FTA RESEARCH

FEDERAL TRANSIT ADMINISTRATION

# Considerations for Evaluating Automated Transit Bus Programs

DECEMBER 2019

FTA Report No. 0149
Federal Transit Administration

**PREPARED BY**

Joseph Luna, Elizabeth Machek, Sean Peirce
Advanced Vehicle Technology Division
John A. Volpe National Transportation Systems Center

U.S. Department of Transportation
**Federal Transit Administration**

## COVER PHOTO

*Courtesy of John A. Volpe National Transportation Systems Center*

## DISCLAIMER

*This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.*

# Considerations for Evaluating Automated Transit Bus Programs

DECEMBER 2019

FTA Report No. 0149

**PREPARED BY**

Joseph Luna, Elizabeth Machek, Sean Peirce
Advanced Vehicle Technology Division
John A. Volpe National Transportation Systems Center
U.S. Department of Transportation
55 Broadway, Cambridge, MA 02142

**SPONSORED BY**

Federal Transit Administration
Office of Research, Demonstration and Innovation
U.S. Department of Transportation
1200 New Jersey Avenue, SE
Washington, DC 20590

**AVAILABLE ONLINE**

*https://www.transit.dot.gov/about/research-innovation*

## Metric Conversion Table

| SYMBOL | WHEN YOU KNOW | MULTIPLY BY | TO FIND | SYMBOL |
|---|---|---|---|---|
| **LENGTH** | | | | |
| **in** | inches | 25.4 | millimeters | mm |
| **ft** | feet | 0.305 | meters | m |
| **yd** | yards | 0.914 | meters | m |
| **mi** | miles | 1.61 | kilometers | km |
| **VOLUME** | | | | |
| **fl oz** | fluid ounces | 29.57 | milliliters | mL |
| **gal** | gallons | 3.785 | liter | L |
| **$ft^3$** | cubic feet | 0.028 | cubic meters | $m^3$ |
| **$yd^3$** | cubic yards | 0.765 | cubic meters | $m^3$ |
| NOTE: volumes greater than 1000 L shall be shown in $m^3$ | | | | |
| **MASS** | | | | |
| **oz** | ounces | 28.35 | grams | g |
| **lb** | pounds | 0.454 | kilograms | kg |
| **T** | short tons (2000 lb) | 0.907 | megagrams (or "metric ton") | Mg (or "t") |
| **TEMPERATURE (exact degrees)** | | | | |
| **°F** | Fahrenheit | 5 (F-32)/9 or (F-32)/1.8 | Celsius | °C |

| REPORT DOCUMENTATION PAGE | *Form Approved*<br>*OMB No. 0704-0188* |
|---|---|

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

| 1. AGENCY USE ONLY | 2. REPORT DATE<br>December 2019 | 3. REPORT TYPE AND DATES COVERED<br>Final Report, December 2019 |
|---|---|---|

4. TITLE AND SUBTITLE
Considerations for Evaluating Automated Transit Bus Programs

5. FUNDING NUMBERS

6. AUTHOR(S)
Joseph Luna, Elizabeth Machek, Sean Peirce

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSE(ES)
John A. Volpe National Transportation Systems Center
U.S. Department of Transportation
55 Broadway
Cambridge, MA 02142

8. PERFORMING ORGANIZATION REPORT NUMBER
FTA Report No. 0149

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Department of Transportation
Federal Transit Administration
Office of Research, Demonstration and Innovation
East Building
1200 New Jersey Avenue, SE
Washington, DC 20590

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
FTA Report No. 0149

11. SUPPLEMENTARY NOTES [https://www.transit.dot.gov/about/research-innovation ]

12A. DISTRIBUTION/AVAILABILITY STATEMENT
Available from: National Technical Information Service (NTIS), Springfield, VA 22161.
Phone 703.605.6000, Fax 703.605.6900, email [orders@ntis.gov]

12B. DISTRIBUTION CODE
TRI-30

13. ABSTRACT
Given the potential of transit-bus automation, it is critical to evaluate the benefits and challenges from early implementations. A well-designed evaluation can quantify such societal benefits as improving travel time, increasing mobility, and raising transit ridership. This guide aims to assist transit stakeholders with designing and implementing evaluations of automated transit-bus programs. In designing evaluations, transit agencies and other stakeholders should identify program goals and audiences affected by the technology; develop a logic model that maps project inputs, activities, and outcomes; choose an appropriate evaluation design; and collect and analyze data on key performance indicators related to their program goals.

14. SUBJECT TERMS
Transit, bus, automation, technologies, research, demonstrations, evaluation, performance measures, key performance indicators

15. NUMBER OF PAGES
30

16. PRICE CODE

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | |

# TABLE OF CONTENTS



# LIST OF FIGURES

## ABSTRACT

Given the potential of transit-bus automation, it is critical to evaluate the benefits and challenges from early implementations. A well-designed evaluation can quantify such societal benefits as improving travel time, increasing mobility, and raising transit ridership. This guide aims to assist transit stakeholders with designing and implementing evaluations of automated transit-bus programs. In designing evaluations, transit agencies and other stakeholders should identify program goals and audiences affected by the technology; develop a logic model that maps project inputs, activities, and outcomes; choose an appropriate evaluation design; and collect and analyze data on key performance indicators related to their program goals.

# EXECUTIVE SUMMARY

Automated transit-bus technologies have great potential, but they can introduce uncertainties for transit agencies and the traveling public. To assess the impacts of automated transit-bus technologies and reduce the uncertainties, the transit industry will need to evaluate early transit bus automation projects/pilots/demonstrations and share the results. This document offers guidance for transit agencies' consideration in evaluating deployments of transit bus automation technologies.

Key findings from this report include the following:

- **Identify program goals and audience**. It is critical to identify transit program goals for deployment of automated transit buses. Such goals illustrate what a transit agency aims to accomplish and why the program is needed. Some goals for deploying an automated transit bus technology could include improving the operator's experience, enhancing mobility, and increasing safety. In addition to goals, agencies should identify the audiences who will be impacted by a project. Those impacted could include riders, persons with disabilities, motorists, agency staff, and local businesses.
- **Develop logic model.** After identifying program goals, it is helpful for agencies to develop a logic model. As described in this report, logic models summarize how a program's inputs and activities achieve intended goals. In addition to creating a logic model, agencies should also consider external factors that may affect a technology's deployment or observed outcomes. Such external factors can include changes in legislation and declines in the broader economy.
- **Choose evaluation design**. Program goals and the logic model inform the questions that an evaluation seeks to answer. Evaluation questions should be clear and specific, and the terms used in the questions should be readily defined and measurable. An evaluation design is the overall strategy used to answer evaluation questions. Case-study designs allow evaluators to explore issues in depth and are suitable for both qualitative and quantitative data gathering. However, case studies are typically limited to a small sample size. Statistical-analysis designs offer a variety of quantitative methods for identifying the ways in which a program led to its observed outcomes. However, care must be taken to explain the causal relationships (why did X lead to Y?) that inform statistical results.
- **Collect and analyze data.** Once an evaluation design is selected, evaluators should choose appropriate qualitative and quantitative methods for collecting and analyzing data. Such methods could include administering surveys and questionnaires, deploying roadside and in-vehicle sensors, examining agency records, and leading interviews and focus groups.
- **Additional considerations.** At the earliest possible stage, transit agencies should confirm with private-sector and other partners how data will be protected and shared. Such data may include commercially sensitive or

personally identifiable information that cannot be publicly shared. In addition, evaluation teams should ensure that they periodically validate data collection. Data validation ensures that problems can be fixed early with little impact to the final results.

This guide seeks to inform transit agency officials on how to think about and design an appropriate evaluation while also remaining aware of the constraints faced by agencies. The guide also emphasizes important considerations agencies should take with respect to validating data, protecting sensitive information, and developing communications plans.

SECTION

1

# Introduction

As described in the Strategic Transit Automation Research (STAR) Plan, the Federal Transit Administration (FTA) is sponsoring research and demonstrations of transit bus automation to help transit agencies, stakeholders, and industry make informed decisions. Given both the potential of transit automation and the unknowns associated with it, the benefits, challenges, and lessons learned from early demonstrations need to be evaluated and shared. A well-designed evaluation can quantify such societal benefits as improving travel time, reliability, and throughput; increasing mobility (spatial and temporal); enhancing safety; raising transit ridership; and saving money on operations and maintenance. Evaluation also demonstrates agency commitment to accountability and offers agencies the opportunity to engage the public and identify unforeseen areas for improvement. Ultimately, evaluation advances knowledge. As agencies share experiences with each other, the benefits and cost savings multiply. Evaluation and knowledge-sharing help agencies plan for future deployments and better position themselves to advocate for public-transportation funding.

With advances in technology and data gathering, it is expected that program evaluation will be conducted rigorously. This guide provides recommendations on designing and implementing a useful, effective evaluation of a transit bus automation project/pilot/demonstration to measure its impacts and record key lessons learned. However, this guide recognizes that transit agencies face time and budget constraints. This guide highlights important, general principles that can be applied to evaluations of various transit-automation projects. Given the number of factors that affect the quality of evaluations, FTA recommends planning evaluation activities from the start of a program. Figure 1-1 illustrates the recommended steps for designing and implementing an evaluation, and this guide explains each step. For a checklist of key evaluation components, please refer to Appendix A.

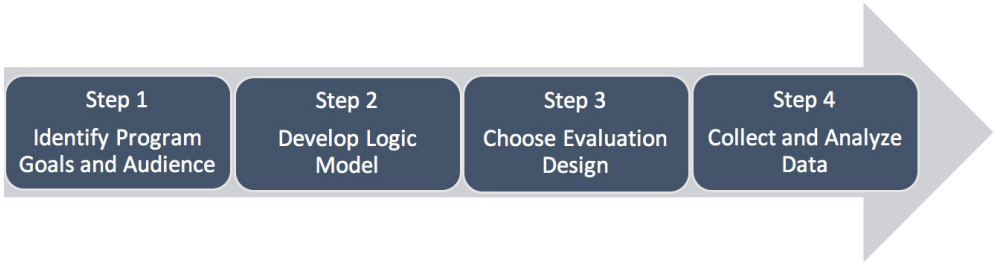



**Figure 1-1**

*Recommended steps for designing and implementing an evaluation*

SECTION

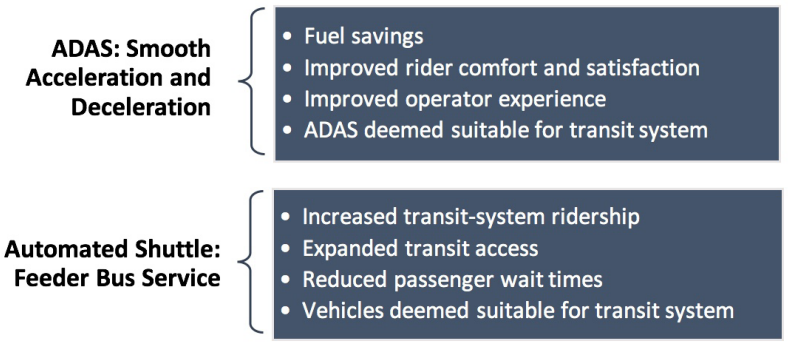
2

# Step 1: Identify Program Goals and Audience

It is crucial to identify first a transit program's goals for a given deployment. What does your program aim to accomplish? Why is this program needed? Figure 2-1 presents example program goals for deployments of a smooth acceleration and deceleration advanced driver-assistance system (ADAS) and automated feeder bus service.

**Figure 2-1**

*Example program goals for ADAS and automated shuttle deployments*




Identifying program goals establishes direction for a given program. Although goals may evolve during program implementation as a result of unforeseen circumstances, establishing clear goals at an early stage can guide any program changes.

Along with identifying program goals, it is crucial to pinpoint the audiences that will be impacted by a project. Who would benefit from the new technology? Who might be negatively impacted by the new technology? Potential audiences include:

- Users – regular riders (e.g., commuters), infrequent riders, persons with disabilities
- Non-users – motorists, pedestrians, bicyclists, local businesses
- Agency staff – drivers, managers and supervisors, maintenance, dispatchers, planners, unions

Listing potentially-impacted audiences at an early stage not only helps to clarify goals, but also identifies groups to interview to measure whether goals are being achieved.

SECTION

3

# Step 2: Develop Logic Model

Once program goals have been identified, it is useful for program managers, in conjunction with evaluators and other agency staff, if possible, to develop a logic model. Logic models summarize how a program achieves its goals; that is, how do a program's inputs and activities achieve the outcomes observed?

Figure 3-1 presents an example logic model.[1] As depicted, typical logic models consist of Inputs, which feed into Activities, which result in Outcomes (both short- and long-term).

**Figure 3-1**

*Example logic model*

| Inputs | Activities | Short-Term Outcomes | Long-Term Outcomes |
|---|---|---|---|
| • Staff<br>• Volunteers<br>• Time<br>• Money<br>• Materials<br>• Equipment<br>• Technology<br>• Partners | • Workshops<br>• Meetings<br>• Counseling<br>• Facilitation<br>• Assessments<br>• Product development<br>• Media work<br>• Recruitment<br>• Training | • Learning<br>• Awareness<br>• Knowledge<br>• Attitudes<br>• Skills<br>• Opinions<br>• Aspirations<br>• Motivations | • Behavior<br>• Practice<br>• Decisions<br>• Policies<br>• Social action<br>• Economic impacts<br>• Civic impacts<br>• Environment impacts |

- **Inputs** consist of the financial, organizational, and human resources that a program has available to meet its goals.
- **Activities** refer to what a program actually does. These activities could comprise new processes, research, tools, technology, events, outreach, and so forth. Activities help a program achieve its goals.
- **Outcomes** correspond to changes in knowledge and behavior of a program's target audiences. Short-term outcomes could include improved public awareness and operator acceptance of a new technology. Long-term outcomes could include improvements in safety, agency cost savings, and operational performance.[2]

---

[1]Logic model adapted from Government Accountability Office, 2012, "Designing Evaluations, 2012 Revision," GAO-12-208G. For further discussion of logic models and evaluation design, please refer to this document.

[2]These definitions for inputs, activities, and outcomes are adapted from W. K. Kellogg Foundation (2004), "Logic Model Development Guide," https://www.wkkf.org/resource-directory/resource/2006/02/wk-kellogg-foundation-logic-model-development-guide. For additional discussion of logic models, please refer to this document.

In addition to inputs, activities, and outcomes, it is also critical to consider the external factors that might affect a program's intended goals, e.g., changes in legislation, declines in the broader economy, harsh weather, etc.[3] For example, a weakening job market might reduce overall transit ridership, cancelling out ridership gains expected from deployments. As part of evaluation planning, program managers and evaluation staff should brainstorm and identify such external factors.

Taking the example of adopting AVs for feeder services, a program logic model could resemble (Figure 3-2):

**Figure 3-2**

*Example program logic model for AV deployment*

| Inputs | Activities | Short-Term Outcomes | Long-Term Outcomes |
|---|---|---|---|
| • Federal funds<br>• Local funds<br>• Agency expertise<br>• Lessons from prior deployers<br>• Automated vehicle technology<br>• Private and academic partnerships | • Public-relations campaign<br>• Agency workshops<br>• Staff training<br>• Technology deployment<br>• Data gathering<br>• Evaluation | • Improved public awareness<br>• Agency acceptance<br>• Proof of technological feasibility<br>• Enhanced performance database | • Public acceptance of transit AVs<br>• AVs deemed suitable for transit<br>• AV policy development<br>• Increased system ridership<br>• Expanded transit access |

The logic model depicted above includes several program goals in its outcome columns, such as proving suitability of new technology, increasing system ridership, and expanding transit access. A program's goals should be reflected in its program logic model.

[3]For a detailed discussion of outside issues and challenges facing deployments of automated transit technologies, please refer to Intelligent Transportation Systems Joint Program Office (2018), "Low-Speed Automated Shuttles: State of the Practice Final Report," https://rosap.ntl.bts.gov/view/dot/37060.

SECTION

4

# Step 3: Choose Evaluation Design

There are several components to consider when choosing an evaluation design, including evaluation questions, evaluation types (process and outcome), counterfactual scenarios, baseline data, and measures of effectiveness.

## Evaluation Questions and Evaluation Types

Drawing from a program's goals and its logic model, evaluators derive questions that an evaluation seeks to answer. These questions are similar to hypotheses that are tested in a scientific experiment. As examples, the following evaluation questions might be applicable to a test of ADAS, although the specific questions will need to be tailored to the nature of the deployment:

- Did ADAS-equipped buses save fuel relative to non-equipped buses on the same route?
- Did transit drivers use the ADAS as intended? Did they find them useful?
- Did ADAS reduce variability in headway times?
- How effective was the program's public engagement effort in terms of raising public understanding of the project?
- How effectively did the program respond to maintenance challenges during the pilot demonstration (alternative challenges: schedule, procurement, etc.)?

An evaluation question should be clear, specific, objective, and politically neutral; further, the terms in an evaluation question should be readily-defined and measurable, whether quantitatively or qualitatively (GAO 2012).[4] Evaluation questions should be linked to the audiences, activities, and goals laid out by a project.[5]

Evaluation questions that are ambiguously written or include multiple combinations of activities and outcomes can be complicated to measure and may yield misleading recommendations. Such questions to avoid could include, for example:

---

[4]See GAO (2012) for further discussion on developing evaluation questions. With respect to terms suitable for transit-related evaluation questions, this guide recommends referring to the National Transit Database (NTD) Glossary, https://www.transit.dot.gov/ntd/national-transit-database-ntd-glossary.

[5]For a short checklist on fine-tuning evaluation questions, please refer to Centers for Disease Control (2013), "Good Evaluation Questions: A Checklist to Help Focus Your Evaluation," https://www.cdc.gov/asthma/program_eval/assessingevaluationquestionchecklist.pdf.

- Did AVs change the passenger experience?
- Did ADAS reduce collisions and stress amongst transit drivers?
- Given schedule improvements resulting from AVs, does route ridership increase?

The first question is ambiguous and can invite multiple interpretations. What is meant by change? Would such change be positive or negative? Passenger experience can encompass a wide range. The second question is double-barreled: it incorporates two possibilities that may not necessarily be correlated. For instance, it is possible to reduce collisions with ADAS, but at the cost of increased stress on transit drivers if there are many false alarms. The third question is a leading question that could bias the interpretation of data; it assumes that schedule improvements would result from AVs, but such improvements could be influenced by a variety of other factors. As a result, observed ridership increases could be falsely attributed to schedule improvements due to AV technology.[6]

As evaluators derive evaluation questions, it is helpful to keep in mind two main types of program evaluations: process and outcome evaluations. Evaluation questions often fall into one of those two categories. Process evaluations examine the management and execution of a given program's activities. For example, questions on how a program manages its public-engagement activities and how a program responds to challenges during the pilot would pertain to a process evaluation. Process evaluations often occur while a program is in progress; as such, the findings of a process evaluation can be used to improve program management in real time. Alternatively, outcome evaluations focus on a program's outcomes to measure whether a program met its intended goals. Outcome evaluations typically occur once a program has completed its activities.[7] Although some programs can budget for multiple evaluation teams, many agencies can field only one evaluation for a program. For agencies that can field only one evaluation team, the team could include both process and outcome questions in its evaluation. Transit-automation technologies are still new, and there are important questions to address, such as whether a given technology works and whether it is suitable for transit applications. Therefore, evaluations of transit automation should be leveraged to contribute to this growing field of knowledge to ease future deployments.

## Counterfactual Scenarios and Baseline Data

Evaluations compare what a program has accomplished (or is accomplishing) to what would have happened had there been (a) no program at all (no-build

[6]For additional discussion on common evaluation and survey question mistakes, please see University of Sheffield, "Good Practice: Common Question Pitfalls," https://www.sheffield.ac.uk/apse/wp/wpevaluation/pitfalls.

[7]For further discussion, please see W. K. Kellogg Foundation (2004) and GAO (2012).

scenario) or (b) a different program in its place (next-best alternative). If an evaluation focuses only on the outcomes and benefits of a program, it would not be clear whether those outcomes or benefits would have been achieved without the program or perhaps with a less-costly alternative. Would a given transportation situation worsen in the absence of the proposed program? What would happen if a transit system continued to use conventional technologies? Comparing a project's benefits to a no-build scenario (also known as a "counterfactual") or next-best alternative not only provides a more accurate evaluation, but such comparison strengthens justification for further program support.

Once counterfactual or alternative scenarios are established, evaluation teams should assess what a given transportation situation looked like prior to program implementation.[8] Program managers and evaluation teams, consulting their established program goals, should identify the data needed for measuring a baseline transportation situation and changes to that situation that result from the program. The National Transit Database (NTD) Glossary can help program managers and evaluators in selecting consistent metrics for measuring baselines, since these metrics are already collected as part of NTD reporting. However, there may be other relevant metrics that go beyond NTD's scope.

## Measures of Effectiveness[9]

In identifying baseline data and monitoring changes to that data, program managers and evaluators are pinpointing measures of effectiveness (also known as key performance indicators or KPIs) to determine whether a program is meeting its stated goals. Changes in values of these KPIs can help an agency determine whether a particular investment in automated transit bus programs has "moved the needle" in such areas as customer satisfaction, safety, and so forth. Please refer to Appendix B for KPIs drawn from a variety of reports related to automated transit. Although these KPIs are offered as examples for reference, KPIs should be tailored to each technology deployment. Other potential measures/KPIs can be drawn from the NTD, the Transit Cooperative Research Program (TCRP), and the National Cooperative Highway Research Program (NCHRP).

## Evaluation Designs

An evaluation design is the overall strategy that is used to answer evaluation questions; data collection and analysis methods (discussed in the next section) are tactics used in executing the evaluation design. There are many different designs that a program evaluation can adopt. GAO (2012) notes that good evaluation design should be appropriate for the evaluation questions and context, adequately address the evaluation questions, fit available time and resources, and rely on

[8]Ideally, program managers should measure this situation prior to program implementation.
[9]As used in this document, measures of effectiveness are synonymous with metrics, performance measures, and KPIs.

sufficient, credible data.[10] Descriptions of a few common evaluation designs, adapted from GAO (2012), are summarized below. Many evaluations mix different designs.

- **Case studies** allow evaluators to explore issues in depth, from both qualitative and quantitative perspectives. Case studies are particularly suitable for process evaluations, but they are also relevant for outcome evaluations. Case studies typically have a smaller sample size—that is, they focus on only a few projects or components—than other evaluation designs, but allow for deeper study of each project or component. Because of their small sample size, case studies are generally not statistically representative, but nevertheless should be chosen carefully to ensure that there is representation across the relevant variables of interest.
- **Randomized experiments** are considered the ideal form of evaluation. In a randomized experiment, the "treatment" (program intervention, such as a funding grant) is assigned to participants (e.g., transit agencies, State transportation departments) randomly. This random assignment controls for any biases in the population that could affect outcomes. The group of participants that receives the program intervention would be known as the "treatment group," and those that did not receive the intervention would be called the "control group." Evaluators would then compare the outcomes observed for the treated group with the control group that did not receive the intervention to establish the effectiveness of a given program. Randomized experiments need to be designed and implemented at the start of a program, and random assignment can be difficult to implement in real-world transportation settings. Further, these experiments are time- and resource-intensive, so they often are not suitable for many scenarios.
- **Quasi-experiments** offer a compromise solution to randomization. In many quasi-experiments, the treatment is not assigned randomly. However, evaluators can compare those that have been affected by a program to a control group that is similar to the treatment group—but that have not been exposed to the treatment. For example, if a new transit technology is deployed on two routes, evaluators could compare outcomes on those two routes with two other routes (lacking the new technology) that have similar ridership, length, traffic patterns, etc., to establish a new technology's effect.[11]
- **Statistical analysis** offers another design possibility where randomized and quasi-experiments are not possible. Such analysis can be done via a variety of quantitative methods that describe the relationship between a program and its outcomes or that identify ways in which a program specifically led to outcomes. Although statistical analysis can demonstrate

[10]Government Accountability Office, 2012, "Designing Evaluations," https://www.gao.gov/assets/590/588146.pdf, accessed March 8, 2019.
[11]In this example, evaluators could compare a treated route with its own baseline (and counterfactual scenario) as well as with a similar route that was not treated with the program intervention.

different relationships amongst observed data, such analysis should be viewed cautiously because it can be difficult to establish how a data relationship was caused.[12]

[12]Due to this difficulty in establishing causal relationship, project sponsors can consider supplementing statistical analysis with case studies and/or qualitative data gathering to address causality.

SECTION

5

# Step 4: Collect and Analyze Data

Once an evaluation design (e.g., case studies, quasi-experiment, statistical analysis) is selected, evaluators should choose appropriate data-collection methods to assess a program's measures of effectiveness. Some typical methods include:

- Surveys and questionnaires – to assess perceptions of passengers and agency staff
- Sensors (e.g., on-vehicle LIDAR, cameras, roadside sensors) – to gather safety- and operations-related data
- Agency records – to capture impacts on transit run times, ridership, safety incidents, labor costs, and other elements that typically are recorded
- Interviews and focus groups – to ascertain in-depth opinions of passengers and agency staff

GAO (1993) notes that questionnaires are useful when a large amount of standardized information must be collected, when different sets of people are involved, and when those people are located in widely separated locations.[13] Questionnaires can collect a wide variety of information, from facts to statistics to opinions. However, evaluators must consider several elements to design a valid questionnaire. Has the survey sample been chosen in an unbiased manner? Have the questions been written appropriately? Please refer to GAO (1993) for resources on designing and deploying survey questionnaires.

Within questionnaires, a Likert-scale question is a common method for assessing the extent to which respondents agree with a given item. For example, a transit driver may be asked the following question with these Likert-scale responses:

- Overall, how satisfied or dissatisfied were you with the ADAS user interface? (Choose one.)
  - Very satisfied
  - Satisfied
  - Neither satisfied nor dissatisfied
  - Dissatisfied
  - Very dissatisfied

For Likert-scale questions, it is important to ensure that questions do not "lead" respondents toward one answer or another. Such questions must also be written clearly, avoiding language that would be confusing to a respondent.

[13]General Accounting Office (1993), "Developing and Using Questionnaires," https://www.gao.gov/assets/80/77270.pdf, accessed March 8, 2019.

Regardless of the data-collection method chosen, evaluators should take steps to reduce bias. For instance, in questionnaires respondents may give an answer that they think will please the interviewer (social-desirability bias). In other cases, respondents will lack knowledge sufficient to provide a response but will provide a response anyway. There are several methods for reducing respondent bias. One such method is conjoint analysis, where respondents review pairs of scenarios in which key criteria have been randomized. Respondents then indicate which of the paired scenarios they prefer. Such a design can allow researchers to determine which criteria are most important to respondents while reducing potential bias.[14] Anchoring vignettes are another method for reducing respondent bias; these are short, hypothetical stories that help "anchor" respondent responses to normative questions. Because respondents may have different definitions of how much they agree with a given item, anchoring vignettes normalize responses across respondents.[15] Depending on available time and budget, it is recommended that evaluation teams "pilot test" questionnaires and other data-collection methods with a small sample of respondents or experts. Such pilots can catch and mitigate response biases prior to full deployment.

In terms of analyzing data once it has been gathered, there are two main categories of analysis—descriptive and inferential. Descriptive analysis presents characteristics of data without necessarily discussing how those characteristics came about. For instance, such data characteristics as mean, median, range, variance, and mode are considered descriptive. Descriptive statistics are often visualized through histograms, line graphs, scatter plots, and various other graphics.

Inferential analysis, on the other hand, intends to establish causality—that is, how did a particular set of findings come about? Methods such as quasi-experiments and statistical regression are employed in inferential data analysis, but such methods should be used cautiously and with a strong understanding of confounding factors. Qualitative data gathering, such as through surveys and interviews, can help to establish a causal story on top of data analysis. Such qualitative data from surveys and interviews also can be statistically analyzed through content-analysis software.

---

[14]For a discussion of conjoint analysis, please see Hainmueller et al. (2013), "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments," *Political Analysis*.

[15]For discussion and further resources related to anchoring vignettes, please refer to "Anchoring Vignettes Overview," https://gking.harvard.edu/vign, accessed March 8, 2019.

SECTION

6

# Additional Considerations

## Periodic Data Validation

Regardless of whether data collected are quantitative or qualitative, evaluation teams should periodically analyze samples of their data during data collection to assess the extent of any missing or corrupted data. If such issues are caught early, data collection can be restarted or revised with little impact on the ultimate evaluation analysis. However, waiting until the end of data collection to analyze all data risks the surfacing of unforeseen problems that can negatively impact analysis. Further, agencies should continuously monitor the status of key equipment to ensure that data are captured correctly and that equipment is in good operating condition or fixed or replaced as necessary if it is malfunctioning. It is recommended that evaluation teams also develop a risk matrix of potential challenges to data collection and develop appropriate mitigation strategies (e.g., collect qualitative data when quantitative data are unavailable).

## Data-Sharing Protocol

Given that transit demonstrations of automated driver assistance systems and automated vehicles involve private-sector partners, transit agencies should negotiate a data-sharing protocol. Private-sector partners are likely to be collecting data through sensors and other means, and such data might have useful applications for transit agencies. However, private partners may view data as proprietary, and the release of data could put them at a competitive disadvantage relative to other companies operating in the automated transit space. As such, transit agencies should identify early on their data needs and discuss with private partners how that data can be shared, with appropriate protections for the private partner.[16]

## Data Protection

In addition to establishing a data-protection protocol, transit agencies should protect the data they gather and use. With surveys, interviews, and camera- and sensor-based data, participants may have concerns about personally identifiable information (PII). Who can access this data? How will identities be protected? To ensure reliable participation in data collection, transit agencies must demonstrate to their audiences that data will be kept confidential, such as through the generation of randomized identification numbers (to anonymize PII) and firewalled servers. Unauthorized data releases could present safety and

[16]For further discussion of the importance of establishing a data-sharing protocol, please see Intelligent Transportation Systems Joint Program Office (2018), "Low-Speed Automated Shuttles: State of the Practice Final Report," https://rosap.ntl.bts.gov/view/dot/37060.

security risks for a given project and harm an agency's reputation. It is essential for project and evaluation teams to be aware of any regulations that might pertain to data gathering, including the need for non-disclosure agreements, institutional review board (IRB) review, data agreements, protected data storage, PII protections, and so forth.

## Project Updates

In an ideal evaluation, a technology being evaluated would not change during the course of the intervention. Changes to a technology would complicate the causal chain of how an intervention achieves its goals and impacts society. Evaluators recognize that such an assumption is not always feasible in real-world environments. Should a program learn about critical safety or operational improvements over the course of a pilot, then a transit agency would be obliged to update its program to safeguard the public. However, for the purpose of an evaluation, program managers should maintain clear records of when hardware and software are updated—or operational or other practices changed—during a demonstration to identify which outcomes could be attributed to those program changes. Updated records provide essential qualitative information that allows evaluators to create a clear picture of how a technology, and updates to that technology, achieved an agency's goals.

## Communications Plan

Ultimately, an evaluation is only as good as its distribution. Evaluations provide lessons learned, and whereas those lessons learned do not necessarily have to be advertised to the public, they should be presented to key decision-makers to improve a program. Evaluations also generate important information for peer entities, and sharing that knowledge can advance technological innovations around the world. Finally, the information generated by evaluation is important for the public. With the rapid pace of technological change in transportation, many in the public are curious, excited, and apprehensive about changes to the status quo. Evaluation results, if well presented, can demonstrate potential benefits and invite further public engagement to improve the transportation enterprise.

APPENDIX

# A

# Evaluation Design and Implementation Checklist

The following checklist indicates the four key areas of evaluation design and implementation and includes questions within each area that evaluation teams can consider.

**Identify Program Goals and Audience**

- What is the program trying to accomplish?
- Have goals related to safety been identified? Operations? Mobility? Agency acceptance?
- Are there other important goals to consider?
- Have all potential audiences that would be affected (both positively and negatively) by the technology deployment been identified?

**Develop Logic Model**

- Have all program inputs, including costs, been listed? Activities? Short-term outcomes? Long-term outcomes?
- Does the logic model reflect the program's goals?
- Has the logic model been validated with the program's managers?

**Choose Evaluation Design**

- Have evaluation questions been derived from the logic model? Are the questions clear and objective?
- Have counterfactual scenarios been identified?
- What baseline data should be collected? Can that data continue to be collected over the evaluation period?
- Have clear measures of effectiveness been identified?
- Has the most appropriate evaluation design or combination of designs been selected to best answer the evaluation questions?

**Collect and Analyze Data**

- Have appropriate data-collection methods been chosen? Will these methods accurately assess the chosen measures of effectiveness?
- Have potential sampling biases been considered?
- Have data-collection methods been piloted to check for potential biases?
- Have appropriate data-analysis techniques been chosen?
- Is the data being periodically validated?
- Have data risks been considered and mitigation strategies developed?

APPENDIX

B

# Sample Key Performance Indicators (KPIs)

This appendix presents example measures of effectiveness from FTA-sponsored evaluations and considerations from other sources.

## Publications

**Federal Transit Administration (2016), "Vehicle Assist and Automation (VAA) Demonstration Evaluation Report," FTA Report No. 0093**

This report summarizes an evaluation of vehicle assist and automation (VAA) technologies deployed by Lane Transit District in Eugene, Oregon, for its Emerald Express Bus Rapid Transit. The demonstration used magnetic sensors for precision docking at three stations and lane guidance between stations. KPI areas include:

- Bus driver satisfaction
- Customer satisfaction – rating of ride quality, rating of precision docking
- Efficiency/productivity
- Technical performance
- Maintenance
- Safety

For further information, please see: https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/FTA_Report_No._0093.pdf

**Federal Transit Administration (2019), "Driver Assist System (DAS) Technology to Support Bus-on-Shoulder Operations," FTA Report No. 0135**

This report summarizes project activities and results of the Generation 2 DAS deployed by the Minnesota Valley Transit Authority (MVTA) for bus shoulder operations. The system provides warnings for lane departure, side collision, and forward collision. KPI areas include:

- Route system performance
- Customer satisfaction
- Bus operator satisfaction
- Maintenance
- Safety

For further information, please see: https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/research-innovation/132941/driver-assist-system-technology-support-bus-shoulder-operations-ftareport0135.pdf

### Innamaa, Satu, Scott Smith, and Salla Kuisma (2018), "Key Performance Indicators (KPIs) for Assessing the Impacts of Automation in Road Transportation"

This paper presents survey results that measured expert views on the importance of various KPIs to understanding the impact of automation in road transportation. KPI areas include:

- Vehicle operations
  - Number of instances where the driver must take manual control/1000km or miles
  - Mean and maximum duration of the transfer of control between operator/driver and vehicle when requested by the vehicle
  - Mean and maximum duration of the transfer of control between operator/driver and vehicle when turning automated driving system on/off (manual override)
- Use of automated driving
  - Number of instances where the driver must take manual control/1000km or miles
  - Use of automated driving functions (% of km of maximum possible use)
  - Comprehensibility of user interface (expressed on a Likert scale, e.g. 1–9, low–high)
- Safety
  - Number of crashes (distinguishing property damage, and crashes with injuries and fatalities), in total and per 100 million km or miles
  - Number of instances where the driver must take manual control/1000 km or miles
  - Number of conflicts encountered where time-to-collision (TTC) is less than a pre-determined threshold/100 million km or miles
- Energy or environment
  - Energy consumption of a vehicle (liters/100km or miles per gallon or electric equivalent)
  - Tailpipe carbon dioxide ($CO_2$) emissions in total per year and per vehicle-km or mile
  - Tailpipe criteria pollutant emissions (NOX, CO, PM10, PM2.5, VOC) in total per year and per vehicle-km or mile
- Personal mobility
  - Type and duration of in-vehicle activities when not operating the vehicle (high levels of automation)
  - User perceptions of travelling quality (expressed on a Likert scale, e.g., 1–9, low–high)

  - User perceptions of travelling reliability (expressed on a Likert scale, e.g., 1–9, low–high)
- Travel behavior
  - Share of transport modes (modal split) per week (based on number of trips)
  - Number and type of trips per week (in total and per inhabitant)
  - Total duration of trips per week (in total and per inhabitant)
- Network efficiency
  - Throughout, i.e., number of vehicles per hour through a particular road section or intersection approach, normalized to number of lanes and proportion of green time (where relevant)
  - Maximum road capacity (for a given road section)
  - Peak period travel time along a route
- Asset management
  - Vehicle-to-infrastructure (V2I) infrastructure for automation
  - Frequency of pothole occurrence (number of potholes per 100km or miles)
  - Use of hard shoulder (for hard-shoulder running or as emergency stop area for mal-functioning automated vehicles)
- Costs
  - Capital cost per vehicle for the deployed system (infrastructure, monetary value)
  - Cost of purchased automated vehicle (market price, monetary value)
  - Operating cost for the deployed system (per vehicle-hour or per vehicle-km or mile, monetary value)
- Public health
  - Modal share (%) and total mileage travelled (kms) by active modes of transportation (walking and bicycle)
  - Number of fatalities and injuries per year per million inhabitants
  - Proportion of people with improved access to health services
- Land use
  - Number of parking slots
  - Density of housing
  - Location of parking
- Economic impacts
  - Socio-economic cost benefit ratio
  - Work time lost from traffic crashes (hours per year, overall and per capita; monetary value)

For further information, please see: https://connectedautomateddriving.eu/wp-content/uploads/2018/03/KPS-for-Assessing-Impact-CAD_VTT.pdf

### Federal Transit Administration and Intelligent Transportation Systems (ITS) Joint Program Office, Mobility on Demand (MOD) Sandbox Demonstrations Independent Evaluation (IE)

The FTA and ITS Joint Program Office sponsored an independent evaluation of the FTA MOD Sandbox Demonstration projects. Although they do not include transit bus automation, the projects, like automation, include enabling technologies and innovative approaches to improve public transportation. The IE is comprehensive and includes a broad range of measures of effectiveness, which may be considered for and could be applied to transit bus automation projects and programs. Performance measure categories in the MOD Sandbox IE include, but are not limited to, the following:

- Traveler behavior
- System performance
- Capital, operating, and user costs
- Accessibility
- Impacts (e.g., benefits) on disadvantaged populations (e.g., persons with disabilities, low income and unbanked/underbanked populations)
- User experience/satisfaction

For further information, please see: https://www.transit.dot.gov/research-innovation/mobility-demand-mod-sandbox-program

## Other Considerations

Appropriate data will vary according to the scope and goals of a particular project. Participating organizations should consider the feasibility of obtaining and sharing data in evaluating their transit bus automation project/pilot/demonstration to measure its impacts. The following is an illustrative list of possible data types and elements that may be considered:

- **Vehicle performance data** with respect to operations and maintenance (e.g., dwell time, total service provided (vehicle-miles/vehicle-hours), percentage of automated vs manual operation, fuel efficiency, battery life, emissions, travel times, and average vehicle speed)
- **Automation component and system data** (e.g., number and types of sensors and actuators, human-machine interface [HMI] design, confidence information in object detection and classification)
- **Safety data** (e.g., notifications, disengagements, emergency driver takeover, incidents, and edge cases/near-misses, rules-of-the-road compliance; boarding and alighting incidents)
- **Costs** (e.g., vehicle procurement, operation, maintenance, storage; infrastructure improvements; labor and training; other ongoing operational costs; vehicle out-of-service time)

- **Mobility impacts** (e.g., passenger counts, percentage of scheduled trips completed, on-time arrival, average passenger wait time, and major origin-destination patterns, and rider demographics)
- **Human factors** (e.g., on-board attendant experience and alertness, customer experience and satisfaction, accessibility metrics, and passenger safety metrics)
- **Data from ancillary systems** that support non-driving bus operator functions (e.g., fare collection, ramp deployment and retraction, wheelchair securement, occupant detection, and passenger information assistance)
- **Infrastructure and system performance data** (e.g., vehicle-to-infrastructure [V2I] communications and equipment, congestion, average traffic speed)
- **Cybersecurity** (e.g., cybersecurity assessments [which may include threat analysis and risk assessment (TARA) results, cybersecurity mitigation measures, penetration testing results, etc.], incident frequency and response time, vulnerability data, staffing/training levels)

# REFERENCES


Centers for Disease Control (2013). "Good Evaluation Questions: A Checklist to Help Focus Your Evaluation," https://www.cdc.gov/asthma/program_eval/assessingevaluationquestionchecklist.pdf.

Federal Transit Administration (2016). "Vehicle Assist and Automation (VAA) Demonstration Evaluation Report." FTA Report No. 0093, https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/FTA_Report_No._0093.pdf.

Federal Transit Administration (2019). "Driver Assist System (DAS) Technology to Support Bus-on-Shoulder Operations Evaluation Report." FTA Report No. 0135, https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/research-innovation/132941/driver-assist-system-technology-support-bus-shoulder-operations-ftareport0135.pdf.

Federal Transit Administration (2019). "National Transit Database Glossary," https://www.transit.dot.gov/ntd/national-transit-database-ntd-glossary.

General Accounting Office (1993). "Developing and Using Questionnaires," https://www.gao.gov/assets/80/77270.pdf.

Government Accountability Office (2012). "Designing Evaluations, 2012 Revision," GAO-12-208G, https://www.gao.gov/assets/590/588146.pdf

Hainmueller, Jens, Daniel Hopkins, and Teppei Yamamoto (2013). "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments," *Political Analysis*, 22(1): 1-30.

Innamaa, Satu, Scott Smith, and Salla Kuisma (2018). "Key Performance Indicators (KPIs) for Assessing the Impacts of Automation in Road Transportation," https://connectedautomateddriving.eu/wp-content/uploads/2018/03/KPS-for-Assessing-Impact-CAD_VTT.pdf.

Intelligent Transportation Systems Joint Program Office (2018). "Low-Speed Automated Shuttles: State of the Practice Final Report," https://rosap.ntl.bts.gov/view/dot/37060.

King, Gary (n.d.). "Anchoring Vignettes Overview," https://gking.harvard.edu/vign.

University of Sheffield (n.d.). "Good Practice: Common Question Pitfalls," https://www.sheffield.ac.uk/apse/wp/wpevaluation/pitfalls.

W. K. Kellogg Foundation (2004). "Logic Model Development Guide," https://www.wkkf.org/resource-directory/resource/2006/02/wk-kellogg-foundation-logic-model-development-guide.

U.S. Department of Transportation

**Federal Transit Administration**

U.S. Department of Transportation
Federal Transit Administration
East Building
1200 New Jersey Avenue, SE
Washington, DC 20590
*https://www.transit.dot.gov/about/research-innovation*