of randomization. Every unit of the population has a calculable (non-zero) probability of being selected.

Process evaluation: The systematic collection of information to document and assess how a strategy was implemented and operates.

Program/Strategy evaluation: The systematic collection of information about the activities, characteristics, and outcomes of strategy to make judgments about the strategy, improve strategy effectiveness, and/or inform decisions about future strategy development.

Program/Strategy goal: A statement of the overall mission or purpose(s) of the strategy.

Qualitative data: Observations that are categorical rather than numerical, and often involve knowledge, attitudes, perceptions, and intentions.

Quantitative data: Observations that are numerical.

Quasi-experimental design: Study structures that use comparison groups to draw causal inferences but do not use randomization to create the treatment and control groups. The treatment group is usually given. The control group is selected to match the treatment group as closely as possible so that inferences on the incremental impacts of the strategy can be made.

Randomization: Use of a probability scheme for choosing a sample. This can be done using random number tables, computers, dice, cards, and so forth.

Reliability: The extent to which a measurement, when repeatedly applied to a given situation consistently produces the same results if the situation does not change between the applications. Reliability can refer to the stability of the measurement over time or to the consistency of the measurement from place to place.

Sample size: The number of units to be sampled.

Sampling error: The error attributed to sampling and measuring a portion of the population rather than carrying out a census under the same general conditions.

Sampling method: The method by which the sampling units are selected (such as systematic or stratified sampling).

Sampling unit: The unit used for sampling. The population should be divisible into a finite number of distinct, non-overlapping units, so that each member of the population belongs to only one sampling unit.

Secondary data: Data collected and recorded by another (usually earlier) person or organization, usually for different purposes than the current evaluation.

Selection and program/strategy interaction: The uncharacteristic responsiveness of strategy participants because they are aware of being in the strategy or being part of a survey. This interaction is a threat to internal and external validity.

Selection bias: When the treatment and control groups involved in the strategy are initially statistically unequal in terms of one or more of the factors of interest. This is a threat to internal validity.

Setting and program/strategy interaction: When the setting of the experimental or pilot project is not typical of the setting envisioned for the full-scale strategy. This interaction is a threat to external validity.

Stakeholders: People or organizations that are invested in the strategy or that are interested in the results of the evaluation or what will be done with results of the evaluation.

Standard: A principle commonly agreed to by experts in the conduct and use of an evaluation for the measure of the value or quality of an evaluation (e.g., accuracy, feasibility, propriety, utility).

Standard deviation: The standard deviation of a set of numerical measurements (on an "interval scale"). It indicates how closely individual measurements cluster around the mean.

Statistical analysis: The manipulation of numerical or categorical data to predict phenomena, to draw conclusions about relationships among variables or to generalize results.

Statistical model: A model that is normally based on previous research and permits transformation of a specific impact measure into another specific impact measure, one specific impact measure into a range of other impact measures, or a range of impact measures into a range of other impact measures.

Statistically significant effects: Effects that are observed and are unlikely to result solely from chance variation. These can be assessed through the use of statistical tests.

Stratified sampling: A probability sampling technique that divides a population into relatively homogeneous layers called strata and selects appropriate samples independently in each of those layers.

Subjective data: Observations that involve personal feelings, attitudes, and perceptions. Subjective data can be measured quantitatively or qualitatively.

Surveys: A data collection method that involves a planned effort to collect needed data from a sample (or a complete census) of the relevant population. The relevant population consists of people or entities affected by the strategy (or of similar people or entities).

Testing bias: Changes observed in a quasi-experiment that may be the result of excessive familiarity with the measuring instrument. This is a potential threat to internal validity.

Treatment group: In research design, the group of subjects that receives the strategy. Also referred to as the experimental or strategy group.

Utility: The extent to which an evaluation produces and disseminates reports that inform relevant audiences and have beneficial impact on their work.

Overview

The purpose of this document is to provide guidance to traffic safety practitioners about evaluating traffic safety culture strategies. It begins with a description of traffic safety culture strategies and is followed by a summary of evaluation types, components of effective evaluations, and steps to follow to complete an evaluation. It concludes with an evaluation example from an actual project to improve traffic safety culture.

Evaluation is a large, diverse field of scientific research. Clearly, this guidance document cannot cover all there is to know about evaluation. However, it can promote an idea called "evaluative thinking."[1] Evaluative thinking is a problem-solving approach to designing, selecting, and allocating resources to traffic safety strategies. It seeks credible evidence to provide answers about the effectiveness and sustainability of traffic safety strategies.

> *Evaluative thinking is a cognitive process in the context of evaluation, motivated by an attitude of inquisitiveness and a belief in the value of the evidence, that involves skills such as identifying assumptions, posing thoughtful questions, pursuing deeper understanding through reflection and perspective-taking and making informed decisions in preparation for action.*[2]

This guide can help traffic safety practitioners bolster their knowledge about evaluation and include evaluation in their proposal requests and activities – in other words, promote evaluative thinking. This guide is not intended to teach practitioners how to conduct extensive evaluations themselves. Instead, it provides guidance so that more evaluation activities are included in efforts to grow positive traffic safety culture – thereby improving effectiveness of these strategies.

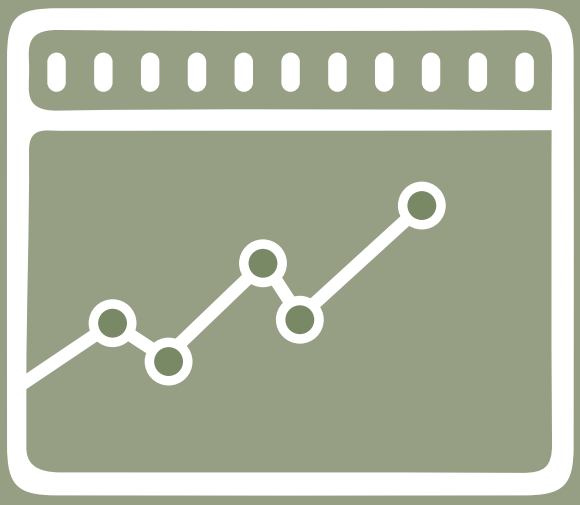
After studying this guide, traffic safety practitioners will be able to:

A. Discuss the importance of evaluating traffic safety culture strategies
B. Understand types of evaluations and components of effective evaluations
C. Ask appropriate questions of evaluation proposals to select effective evaluations
D. Better understand and make meaning of completed evaluations

PROMOTING EVALUATIVE THINKING

Many traffic safety practitioners and stakeholders already engage in forms of evaluative thinking. Discussing the value of evaluative thinking within traffic safety will grow its importance. Here are talking points to foster conversations about the importance of evaluative thinking.

1. Fatal crashes and serious injuries have a significant impact on public health.
2. Zero traffic fatalities and serious injuries is the only acceptable goal.
3. To be successful in reaching this goal, we must learn to use innovative strategies and grow evidence of their effectiveness.
4. Evaluations inform which strategies are effective and generate knowledge about how to make strategies more effective and sustainable.
5. Traffic safety practitioners can seek opportunities to include process, outcome, and impact evaluations in the projects they implement, manage, and fund.
6. Effective evaluations require quality data and appropriate comparisons.
7. Evaluations should include engaging stakeholders, developing careful descriptions of strategies, and identifying quality data and appropriate comparisons.
8. Traffic safety practitioners can create opportunities to review and discuss evaluation results with stakeholders to gather lessons learned and identify opportunities for improvement in future efforts.
9. More consistent and rigorous evaluations will accelerate learning and effectiveness of strategies in improving traffic safety.
10. Investing in training to help staff become more familiar with evaluation design and contracting with evaluators will improve the effectiveness of strategies and ultimately traffic safety.



Traffic safety practitioners invest significant time and resources in strategies to improve traffic safety. Everyone wants these resources to be invested in strategies that actually make a difference. Evaluation can ensure these investments are effective.

According to the American Evaluation Association, "evaluation involves assessing the strengths and weaknesses of strategies" (including policies and laws) "to improve their effectiveness."[3] By understanding the strengths and weaknesses of a strategy, those implementing the strategy can make adjustments to make the strategy more effective.

Information about if and how a strategy works provides important evidence. This evidence becomes the basis for considering a strategy as "evidence-based." Evidence is critical to making good decisions (e.g., evidence-based decision making).

The Centers for Disease Control and Prevention (CDC) lists several important reasons for evaluating strategies:[4]

- To assess effectiveness and inform good management practices by
  - comparing actual outcomes with intended outcomes,
  - comparing outcomes with those of previous years, and
  - establishing realistic intended outcomes (standards) for future performance.
- To foster sustained improvements in traffic safety by
  - focusing attention on issues important to the effectiveness of the strategy,
  - promoting a strategy by documenting and sharing its effectiveness,
  - recruiting new partners (who want to join in contributing to effective strategies),
  - enhancing the image of the strategy,
  - sustaining or increasing funding,
  - providing direction and informing training for staff and partners to implement the strategy effectively in the future,
  - informing what training and technical assistance is needed to improve effectiveness,
  - informing long-range planning, and
  - justifying the investment of resources by legislators or other stakeholders by showing the strategy is effective.

Evaluating Traffic Safety Culture Strategies

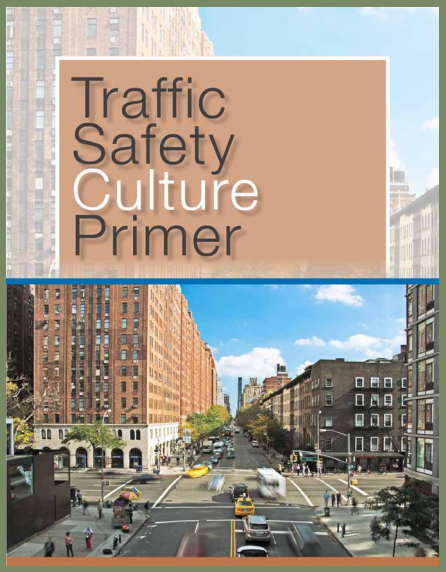
Traffic safety culture can be defined as the beliefs shared by a group of road users or stakeholders that influence their behaviors that impact traffic safety. This definition of culture establishes a relationship between beliefs and behaviors. Specifically, when individuals have certain beliefs, they are more likely to engage in certain behaviors. For example, if people believe that it is safe to have hands-free cell phone conversations while driving, they are more likely to engage in this risky driving behavior.

There are several basic types of beliefs including:

- Expectations about the consequences of behavior (e.g., "If I drive after using cannabis, I am more likely to cause a crash.")
- Perceptions about how common a behavior is (e.g., "I believe most people speed.")
- Perceptions about how acceptable or expected a behavior is (e.g., "My spouse expects me to use a seat belt.")
- Perceptions about an individual's ability to perform the behavior (e.g., "I am comfortable not answering my cell phone while driving.")

LEARNING MORE ABOUT TRAFFIC SAFETY CULTURE

Learn more about traffic safety culture by reading the Traffic Safety Culture Primer (https://www.mdt.mt.gov/other/webdata/external/research/docs/research_proj/tsc/TSC_PRIMER/PRIMER.pdf) or Google "traffic safety culture primer."



Traffic safety culture strategies focus on changing beliefs like these. When these beliefs change, people's behaviors are likely to change, and this change in behavior is more likely to be sustained.

In contrast, a speed bump is a physical way of changing behavior. People tend not to speed over speed bumps but will resume their speed when the speed bumps are no longer present. A speed bump does not change underlying beliefs about speeding and therefore does not result in sustained behavior change.

Traffic safety culture strategies use specific experiences designed to change beliefs. For example, workplace traffic safety training is a specific experience designed to change a worker's beliefs about specific driving practices. The training might discuss the increased risk for crashing while talking on a cell phone when driving. The training could provide information about how most employees do not drive while using a cell phone and that leadership, management, and supervisors expect drivers not to use their cell phones while driving. A workplace policy prohibiting cell phone use while driving could be reviewed. Supervisors could meet with employees and discuss how work procedures will take place without using cell phones while driving. By growing healthy beliefs among workers, the likelihood of risky driving is reduced. As fewer drivers engage in risky driving (e.g., distracted driving), fewer crashes will occur, and traffic safety will improve. This process is summarized in Figure 1.

Figure 1. Diagram of how a traffic safety culture (TSC) strategy leads to improved traffic safety

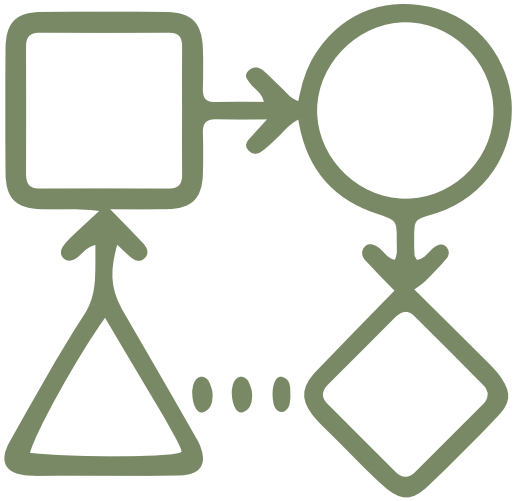
Understanding how a traffic safety culture strategy leads to improving traffic safety is important when considering evaluating a traffic safety culture strategy. There are many potential problems that could result in a traffic safety culture strategy being ineffective.

Using the same workplace training example shared previously, imagine what would happen if only 10% of the workers were trained. Only training 10% of the workers would significantly reduce the likelihood that beliefs across the workforce would change, thus reducing the likelihood that behaviors across the workforce would change, thus reducing the likelihood that crashes would be reduced.

Suppose everyone participated in the training, but the training was poorly implemented and did not change people's beliefs. If beliefs did not change, it would be unlikely that behaviors would change, and traffic safety would not improve.

Suppose everyone participated in the training, and the training changed beliefs, but it changed the wrong beliefs – beliefs that did not matter or did not influence the behavior. Behavior would not change, and traffic safety would not improve.

Understanding how a traffic safety culture strategy leads to improving traffic safety will inform how a traffic safety culture strategy should be evaluated. Specifically, the evaluation should verify the process of change underlying the strategy. For example, an evaluation might capture what percentage of the workers participated in the training, to what degree the training changed beliefs, how much subsequent risky driving behaviors changed, and whether crashes were reduced. This simple example shows that there may be several ways to evaluate a traffic safety culture strategy.

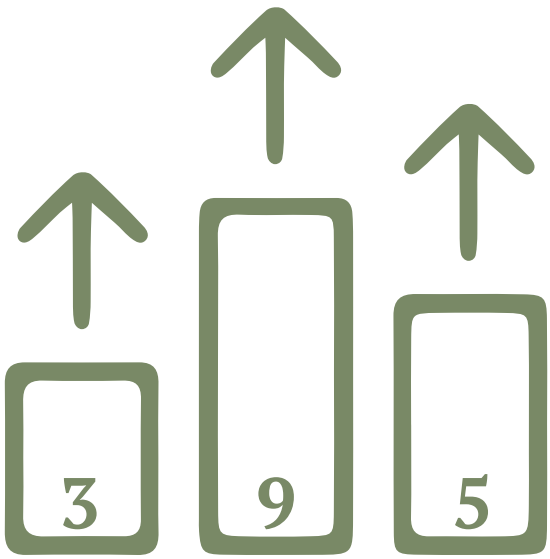


Evaluation Types

Because there are different factors impacting the effectiveness of a strategy, there are different types of evaluation. Each evaluation type provides important information to make a strategy more effective. The CDC summarizes three types of evaluation:[4]

- **Process evaluations** examine the way the strategy was implemented.
    - Was the strategy (e.g., workplace training) implemented exactly how it was designed? This is also referred to as implementing a strategy with fidelity. For training, this might include assessing the number of sessions the training required, how many sessions were completed, how much of the content was covered, etc.
    - Did the strategy reach a sufficient portion of the population (e.g., percentage of workers) to make a difference?
- **Outcome evaluations** measure the effect of the strategy on creating change. An outcome evaluation could assess to what degree the strategy (e.g., workplace training) changed beliefs. An outcome evaluation could also assess to what degree there was a change in behaviors.
- **Impact evaluations** assess the consequences of changes that result in improved public health. An impact evaluation could assess to what degree there was a reduction in distracted driving related crashes (and injuries) in a workplace.

All three types of evaluation require the gathering of data and then using those data to make comparisons and draw conclusions. Therefore, effective evaluations require using quality data and making appropriate comparisons.



Key Components of Effective Evaluations

The quality of the information (or evidence) provided by an evaluation depends on two key components of the evaluation design: the quality of the data and the way data are compared.

Data Quality

Information about the process, outcomes, and impacts of strategies is assessed from many kinds of data gathered from different sources. These data become the basis for drawing conclusions. Therefore, the quality of the data determines the quality of the conclusions made based on that data.

There are two important aspects of data quality.

1. **Data must be reliable – the data are accurate and have consistent measures.**

   For example, suppose we ask two participants in a workplace training how many people were present at the training. One person says 4, and the other says 10. This is not a reliable measure of how many people attended the training. Another data source needs to be used to provide a reliable measure of how many people attended the training (perhaps a sign-in sheet).

   Reliability can also be compromised on measures that may change over time. For example, suppose we want to know how people feel, on average, over the period of a week. One way to measure this might be to ask people once how they feel "right now." Another way might be to ask people several times a day over several days and average these responses. The second method will mostly likely create a more reliable because people's feelings can fluctuate during a week so asking them at only one point of time may measure an unusual feeling that is not representative of how they usually felt during the week.

2. **Data must be valid – the data truly represent the concepts that are being measured.**

   For example, suppose we want to assess whether a strategy changed beliefs about distracted driving. A survey question might ask, "Do you think driving distracted is dangerous - yes or no?" The person might not know what "driving distracted" means, or they might want to answer "sometimes," but that is not an option. The results from asking this question may not be a valid indicator of people's beliefs about distracted driving. A better question might be, "How dangerous is driving while having a conversation on a hands-free cell phone?" with five choices ranging from "not at all dangerous" to "extremely dangerous."

   Another challenge is when data do not represent what we think they represent. For example, we compare the number of distracted driving citations between two communities and draw the conclusion that one community has more distracted driving than another community. The data indicate how many citations were written – which may or may not reflect the prevalence of distracted driving in the community. In fact, the number of citations written may be a better indicator of enforcement activity in each community. In this case, the number of citations is not a valid measure of distracted driving.

To help illustrate concerns about reliability and validity of data, Table 1 shows examples of data that could be measured in different types of evaluation.

Table 1. Examples of Reliability and Validity Concerns with Possible Measures Used in Different Types of Evaluation

Evaluation Type Process	Source and Method for Gathering Data	Concerns About Reliability and Validity	Recommendations
Process			
Was the strategy implemented as designed?	Self-reports by those implementing the strategy indicating if they implemented it as designed	Those implementing the strategy may not know what is required during implementation to assure effectiveness. They may also have a biased opinion regarding the quality of implementation.	Develop checklists to assess implementation and, if possible, use observers to complete the checklists. Require documentation of activities (e.g., publication affidavits for placed media, number of sessions conducted with attendance records, etc.)
Did the strategy reach the intended audience?	Information provided by those implementing the strategy	While those implementing the strategy can perceive they reached most people, there could be gaps.	Collect information directly from people who participated in the strategy. This could include questions on annual workplace surveys (or personnel evaluations) about attendance at safety training or random sample surveys of larger populations asking how often they heard a message or received information.
Outcome			
Did beliefs change as expected by the strategy?	Interviewers use focus groups to assess people's beliefs	The interviewers may have different interviewing styles, which affects responses. The presence of the interviewer and other participants may create social pressure for respondents to provide only desirable answers.	Use tested or standardized self-report surveys that ensure respondent anonymity.
Did behavior change as expected by the change in beliefs?	Citation or arrest data Data reported by others without using strong methodologies (like workplace supervisors reporting prevalence and frequency of behaviors among direct reports)	Citation or arrest data typically assess law enforcement activity and may not be an indication of underlying behaviors. People's perceptions of other people's behaviors are typically very inaccurate.	Use tested or standardized self-report surveys that ensure respondent anonymity. Use observational studies with well-designed methodologies using trained observers.
Impact			
Did crashes related to the behavior decrease? Did fatalities and serious injuries decrease?	Crash databases	Limitations and inconsistencies in how crash data are collected and reported may limit validity and reliability. For example, assessing distraction after a crash may be problematic resulting in underreporting of distraction-related crashes. Furthermore, assessing impairment from drugs may be limited by lack of valid biological tests.	Understand the limitations of crash databases and use extra care when comparing results across systems that may use different methods.

Data Comparisons

Once reliable and valid data are collected, comparisons are used to make meaning of the data. There are at least four ways to make comparisons.

- **Benchmark-based evaluations** compare data with a stated reference. These references may be specified by stakeholders or based on previous implementation of the strategy. Examples include:
  - 80% of the employees will agree that not using a seat belt violates company policy.
  - Less than 10% of the population will report driving within two hours of consuming alcohol.
  - There will be fewer than 35 speed-related crash fatalities this year.

  Because benchmark-based evaluations compare a data measure with a set value, this type of comparison does not measure change. Therefore, it is unclear if the strategy resulted in meeting the benchmark or something else (or even if the benchmark was met prior to implementing the strategy). Benchmark-based evaluations cannot claim the strategy caused the change in the outcome or impact.

- **Time-based evaluations** compare data across different time points. The time points could be at the beginning and end of implementing the strategy or could be on a regular basis (like every year). The data between these time points are then compared to assess change. Examples include:
  - Beliefs about impaired driving are compared at the beginning of the first session and at the last session of a class for individuals cited for repeatedly driving under the influence of alcohol.
  - Seat belt use (as measured by observational studies) is compared to previous years.
  - Crash statistics from a specific area over the summer months (i.e., May to September) are compared from year to year.

  It is critical to acknowledge that many other factors besides the strategy can affect outcomes. Time-based evaluations (like these) cannot claim the strategy was the cause of the change (e.g., a change in driving behaviors could be due to the economy). Similarly, a lack of change may not necessarily be the consequence of an ineffective strategy; other factors may have changed such as a significant change in the population and age of drivers.

- **Place-based comparisons** compare data across different locations, usually within the same time period. To isolate the effect of the strategy, one location (test site) is the place where the strategy is implemented. The other place is similar but does not have the strategy implemented (control site). For example:
    - One county in a state implements a new strategy (test site) while another county with similar characteristics (e.g., road type, population, etc.) does not implement the strategy (control site). Outcome measures (e.g., observed seat belt use or crash data) for the same time period are compared.

  If the two sites are different only in terms of the implemented strategy, then any measured differences between these sites might be attributable to the strategy itself. However, this assertion depends on how comprehensively these sites were matched. It is very difficult to find two sites that match perfectly.

- **Combined time-place evaluations** use a combination of the time-based and place-based evaluation methods by comparing different places at two points of time. The "before" and "after" measures for each place are compared. The changes assessed in each place are then compared against each other.
    - One county in a state implements a new strategy (test site) while another county with similar characteristics (e.g., road type, population, etc.) does not implement the strategy (control site). Several measures (beliefs, behaviors, crash data) are collected in the same way before and after the strategy is implemented. In the "control site" county, none of the measures show any statistically significant changes (as expected, since the strategy was not implemented in this county). There are changes in the measures in the "test site" county.

  This evaluation design has the advantage of permitting multiple comparisons, which can reinforce conclusions about whether the strategy caused the change in measures (this is called "causality"). Notably, if the test site had similar speeding compared to the control site before the strategy was implemented AND speeding reduced at the test site but no changes were measured over the same time period at the control site after the strategy was implemented, we can be more confident in claiming that the strategy caused the change.

  It is important to note that other factors could still explain the change. For example, assume that during the implementation period, a large number of young adult males leave the test county (e.g., for work, join the military, or go to university). Such a change in the population could also cause a change in speeding behaviors in the test site – a change that was not caused by the strategy.

Steps to Plan, Implement, and Make Meaning of an Evaluation

The CDC promotes five core steps for implementing evaluations.[4]

1. **Identify, Recruit, and Engage Stakeholders.** Stakeholders include people responsible for the strategy (e.g., funders, contractors, etc.), people affected by the strategy (e.g., general population, workplaces, etc.), and those who will use the evaluation results. Early participation by stakeholders is necessary to identify questions and concerns and support access to quality data to ensure an effective evaluation. Those affected by the strategy should be included to measure exposure to the strategy and help identify unintended consequences including potential harms.

> A key purpose of stakeholder involvement is to specify "standards" for effectiveness. What does an effective evaluation mean for this strategy in this context? How does each stakeholder define and envision success? What outcomes are important to the needs of each stakeholder? Should the evaluation bolster a sense that the strategy caused the change in outcomes or is it OK just to assess change? It is important to understand these distinct perspectives to align expectations about potential interpretations of the evaluation results.

2. **Describe the Strategy.** Before starting an evaluation, it is necessary to agree on a detailed description of the strategy, including the conditions necessary for its implementation: "a comprehensive [strategy] description clarifies all the components and intended outcomes of the [strategy], thus helping you focus your evaluation on the most central and important questions."[4] Understanding how the strategy causes a change in outcomes (and subsequent positive impact to traffic safety) is critical to designing an evaluation. This understanding will inform potential process measures (e.g., how many people experienced the strategy), intermediate outcome measures (e.g., which beliefs and behaviors to examine for change), and impact measures (e.g., crash types) as well as provide insights as to how much time the strategy will take to cause changes. Practitioners can reach out to the strategy developer and ask for the "theory of change" for the strategy (the theory of change lays out the science behind how a strategy has been shown to cause the expected outcomes). Additionally, a practitioner could require a contractor implementing a strategy to articulate how the strategy causes the expected outcomes.

3. **Identify Data Measures and Comparisons to Be Performed.** In this step, the stakeholders identify the reliable and valid data that measure the process, outcome, and impact of the strategy. The sources and methods to collect these data are also identified. A comprehensive plan is developed for the evaluation, which includes the type(s) of planned comparison(s). As discussed previously, carefully consider the type of comparison, because it affects the ability to draw conclusions about strategy.

4. **Make Meaning.** This step involves analyzing the data and interpreting the results. Considerations addressed in the first step can inform efforts in this step. Too often, evaluations are reduced to one simple question: "Did the strategy work?" Often the answer is: "Yes and no." Making meaning of the evaluation should allow for greater learning to inform how to make the strategy more effective. Additional questions to ask include:

    a. Was the strategy implemented as intended? Why or why not? What could be done better next time?
    b. Did the strategy reach the intended audience? Why or why not? What could be done better next time?
    c. Did the strategy result in the intermediate outcomes (e.g., in beliefs and behaviors) expected? Why or why not? What could be done better next time?
    d. Did the strategy result in the desired impact? Why or why not? What could be done better next time?
    e. The evaluation may not inform all these questions. Asking these questions may guide how future evaluations may be modified to answer additional questions. The intent is to use the evaluation results to improve effectiveness over time by enhancing learning.

5. **Accumulate and Share Wisdom (e.g., lessons learned)**. A single evaluation, if explored and discussed by stakeholders, can generate many lessons that can inform future actions. These lessons are often much more valuable than simply answering the question "Did the strategy work?" Stakeholders should allocate time to review and discuss the evaluation results and gather lessons to share with other stakeholders.

    An evaluation can have greater impact if the lessons learned reach a variety of audiences that need the information to make decisions about strategies, planning, funding, etc. To be accessible and usable, lessons should use language familiar to stakeholders.

    It is also important to accumulate lessons learned and evidence for a strategy over time, because a single evaluation may not be enough to truly understand how best to implement a strategy or to convince stakeholders to continue support for the strategy.

Evaluation Example: A Case Study

This section presents a case study of an evaluation of a project completed by the Center for Health and Safety Culture for the Idaho Transportation Department (ITD) to decrease alcohol-impaired driving by encouraging people to intervene and prevent others from driving when impaired. The case study reviews each of the five steps described previously.

Background

In a previous research project, the Center for Health and Safety Culture ("the Center") identified beliefs associated with bystanders speaking up to prevent others from driving after drinking. This research identified potential messages that could be used in a media campaign to increase bystander engagement. The purpose of the project described in this case study was to test these messages using a universal media campaign, engage local stakeholders to use the media to reduce impaired driving, and evaluate the strategy's impact on beliefs and behaviors about bystander engagement as a way to reduce alcohol-related crashes.

Step 1. Identify, Recruit, and Engage Stakeholders

The initial stakeholders included leaders within the Idaho Transportation Department (including the Office of Highway Safety) and the researchers from the Center working on the project. These stakeholders met regularly to discuss the project design that would best meet the purpose of the project. The group decided to engage three communities to implement the strategy and use the remainder of the state as a comparison.

The three communities identified to implement the strategy were Blackfoot, Lewiston, and Twin Falls. These communities were selected because of their geographic distribution across the state, diversity of size, and their high rates of alcohol-impaired driving crashes.

Stakeholders from the three communities were identified and recruited to participate in initial training about the project. Twenty-one individuals from the three communities participated in a two-day training that reviewed the background for the strategy, how the strategy would be evaluated, and potential ways they could support the strategy.

Step 2. Describe the Strategy

The strategy was a media campaign to be augmented with additional supportive materials that could be used by local stakeholders in each of the three communities. The messages for the media campaign were designed to grow specific beliefs associated with bystander engagement including:

- "Most Idaho adults do not drink and drive."
- "Most Idaho adults agree they should try and prevent a stranger from driving after drinking."
- "Most Idaho adults agree they would try and prevent a stranger from driving after drinking."
- "Most Idaho adults agree with strongly enforcing impaired driving laws."

Additionally, media was created to demonstrate what it looked like to actually speak up to prevent impaired driving. These examples were captured in short video messages designed for placement on television. The media campaign was branded "Courageous Voices Create Safe Roads." Media including television and radio ads were developed using this brand and placed in these three communities from late 2013 to late 2014. Supportive materials including a brochure, speaking points, sample presentation, press releases, and a website landing page were also created. A media buyer was contracted to place the media to reach the three communities. The media buyer worked with stakeholders from ITD and the Center to develop a media plan for the media placement.

Step 3. Identify Data Measures and Comparisons To Be Performed

Process measures to assess the placement of the media included

- affidavits from the media buyer on exact placement locations, times, etc.;
- earned media placements in local newspapers (letters to the editor, etc.); and
- distribution of the supportive materials.

Outcome measures to assess changes in beliefs and behaviors included survey responses by adults. Paper surveys were mailed to a random sample of households in each of the three pilot communities as well as across the rest of the state before and after the media campaign. The responses in each sample were compared to reveal change. The three communities acted as the test sites, and the rest of the state acted as the control site.

Alcohol-related crashes in the three communities and the state before and after the campaign were used as a measure of impact.

Combined time-place comparisons were made to assess change. The responses in the three test sites were aggregated together. The means measuring each of the beliefs and behaviors of the responses before and after the media campaign were compared. Similar comparisons were made for the sample representing the control site (i.e., the rest of the state).

Step 4. Make Meaning

Process measures indicated that the media were placed as planned by the media buyer. The affidavits showed placements on radio stations, television stations, newspapers, and billboards. However, there was no earned media generated (no letters to the editor, articles written, etc.), and none of the supportive material created was ever distributed in bars or restaurants that serve alcohol.

Outcome measures included responses to surveys. Comparisons of survey responses before and after the media placement showed statistically significant improvements in beliefs addressed in the media messages in the test sites. Specifically, agreement with the belief that most adults think people should try to prevent a stranger from driving after drinking enough alcohol to be impaired and agreement with the statement that "I should try to prevent a stranger…" statistically significantly increased ($p<0.001$ and $p=0.008$, respectively).

Furthermore, the perception that most people would support individuals who chose to prevent a stranger from driving after drinking too much increased ($p<0.001$) as did the perception that most people would try to intervene ($p<0.001$). Other related beliefs increased as well.

Beliefs not addressed in the media messages showed no changes. No changes were seen in responses outside of the three test sites (i.e., the control site) supporting the notion that the media messages caused the changes measured in the test sites.

The surveys revealed no changes (in either the test sites or the control site) in self-reported behaviors about intervening to try and prevent a stranger from driving after having too much to drink, calling 911 to report a potentially impaired driver, or driving within two hours of drinking.

Because these behaviors are rare (most people do not drive after having too much to drink, therefore, few people can intervene), measuring changes in these behaviors can be challenging with small survey samples.

Impact measures included crash data. Crash reports indicated a slight reduction in alcohol-related crashes during the year of the campaign. However, the alcohol-related crash reduction in the pilot communities occurred at a rate similar to the reduction at the state level. Thus, the messages did not appear to have reduced alcohol-related crashes.

Step 5. Accumulate and Share Wisdom

The process measures revealed that there was not enough stakeholder engagement within the three communities to use the supportive media created. There was no known engagement (such as using the supporting media materials, working to change local practices or policies, or engaging specific groups such as schools or community groups) by local stakeholders to support the media strategy in the test sites after participating in the training. An important lesson learned was just how much effort is required to encourage local stakeholders to engage in media strategies.

The outcome measures (based on analyses of the surveys in the test sites and control site) indicated the media strategy changed the targeted beliefs even with such a short implementation period (about 12 months).

Neither changes in behaviors (outcome measure) nor reductions in alcohol-related crashes (impact measure) were found. These results are consistent with previous efforts conducted by the Center for Health and Safety Culture in which behavior change often requires several years of intense messaging and is more likely to occur when supported by other strategies at the local level.

As a result of the evaluation, recommendations were made to improve the effectiveness of the strategy:

- Continue leveraging the existing positive norms at the community level that can provide energy to foster local coalitions to take additional steps to address traffic safety.
- Use highly targeted media to reach those most in a position to act. For example, use the media developed for placement in alcohol retail establishments in future efforts to address impaired driving.
- Invest more in local involvement and leverage the media to engage action and policy at the community level. This may require "seed" funding and/or partnerships with existing entities at the community level. Local stakeholders can use the media as a catalyst to promote family engagement, school or driver education programs, workplace safety programs, enforcement strategies, and local policy change.
- Shift from viewing media campaigns as only a tool for behavior change to viewing campaigns as a catalyst to support local efforts to address traffic safety thus resulting in sustained, long term change in traffic safety culture. While sustained media efforts can impact behavior, augmenting media strategies with local efforts using multiple strategies is more likely to result in greater and sustained change.

The results and recommendations were compiled in a report and presentation. The presentation was shared with key stakeholders including the public board that oversees the Idaho Transportation Department.

Conclusion

Reaching zero traffic-related deaths and serious injuries will require new thinking – including evaluative thinking. Evaluative thinking is a problem-solving approach to designing, selecting, and allocating resources for traffic safety strategies. It seeks credible evidence to provide answers about the effectiveness and sustainability of traffic safety strategies.

Traffic safety culture strategies focus on changing beliefs that influence behaviors related to traffic safety. For such strategies to become more widely used, we need more evidence that they are effective and more knowledge about how to implement them effectively.

Traffic safety practitioners can use process, outcome, and impact evaluations to grow evidence and knowledge. For evaluations to be useful, they must use quality data and make appropriate comparisons. Stakeholders should be involved in developing an evaluation. After developing a clear description of the strategy, quality data and appropriate comparisons can be identified for use in the evaluation.

Once evaluation results are gathered and analyzed, stakeholders should make meaning of the results, accumulate wisdom (i.e., lessons learned), and identify opportunities to apply the knowledge in the future.

References

1 Schwandt, T. A. (2018). Evaluative Thinking as a Collaborative Social Practice: The Case of Boundary Judgment Making. New Directions for Evaluation, 2018(158), 125–137.

2 Archibald, T. (2013). "Evaluative Thinking." Free Range Evaluation, WordPress, Retrieved July 27, 2020 from https://tgarchibald.wordpress.com/2013/11/11/18/.

3 American Evaluation Association. An Evaluation Roadmap for a More Effective Government. Washington, D.C. Retrieved on July 24, 2020 from: https://www.eval.org/evaluationroadmap

4 U.S. Department of Health and Human Services Centers for Disease Control and Prevention. Office of the Director, Office of Strategy and Innovation. Introduction to program evaluation for public health programs: A self-study guide. Atlanta, GA: Centers for Disease Control and Prevention. (2011). Retrieved from: https://www.cdc.gov/eval/guide/cdcevalmanual.pdf