Evaluating the Effectiveness of Stormwater Education and Outreach: Permittee Guidance for Addressing Challenges through Behavior Change

Evaluation Guidance Manual

January 12, 2023

Prepared by:

Evergreen StormH2O
PO Box 19812
Spokane, WA 99228

SAM Stormwater Action Monitoring
Washington State University
Washington Stormwater Center Assistance · Research · Training
Evergreen StormH2O

ACKNOWLEDGEMENTS

This Manual was made possible by the Stormwater Action Monitoring (SAM) program, funded by Permittees and administered by the Washington State Department of Ecology (Ecology). The Manual is one of several deliverables developed for the following project: *Evaluating the Effectiveness of Stormwater Education and Outreach: Permittee Guidance for Addressing Challenges through Behavior Change*. The Manual authors are noted below. Special thanks to McKenna Lindberg, Osborn Consulting, Inc., and Nicole Chen, EIT, Port of Seattle, for their supporting contributions to the Manual development.

Aimee Navickis-Brasch, PhD, PE
Evergreen StormH2O

Francesca White, PE
Evergreen StormH2O

TABLE OF CONTENTS

1.0 Introduction

1.1 Manual Purpose and Background

In Washington State, discharges from Municipal Separate Storm Sewer Systems (MS4s) are regulated under a combined National Pollutant Discharge Elimination System (NPDES) and State Waste Discharge General Permit (MS4 Permit). The MS4 Permits require Permittees to implement a Stormwater Management Program (SWMP) that includes an Education and Outreach (E&O) Program that is designed to meet specific goals. These goals vary depending on whether the permit is issued to a Western Washington (WWA) Phase I or Phase II jurisdiction or an Eastern Washington (EWA) Phase II jurisdiction. **Table 1-1** Provides a summary of the goals based on the different types of permits.

Table 1-1 MS4 E&O Program Design

Education and Outreach in MS4 Permit	WWA Phase I	WWA Phase II	EWA Phase II
Build general awareness about methods to address and reduce stormwater runoff.	✓	✓	
Effect behavior change to reduce or eliminate behaviors and practices that cause or contribute to adverse stormwater impacts.	✓	✓	
Create stewardship opportunities that encourage community engagement in addressing the impacts from stormwater runoff.	✓	✓	
Educate target audiences about the impacts of stormwater discharges to water bodies and the steps to take to reduce pollutants in stormwater.			✓

In the 2019–2024 WWA and EWA MS4 Permits, all three included requirements for evaluating and reporting on behavior change campaigns (WWA) and E&O programs (EWA) for changes in the understanding and adoption of target behaviors. The specific requirements vary between the WWA and EWA Permits, as shown in **Figure 1-1** and **Figure 1-2**. Most notably, the WWA Permits explicitly state to "*follow social marketing practices and methods, similar to community based social marketing (CBSM) and develop a campaign that is tailored to the community, including development of a program evaluation plan. Each Permittee shall: develop a strategy and schedule (a) to more effectively implement the existing campaign or; (b) to expand the existing campaign to a new target audience or BMP; or (c) for a new target audience and BMP behavior change campaign"* [WWA Phase I S5.C.11.a.iv. and Phase II S5.C.2.a.ii.(c)]. In contrast, the EWA Permit does not mention social marketing or CBSM [EWA Phase II S5.B.1.]. However, in all three permits, Permittees have similar requirements for evaluating and reporting on the understanding and adoption of targeted behaviors.

The specific WWA Permit language [Phase I S5.C.11.a.vi. and Phase II S5.C.2.a.ii.(e)] is as follows:

No later than March 31, 2024, evaluate and report on:

1. *The changes in understanding and adoption of targeted behaviors resulting from the implementation of the strategy; and*

2. *Any planned or recommended changes to the campaign in order to be more effective; describe the strategies and process to achieve the results.*

The purpose of this Manual is to assist Permittees with meeting the evaluation requirement, which is due on March 31, 2024, for WWA and was due on December 31, 2021, for EWA. While the EWA MS4 Permit deadline passed before this document was developed, this resource may still be useful for EWA Permittees to meet future E&O MS4 Permit requirements.

This Manual describes professionally recommended approaches and concepts that can be used to conduct an evaluation of the behavior change campaign component of the E&O requirements; however, it is not required that this Manual be used to meet MS4 Permit requirements. It was designed to provide guidance for any size project. Some projects may use information from every chapter to conduct a successful evaluation, and some projects may not need that much detail to conduct a successful evaluation. **For example**, small projects that use photograph comparisons or observational data will likely use very little math and may not find Chapter 5 useful, whereas larger projects that collect survey data may find Chapter 5 useful to analyze their data and support their results. While this Manual provides resources about social marketing and CBSM (Social Marketing and Community-Based Social Marketing Resources [Section 1.4]), it does not provide guidance for developing a behavior change campaign.

Note: MS4 Permit language has been included in this document to provide the reader with context for why this Manual was developed. There are slight variations between the WWA MS4 Phase I and Phase II Permits, and the MS4 Permits are updated and reissued every five years. Please refer to the current version of the MS4 Permit that applies to your jurisdiction for exact permit language.

1.2 Manual Organization

The following provides information about the content of each chapter in this document, and each chapter has examples of how to apply the approaches and concepts described. Chapters 2 through 5 are in the order typically followed when an evaluation plan is developed. Ideally, an evaluation plan should be developed prior to implementing a campaign so that all the data can be developing during the campaign.

- **Chapter 1 Introduction** (this chapter) introduces the reader to the Manual and describes relevant permit requirements and the Manual's purpose and organization. It also provides additional resources that were developed as part of this project, as well as information about social marketing and CBSM resources.
- **Chapter 2 Sample Size Selection** provides an overview of common methods for selecting a minimum sample size to evaluate the campaign.
- **Chapter 3 Evaluation Instruments** provides tools (surveys, observational checklists, etc.) used to measure the target audience's change in the understanding and adoption of the targeted behavior campaign for the E&O program. It also provides information about different evaluation instruments, including suggestions for selecting and validating instruments.
- **Chapter 4 Data Types** introduces the different types of data and provides guidance for organizing both qualitative and quantitative data in preparation for data analysis.
- **Chapter 5 Analysis Methods** provides an overview of common data analysis methods for both qualitative and quantitative data. Discussion about the values of hypothesis testing is also included.

1.3 Additional Resources

As part of the SAM project, two additional tools were developed to assist Permittees with meeting their E&O requirements for evaluating behavior change campaigns. These resources include a report template and a website. The following provides additional details about these resources.

1.3.1 *Behavior Change Campaign Evaluation Report Template*

A report template was developed to support Permittees in meeting their E&O MS4 Permit requirements for reporting on the evaluation of understanding and adoption of targeted behaviors. It is recommended that the report template be used in tandem with this Manual. The report template was developed to streamline report writing by identifying what information is required by the MS4 Permits and providing suggestions for content, which includes informal suggestions from Ecology as basic information to include in a report. The suggestions for content are included because they are common steps in the evaluation process and provide a more complete story of this process. However, all that is required to be submitted to Ecology is what is written in the MS4 Permits, which were written to provide Permittees with flexibility for reporting their process and results.

1.3.2 *Website – Tools and Resources for Behavior Change Programs*

The website waterbehaviorchange.org was created to provide tools and resources for behavior change programs. It is recommended that the website be used as a companion to this document to help jurisdictions assess the effectiveness of existing campaigns around the country. The website compiles every known evaluation of a behavior change campaign in stormwater or water quality. The site provides details on campaign implementation and evaluation and rates the research quality of the evaluation.

1.4 Social Marketing and Community-Based Social Marketing Resources

Mention of social marketing and community-based social marketing (CBSM) first appeared in the 2019–2024 WWA MS4 Permits as methods for developing behavior change campaigns. Social marketing has been around since the early 1970s (Social Marketing Services, Inc., 2008) and is used to promote behavior change that improves public health and prevents injuries. Lee & Kotler define social marketing as *"a process that applies marketing principles and techniques to create, communicate, and deliver value in order to influence target audience behaviors that benefit society (public health, safety, the environment, and communities) as well as the target audience*" (Lee & Kotler, 2011). CBSM blends social marketing with social sciences (social and environmental psychology) and draws from the concept that sustainable behavior change is most effectively achieved when it involves direct contact with people and initiatives delivered at the community level (McKenzie-Mohr, 2011).

Providing guidance for developing a behavior change campaign is beyond the scope of this Manual. Instead, resources below provide more information about social marketing and CBSM . Additional resources provided by Technical Advisory Committee (TAC) members are included in this section. These resources are organized into the following categories: professional organization, practices and methods papers, informational resources (not tied to a for-profit entity), and informational resources that are created and maintained by for-profit subject-matter specialists/practitioners.

1.4.1 *Professional Organizations:*

- **Pacific Northwest Social Marketing**, https://pnsma.org/ – Professional organization made up of members of the social marketing community across the Pacific Northwest
    - Learning forums and events
    - SPARKS Conference (annual)
    - Resources

- **Social Marketing Association of North America** (SMANA), https://smana.org/ – U.S.-based social marketing association serving Central and North America
    - Listserv
    - Webinars, events, training opportunities
    - Conferences
    - Resources & Guiding Principles
    - Social Marketing Quarterly (peer-reviewed journal) for members

- **International Social Marketing Association** (iSMA), https://isocialmarketing.org/
    - Webinars, events, trainings, and news

1.4.2 Practices and Methods Papers:

- **Consensus Definition of Social Marketing**, http://smana.org/wp-content/uploads/2017/04/iSMA-Consensus-definition-of-Social-Marketing-Oct-2013.pdf

- **Global Consensus on Social Marketing Principles, Concepts and Techniques**, http://smana.org/wp-content/uploads/2017/04/ESMA-AASM-SMANA-endorsed-Consensus-Principles-and-concepts-paper.pdf

- **Social Marketing Evidence of Effectiveness, 2018**, http://smana.org/wp-content/uploads/2018/11/Final-List-of-Key-Social-Marketing-Evidence-of-Effectiveness-citations-Nov-2018.pdf

- **DRAFT Social Marketing Statement of Ethics** (issued for consultation in September 2022, with end of consultation February 15, 2023), https://docs.google.com/document/d/1pCpW15DPyL5a5-D9Ip2x1-gOA_Kr-znK/edit

1.4.3 Informational Resources:

- **Social Marketing: Messaging for Behavior Change**, https://www.epa.gov/system/files/documents/2022-11/EPA%20Social%20Marketing%20Training-%2010.25.22%20FINAL.pdf
  150-page module developed for U.S. Environmental Protection Agency

- **Creating Messages that Drive Behavior Change,** https://www.epa.gov/recyclingstrategy/creating-messages-drive-behavior-change, U.S. Environmental Protection Agency
    - What is Social Marketing?
    - Learn How to Create a Social Marketing Program (recorded presentations, presentation slides, guide, and worksheet)

- **Getting Your Feet Wet with Social Marketing,** https://fyi.extension.wisc.edu/wateroutreach/files/2015/12/GettingYourFeetWet1.pdf, A Social Marketing Guide for Watershed Programs, Jack Wilbur, Utah Dept of Agriculture & Food – Designed to walk watershed groups, municipalities, etc., through the process of developing and implementing a watershed outreach campaign, including stormwater focus.

- **Chesapeake Behavior Change** https://www.chesapeakebehaviorchange.org/ Serves as a repository to publish behavior change campaigns and view other organizations' campaigns to encourage collaboration and learning opportunities.

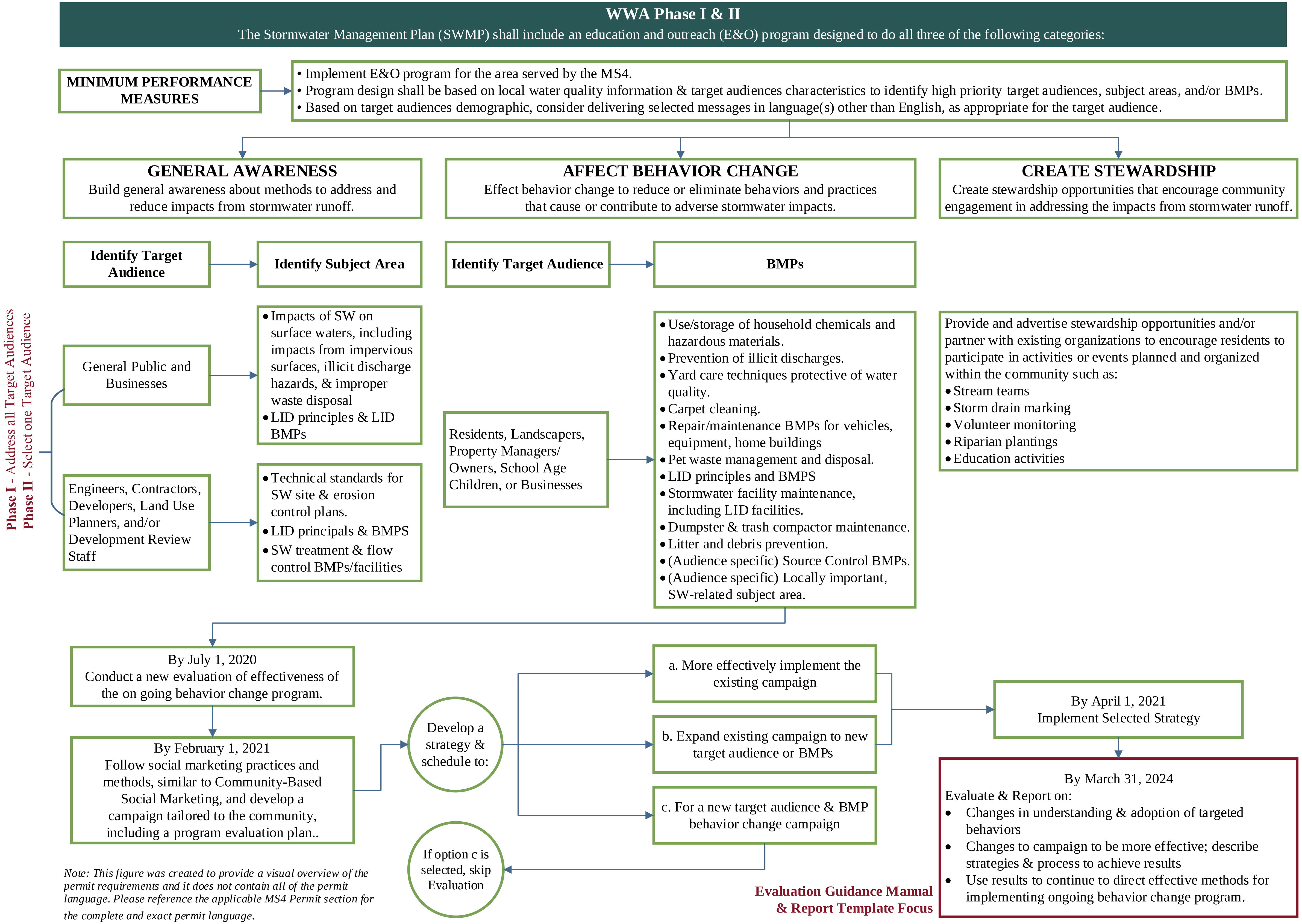
1.4.4 Informational Resources Created and Maintained by For-Profit Subject Matter Specialists/Practitioners:

- **Social Marketing Primer and Step-by-Step Guide and Workbook,** https://cplusc.com/social-marketing-workbook/, Downloadable PDFs created by C+C
  - **Social Marketing Primer,** https://cplusc.com/wp-content/uploads/2022/10/CC-Social-Marketing-Primer.pdf, [linked on EPA's site]
  - **Planning for Effective Social Marketing Campaigns: A Step-by-Step Guide and Workbook**, https://cplusc.com/wp-content/uploads/2021/09/CC-Social-Marketing-Workbook.pdf

- **Tools of Change,** https://toolsofchange.com/en/home/, – Collection of behavior change, social marketing, and community-based social marketing case studies
  - Planning guide
  - Case studies
  - Topic resources
  - Webinars and workshops

- **Community-Based Social Marketing**, https://cbsm.com/ – CBSM founder, Doug McKenzie-Mohr's site includes:
  - Upcoming training opportunities listed on site
  - Resource links to articles, reports, cases, forums, and colleagues
  - Publications: *Fostering Sustainable Behavior* (Third Edition), Doug McKenzie-Mohr, PhD

- **Social Marketing Services, Inc.,** https://www.socialmarketingservice.com/ – Nancy Lee, author of 15 books on social marketing. This website site includes links to:
  - Planning worksheets (free, downloadable PDF provides step-by-step guide through the process of creating a social marketing plan)
  - Journal articles
  - Upcoming learning opportunities are listed on iSMA, SMANA, and Nancy's website.
  - Publications co-authored by Nancy Lee and Philip Kotler:
    - Success in Social Marketing: 100 Case Studies from Around the Globe (2022)
    - Social Marketing: Changing Behaviors for Social Good (Sixth Edition, 2023)
    - Social Marketing to Protect the Environment: What Works (2011)
    - Marketing in the Public Sector: A Roadmap for Improved Performance (2006)



WWA Phase I & II

The Stormwater Management Plan (SWMP) shall include an education and outreach (E&O) program designed to do all three of the following categories:

MINIMUM PERFORMANCE MEASURES

- Implement E&O program for the area served by the MS4.
- Program design shall be based on local water quality information & target audiences characteristics to identify high priority target audiences, subject areas, and/or BMPs.
- Based on target audiences demographic, consider delivering selected messages in language(s) other than English, as appropriate for the target audience.

GENERAL AWARENESS

Build general awareness about methods to address and reduce impacts from stormwater runoff.

AFFECT BEHAVIOR CHANGE

Effect behavior change to reduce or eliminate behaviors and practices that cause or contribute to adverse stormwater impacts.

CREATE STEWARDSHIP

Create stewardship opportunities that encourage community engagement in addressing the impacts from stormwater runoff.

Phase I - Address all Target Audiences

Phase II - Select one Target Audience

Identify Target Audience

Identify Subject Area

General Public and Businesses

- Impacts of SW on surface waters, including impacts from impervious surfaces, illicit discharge hazards, & improper waste disposal
- LID principles & LID BMPs

Engineers, Contractors, Developers, Land Use Planners, and/or Development Review Staff

- Technical standards for SW site & erosion control plans.
- LID principals & BMPS
- SW treatment & flow control BMPs/facilities

Identify Target Audience

BMPs

Residents, Landscapers, Property Managers/ Owners, School Age Children, or Businesses

- Use/storage of household chemicals and hazardous materials.
- Prevention of illicit discharges.
- Yard care techniques protective of water quality.
- Carpet cleaning.
- Repair/maintenance BMPs for vehicles, equipment, home buildings
- Pet waste management and disposal.
- LID principles and BMPS
- Stormwater facility maintenance, including LID facilities.
- Dumpster & trash compactor maintenance.
- Litter and debris prevention.
- (Audience specific) Source Control BMPs.
- (Audience specific) Locally important, SW-related subject area.

Provide and advertise stewardship opportunities and/or partner with existing organizations to encourage residents to participate in activities or events planned and organized within the community such as:

- Stream teams
- Storm drain marking
- Volunteer monitoring
- Riparian plantings
- Education activities

By July 1, 2020
Conduct a new evaluation of effectiveness of the on going behavior change program.

By February 1, 2021
Follow social marketing practices and methods, similar to Community-Based Social Marketing, and develop a campaign tailored to the community, including a program evaluation plan..

Develop a strategy & schedule to:

a. More effectively implement the existing campaign

b. Expand existing campaign to new target audience or BMPs

c. For a new target audience & BMP behavior change campaign

If option c is selected, skip Evaluation

By April 1, 2021
Implement Selected Strategy

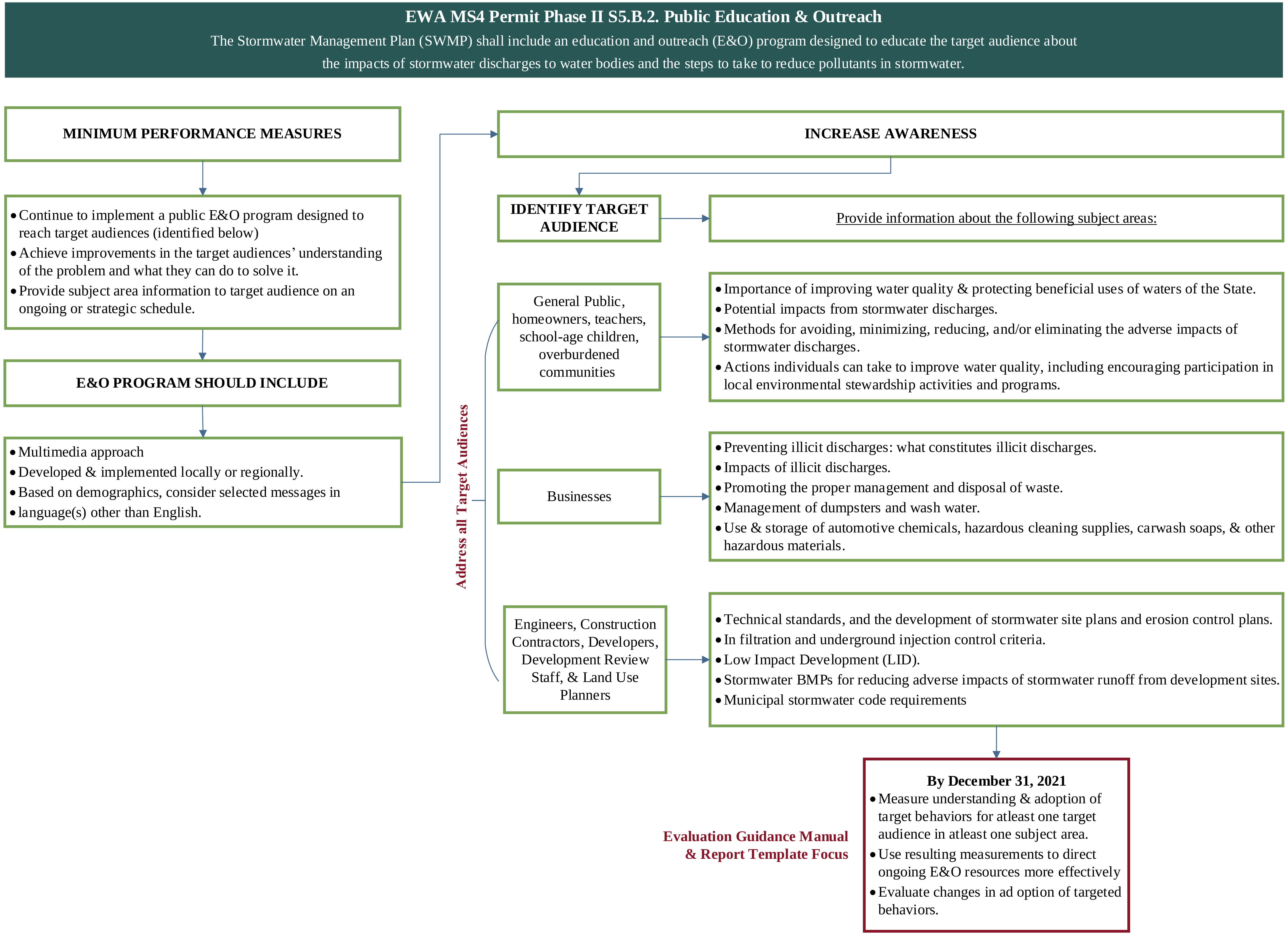
By March 31, 2024
Evaluate & Report on:

- Changes in understanding & adoption of targeted behaviors
- Changes to campaign to be more effective; describe strategies & process to achieve results
- Use results to continue to direct effective methods for implementing ongoing behavior change program.

Evaluation Guidance Manual & Report Template Focus

Note: This figure was created to provide a visual overview of the permit requirements and it does not contain all of the permit language. Please reference the applicable MS4 Permit section for the complete and exact permit language.

Figure 1-1 WWA Phase I & II Illustration of MS4 E&O Program Requirements



EWA MS4 Permit Phase II S5.B.2. Public Education & Outreach

The Stormwater Management Plan (SWMP) shall include an education and outreach (E&O) program designed to educate the target audience about the impacts of stormwater discharges to water bodies and the steps to take to reduce pollutants in stormwater.

MINIMUM PERFORMANCE MEASURES

- Continue to implement a public E&O program designed to reach target audiences (identified below)
- Achieve improvements in the target audiences' understanding of the problem and what they can do to solve it.
- Provide subject area information to target audience on an ongoing or strategic schedule.

E&O PROGRAM SHOULD INCLUDE

- Multimedia approach
- Developed & implemented locally or regionally.
- Based on demographics, consider selected messages in
- language(s) other than English.

INCREASE AWARENESS

IDENTIFY TARGET AUDIENCE

Provide information about the following subject areas:

Address all Target Audiences

General Public, homeowners, teachers, school-age children, overburdened communities

- Importance of improving water quality & protecting beneficial uses of waters of the State.
- Potential impacts from stormwater discharges.
- Methods for avoiding, minimizing, reducing, and/or eliminating the adverse impacts of stormwater discharges.
- Actions individuals can take to improve water quality, including encouraging participation in local environmental stewardship activities and programs.

Businesses

- Preventing illicit discharges: what constitutes illicit discharges.
- Impacts of illicit discharges.
- Promoting the proper management and disposal of waste.
- Management of dumpsters and wash water.
- Use & storage of automotive chemicals, hazardous cleaning supplies, carwash soaps, & other hazardous materials.

Engineers, Construction Contractors, Developers, Development Review Staff, & Land Use Planners

- Technical standards, and the development of stormwater site plans and erosion control plans.
- In filtration and underground injection control criteria.
- Low Impact Development (LID).
- Stormwater BMPs for reducing adverse impacts of stormwater runoff from development sites.
- Municipal stormwater code requirements

Evaluation Guidance Manual & Report Template Focus

By December 31, 2021

- Measure understanding & adoption of target behaviors for atleast one target audience in atleast one subject area.
- Use resulting measurements to direct ongoing E&O resources more effectively
- Evaluate changes in ad option of targeted behaviors.

Figure 1-2 Illustration of EWA MS4 Permit E&O Requirements

2.0 Sample Size Selection

2.1 Chapter Overview

Careful consideration and selection of an appropriate sample size is an important early step in the evaluation planning process. This is because sample size can influence the study design, including the choice of evaluation instruments (Chapter 3) and analysis methods (Chapter 5). In addition, it is important to determine how much data (sample size) is needed from the target audience. The purpose of this chapter is to provide guidance for selecting and justifying a minimum sample size. Specific items included in this chapter are as follows:

- The difference between a target audience and target population (Section 2.2)
- Factors that can influence the sample size selection process (Section 2.3)
- Common strategies for selecting a sample size (Section 2.4)
- How differences between the target sample size and actual samples collected can influence a study approach (Section 2.5); specifically, instrument selection (Chapter 3) and data analysis methods (Chapter 5)
- Discussion about random sample collection (2.6)

2.2 Target Audience vs Target Population

For most evaluations, it is not feasible to collect data from the entire **target audience** (also known as priority audience)[1]. Instead, a subgroup of the target audience is studied that is ideally representative (has similar characteristics) of the target audience. This subgroup is commonly referred to as the “target population” (Takona, 2002). **Example 1** illustrates the difference between the target population and target audience. For this example, behavior change materials would be distributed to the entire target audience, but evaluation data would be collected only from the target population.

Example 1: Several jurisdictions jointly developed a behavior change campaign that targets all Washington State drivers. As part of their MS4 Permit requirements, the jurisdictions have decided to conduct a study to evaluate the campaign. They have decided the sample size for the study will be drivers

[1] Many social marketing and CBSM professionals have moved away from the term “target audience” because it is perceived negatively by populations who have been “targeted” in an adverse way in the past. Instead, many now use the term "priority audience”. Throughout this manual we continue to use the term target audience because a) it is the term currently used in Permits, b) it is the term that other documents and resources developed for this project use, and c) it is the term readers will encounter in most of the existing printed and online resources provided for choosing, designing, and evaluating behavior change campaigns.

from a large city within the state. In this example, drivers from the city would be the target population, while State of Washington drivers would be the target audience.

Audience segmentation is a common approach in social marketing that is conducted to break down the target audience into smaller subgroups based on common interests or characteristics. The target population may end up being the same subgroup as the segmented audience subgroup; however, there are differences between these subgroups. The target population is the size (number of participants needed) of the sample needed to be representative of the target audience, and the size is identified to determine how much data should be collected, whereas audience segmentation is done for the purpose of developing a more tailored campaign. Common ways target audiences are segmented include age, gender, geography, income, habits, etc. (Wilbur, 2006). **Example 2** provides an example of evaluating a program that used audience segmentation. Section 1.4 contains resources with more information about audience segmentation. If it is not feasible to collect data for the entire segmented audience, follow the procedures described in Section 2.4 to select a sample size, and the subgroup included in the evaluation would be considered the target population.

Example 2: A jurisdiction developed a behavior change campaign that would target restaurants and focuses on proper BMPs for disposing of restaurant fats, oils, and grease (F.O.G.). Since the characteristics of restaurants vary substantially (e.g., food trucks, fine dining, fast food) and may influence how the restaurant disposes of F.O.G., the jurisdiction decided to segment the audience to 20 fast-food restaurants (of the total 100 restaurants within city limits). Materials were provided to only fast-food restaurants and evaluation data was only collected from this subgroup.

2.3 Factors Influencing Sample Size

This section outlines how factors can influence sample size. These factors are important to consider, as they directly relate to different types of analysis methods as well as the accuracy of the study results (Israel, 1992a). For example, certain types of statistical analysis require a minimum number of samples or an equal number of participants within the control group and the target population of an evaluation (McKenzie-Mohr, 2011). Additional discussion regarding how sample size can influence instrument selection and the type of analysis methods selected for a study is found in Chapters 3 and 5, respectively.

2.3.1 Level of Precision

Level of precision (e) defines a range in which the true value of the target population is estimated. The range is typically stated as a percentage, such as e = ±5 percent (Olejnik, 2016; Israel, 1992b; Israel, 1992a). For example, if the results indicate that 75% of the target population adopted a new behavior, and a ±5% level of precision was used to select the sample size, then we can assume that 70–80% of the target population have adopted the behavior.

2.3.2 Confidence Level and Interval

Confidence level helps to quantify whether a result is likely due to chance or a factor of interest. In the context of this document, a factor of interest would be a variable such as the campaign strategy that is or is not influencing behavior change. A confidence interval is selected by the researcher, which is used to determine when results are statistically significant (reference Chapter 5, Hypothesis Testing, for additional information about statistical significance). A typical confidence interval (α) is $\alpha = 0.05$, meaning there is a 95% confidence level that the result is real instead of being due to chance. Conversely, there is a 5% chance of concluding that a relationship exists between a variable studied, even though no relationship exists in the target population. For sample size selection, larger sample sizes are typically associated with higher confidence levels (Olejnik, 2016; Israel, 1992b; Israel, 1992a).

2.3.3 *Degree of Variability*

The degree of variability refers to differences in the characteristics of the target population that may influence the study results. If the target population is more heterogeneous (has different characteristics), a larger sample size is required to achieve a greater level of precision. This heterogeneity can be captured in the "standard deviation" of a variable around its average – larger standard deviations indicate more heterogeneity and larger sample sizes needed (see Section 5.2.4 for more details). The more homogenous (similar characteristics) the target population, the smaller the sample size is needed (Olejnik, 2016; Israel, 1992b; Israel, 1992a). **For example**, suppose the target audience for a behavior change campaign is segmented to only fast-food restaurant managers. Because this business type has more similar characteristics, fast-food managers would be considered a more homogenous population.

2.4 Strategies for Selecting Sample Size

There are many different strategies for selecting a sample size (that is, selecting the size of the target population). Five simple methods are described in this section, and **Table 2-1** provides an overview of these strategies along with the recommended applications for when to use these methods. More details about each method, along with examples for applying the method, are included in the following subsections.

Table 2-1 Overview of Strategies for Selecting Sample Size

Sample Size Selection Method	Method	Recommended Applications
Census for Small Populations	The entire target audience is used as the sample size	Studies with small target populations (less than 200)
Sample Size of Similar Study	Sample size is selected based on sample sizes from similar studies	Any size study
Published Tables	Published tables are used to select sample size	Studies that have identified a target level of precision, confidence level, and variability
Formulas	Equations are used to calculate sample size	Studies that have identified a target level of precision, confidence level, and variability
Website Calculators	An online calculator is used to calculate sample size	Studies that have identified a target confidence level and interval

2.4.1 *Census for Small Populations*

This approach proposes to collect data on the entire target audience and is best suited for evaluations where the target audience is small (less than 200). For this strategy, the target audience would also be the target population. The benefit of this method is that it is simple and provides data on the entire target audience, which would eliminate consideration for many of the factors described in Section 2.3. The disadvantage to this method is that it may be cost prohibitive for large target audiences (Israel, 1992a; Israel, 1992b). In **Example 2**, if data was collected from all 20 of the fast-food restaurants, that would be an example of using the census method to select sample size.

2.4.2 *Sample Size of Similar Study*

This method proposes to use the same sample size used in a similar study. The advantage of this method is that it is easy to determine a sample size. The disadvantage is that if errors were made selecting the sample size in the similar study, there is a risk of repeating the same errors in your own study. With this method, it is important to review the procedures the researchers used to select their sample size before applying the same size on a study, to make sure there was justifiable reasoning for the sample size (Israel, 1992a; Israel, 1992b). **For example**, conduct a review of literature of evaluations that are similar to your own study and compare the size of the target audience to the sample size (target population) the researchers selected. (The waterbehaviorchange.org website described in Section 1.3.2 contains many articles focused on behavior change evaluations that could be reviewed.) Then calculate the average sample size from the articles reviewed to identify a "typical" sample size for your study.

2.4.3 Published Tables

Published tables can be used to select a sample size, such as the one shown in **Table 2-2**. Before determining whether this method is appropriate for your project, it is important to consider the Section 2.3 factors influencing sample size (Israel, 1992a; Israel, 1992b). The advantage to this method is that it is easy to use. The disadvantage is that the factors influencing sample size may not be known by the researcher, such as level of precision and confidence level. If these variables are not known, common values used in similar evaluations could be used. **For example**, the target audience is 1,000 people for a behavior change campaign focused on owners of private stormwater facilities. Assuming a 95% confidence level and ± 5% level of precision (common values), the sample size (target population) is 286. This was determined by finding the target audience size in the first column and the level of precision in the third column (± 5%).

Table 2-2 Sample Size Where Confidence Level is 95%

Size of Target Audience	Target Population Sample Size (n) for Level of Precision (e) of:			
	±3%	±5%	±7%	±10%
100	a	81	67	51
125	a	96	78	56
150	a	110	86	61
175	a	122	94	64
200	a	134	101	67
225	a	144	107	70
250	a	154	112	72
275	a	163	117	74
300	a	172	121	76
325	a	180	125	77
350	a	187	129	78
375	a	194	132	80
400	a	201	135	81
425	a	207	138	82
450	a	212	140	82
500	a	222	145	83
600	a	240	152	86
700	a	255	158	88
800	a	267	163	89

Size of Target Audience	Target Population Sample Size (n) for Level of Precision (e) of:			
	±3%	±5%	±7%	±10%
900	a	277	166	90
1,000	a	286	169	91
2,000	714	333	185	95
3,000	811	353	191	97
4,000	870	364	194	98
5,000	909	370	196	98
6,000	938	375	197	98
7,000	959	378	198	99
8,000	976	381	199	99
9,000	989	383	200	99
10,000	1,000	385	200	99
15,000	1,034	390	201	99
20,000	1,053	392	204	100
25,000	1,064	394	204	100
50,000	1,087	397	204	100
100,000	1,099	398	204	100
>100,000	1,111	400	204	100

Table reproduced from the following citation (Israel, 1992a).

a. The entire population should be sampled.

2.4.4 Formulas

There are many formulas that can be used to calculate sample size. Equation 1 was included in this Manual because it is simple. The sample size (target population) is calculated based on the target audience size, selected level of precision, and an assumed confidence level of 95%. The advantage of this method is that it can be used to calculate the sample size (target population) for different levels of precision or target audiences' sizes. The disadvantage is that the level of precision may not be known. As described in Section 2.4.3, the level of precision may be assumed. **For example**, the target audience is 35,000 and a ±6% level of precision was selected by the researcher. Using **Error! Reference source not found.**, N=35,000 and e=0.06, the target sample size (n) is 276. *Note: Equation 1 was also used to calculate the values in Table 2-2.*

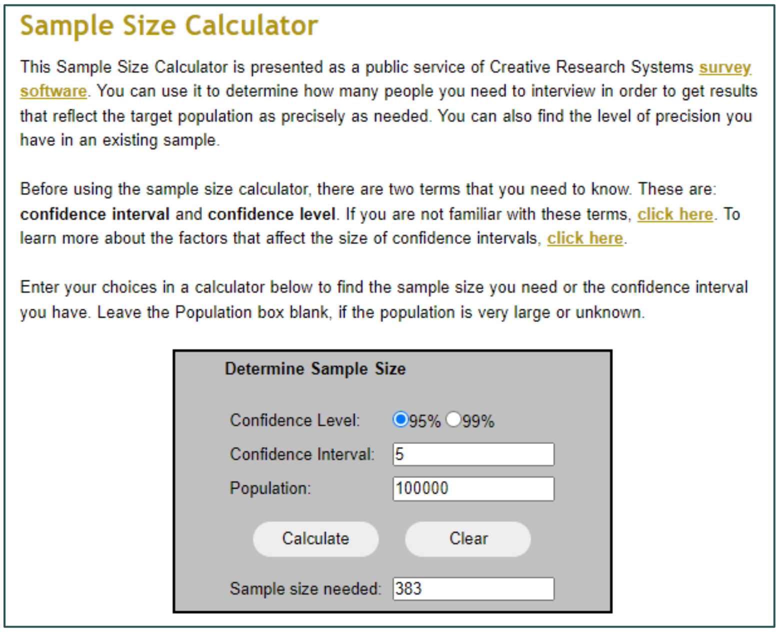
$$n = \frac{N}{1+Ne^2}$$ **Equation 1**

Where:

n	=	Sample size (target population size)
N	=	Target audience size
e	=	Selected level of precision

2.4.5 Website Calculators

There are many websites that have calculators that use formulas to calculate the sample size. A sample size calculator recommended by a TAC member is from the Creative Research Systems website (https://www.surveysystem.com/sscalc.htm). The advantage to this method is that it is easy to use. The disadvantage is that the Section 2.3 factors influencing sample size may not be known by the researcher, such as confidence level and interval. If these variables are not known, common values used in similar evaluations could be used. **For example**, the target audience is 100,000 people and you want to determine the number of people to survey (target population) to collect a representative sample of the target audience. Using the sample size calculator on the Creative Research Systems website and assuming a 95% confidence level and 0.05 confidence interval (common values used on evaluations), the sample size (target population) is 383. **Figure 2-1** provides a screen shot of the calculator and the results. *Note: The reference to the target population in Figure 2-1 is referred to as the target audience in this Manual*.



Sample Size Calculator

This Sample Size Calculator is presented as a public service of Creative Research Systems survey software. You can use it to determine how many people you need to interview in order to get results that reflect the target population as precisely as needed. You can also find the level of precision you have in an existing sample.

Before using the sample size calculator, there are two terms that you need to know. These are: **confidence interval** and **confidence level**. If you are not familiar with these terms, click here. To learn more about the factors that affect the size of confidence intervals, click here.

Enter your choices in a calculator below to find the sample size you need or the confidence interval you have. Leave the Population box blank, if the population is very large or unknown.

Determine Sample Size

Confidence Level: ● 95% ○ 99%

Confidence Interval: 5

Population: 100000

Sample size needed: 383

Figure 2-1 Example of a Website Sample Size Calculator

2.5 Targeted Sample Size vs Actual Sample Size

The sample size selected for a study reflects the number of responses or data collected from the target population. This targeted sample size is typically not the same as the number of surveys mailed or interviews planned. **For example,** Giacalone, et al., implemented a telephone survey to collect information on public perception, knowledge, behaviors, and willingness to get involved in improved stormwater management. Surveys were sent to 1.5 million people located in five different cities and, of those, only 13.4% were willing to complete the survey (Giacalone, Mobley, Sawyer, Witte, & Eidson, 2010). As such, the amount of data planned to be collected (number of surveys mailed, interviews planned, etc.) may need to be increased to compensate for nonresponses; even then, the targeted sample size may not be achieved. Because of this, we recommend starting off an evaluation with the desired target population sample size and then, after the study is complete, comparing the actual sample size to the targeted sample size to estimate the representativeness of the results. **For example**, if your target audience is 100,000 people and a 3% level of precision was selected, based on **Table 2-2**, the target sample size would be 1,099. If after all data has been collected from 100 people, then the final report could describe the results based on a level of precision of ±10% for this sample size. This was determined by locating the target audience size in the first column (100,000) and the target population size in the last column (level of precision ±10%).

2.6 Random Sample Collection

With the exception of collecting data on every member of the target audience (Sec 2.4.1), statisticians advise that where possible you randomly choose which person, business or location to collect data from. By choosing randomly, you maximize the chances that what you learn from the random sample will be representative of the larger group. Without randomization, your evaluation is at risk of "selection bias" and the findings can be misleading.

For example, suppose a pet waste program was evaluated by having a staff member stand near a city-provided station that provided poop bags and a garbage bin. By interviewing only the people who used the station, the researchers inadvertently selected for precisely the people who had successfully made the behavior change. In this case, they would overestimate how successful their campaign had been. By randomly selecting dog walkers along various paths or sidewalks, the researchers would gain a better understanding of the effectiveness of the campaign. Although this example has very obvious selection bias, it can affect evaluations in ways that are often hard to diagnose or anticipate.

Randomization can seem daunting and may not be feasible in some cases. Two simple strategies often suffice. First, if a complete list of the target audience is already available, then one can use a random number generator in Excel (the function RAND) to draw a random sample. **For example**, if the target audience is 100 people and the target population size was determined 51 (Table 2-2), the RAND function could be used to generate a random number for each of the 100, sort the rows by this number, and interview the first 51. Second, where a full list of the target population is unavailable, one can use simple rules like interviewing every fourth dog walker that passes a particular spot, or every seventh house on a street. One can also use an old-fashioned coin toss to determine whether data is collected or not.

3.0 Evaluation Instruments

3.1 Chapter Overview

The purpose of this chapter is to provide an overview of the different types of evaluation instruments (referred to as instruments from this point forward), considerations for selecting and designing instruments, and suggestions for validating instruments. In the context of a behavior change campaign evaluation, an instrument is a measurement device (a survey, interview questions, an observation log, etc.) used to collect data that can be used to assess changes in the target population's understanding and adoption of a targeted behavior. The instruments covered in this chapter include surveys, interviews, focus groups, observations, photos, and drawings. The measurement occurs by comparing data collected using an instrument both before and after implementing a campaign, or by comparing data collected from a control group not exposed to the campaign to data collected from the target population after they were exposed to the campaign.

Instruments fall into two broad categories: researcher-administered and participant-completed. They are distinguished by those the researcher administers versus those that are completed by the participants (target population).

- An example of a researcher-administered instrument would be an observation log completed while observing the target population's behavior.
- An example of a participant-completed instrument is a survey questionnaire that the target population completes following specific instructions (Biddix, 2022).

Instruments may also be classified by the type of data they collect qualitative (i.e., open-ended questions) or quantitative (i.e., multiple-choice surveys). Additional discussion about data types is included in Chapter 4.

3.2 Instrument Types and Selection Considerations

This section provides an overview of the different types of instruments along with considerations for selecting an instrument. Instruments should be selected and developed prior to implementing a campaign to ensure the right data is collected during the campaign evaluation. It is important to note that seldom is only one instrument appropriate for a study. Further, there are typically trade-offs to selecting one instrument over another (Takona, 2002). **For example**, collecting observational data allows the researcher to document participants' actual behavior, whereas a survey is completed by participants who would self-report their behavior.

Research indicates that observational data is typically more accurate than surveys because it documents actual behavior (Kimberlin & Winterstein, 2008; Grove & Fisk, 1992). Because behavior is self-reported by participants in surveys, surveys are subject to social desirability bias where the participant answers the question in a manner, they believe is favorable by others. This can result in over-reporting good behavior or under-reporting undesirable behavior (Grimm, 2010). The trade-off is that, while observational data is typically more accurate than surveys, it is also typically more expensive to collect and analyze compared to survey data, particularly if the survey is administered in an electronic format. That being said, it is not always feasible or appropriate to collect observational data, and every evaluation regardless of the instrument provides value.

An overview of each instrument covered in this chapter is described in **Table 3-1,** followed by a more detailed description of the instrument. Considerations for designing instruments are described in Section 3.3.

Table 3-1 Overview of Instruments and Selection Consideration

Instrument	Description	Considerations for Selection
Surveys	A questionnaire that is typically sent to the target population (participant) who completes the questionnaire following specific instructions. Questions may be closed- or open-ended.	• Any sample size • Low response rate • Less expensive compared to other instruments for large sample sizes
Interviews	An interactive form of data collection that involves an interviewer reading prepared questions to participants and recording their answers. Questions are typically open-ended.	• More suitable for a smaller sample size • Higher response rate • Time-consuming to collect and analyze data
Focus Groups	A small gathering of people who discuss a specific subject under the guidance of a moderator to better understand the target population's perceptions and collect their feedback. Questions are typically open-ended.	• More suitable for a smaller sample size • Inexpensive • Typically requires more than one instrument to conduct a complete evaluation
Observations	Data is collected by the researcher observing and documenting the target population's actual behavior. A predeveloped checklist is typically used to collect data.	• More accurate data because actual behavior is documented • Larger sample sizes can be time-consuming and expensive • May not reveal as much about understanding a targeted behavior
Photos	A camera is used to record any changes in behavior before and after the behavior change campaign takes place.	• Any sample size • Best for documenting inanimate objects such as dumpsters • May increase data management to track where photos were taken and when
Drawings	Drawings developed by the target populations before and after participating in an educational program are used to assess changes in understanding and perceptions.	• Any sample size but typically better suited for smaller sample sizes • Best suited when the target population is school age (K–6) • Time-consuming to analyze data

3.2.1 Surveys

A survey is a process of collecting data that typically involves sending a combination of questions (questionnaire) to the participants who provide responses to the questions. The questions maybe closed- (e.g., multiple-choice, yes/no) or open-ended (Kumar, 2011). The goal of a survey is to learn more about the target population: specifically, about their understanding and adoption of a targeted behavior. Surveys can be administered to any sample size and are typically more cost effective for larger audiences compared to other instruments. Response rates for surveys tend to be lower compared to other instruments. Common methods used to distribute surveys include mail, email, and web links posted on social media or mailed with utility bills.

Because the survey is completed by the participant, the information reported may not be completely accurate due to social desirability bias, discussed above (Grimm, 2010). Careful survey design can minimize or reduce the effects of social desirability bias, as described in Section 3.3.3.

3.2.2 Interviews

An interview is an interactive method of collecting data that typically involves an interviewer reading prepared questions to a participant and then recording the participant's response. The questions are typically open-ended but may also include closed-ended questions. Responses to interview questions typically provide more depth than survey responses because the research can ask more probing questions that provide insights to the participant's responses (Kumar, 2011; Wilbur, 2006). Interviews are more time-consuming to conduct and analyze the data. Consequently, they are better suited for a smaller sample size. In addition, interview response rates tend to be higher than survey response rates (Nehe, 2021). Some evaluations may include both surveys and interviews: the survey is used to learn about the target population and interviews are conducted on a subset of the target population to gain clarification and additional insight on their survey responses. In this case, the interview questionnaire is typically developed based on the survey responses. Interviews are normally conducted face to face, over the phone, or via online video conference.

Similar to surveys, because the interview is completed by the participant, the information reported may not be completely accurate due to social desirability bias (Grimm, 2010). Careful interview and questionnaire design can minimize or reduce the effects of social desirability bias, as described in Section 3.3.3.

3.2.3 Focus groups

A focus group is a small gathering of people in an interactive setting where they discuss a specific subject under the guidance of a moderator or the researcher. The researcher will raise specific questions or issues to stimulate discussion among the focus group participants, and the information collected by the research is used to understand the target population's perceptions and collect their feedback (Kumar, 2011). Typically, data collected from focus groups are used to develop other instruments or aspects of a campaign (Wilbur, 2006). For behavior change campaigns and evaluations, focus groups can be used to develop a better understanding of how the audience perceives a target behavior and provide an opportunity for the audience to discuss in detail their regular behaviors and barriers that prevent them from changing their behavior. Focus groups may also be used to collect feedback on behavior change campaign materials before they are implemented (McKenzie-Mohr, 2011). The sample size for a focus group is usually small, ranging from 6 to 10 people who gather in the same room or an online video conference. Like interviews, focus groups are also subject to social desirability bias. Bias can be reduced by explaining why the participants were chosen to participate in the focus group, what the researcher is wanting to understand about their perceptions or behaviors, and how that information will be used (McKenzie-Mohr, 2011).

Similar to surveys and interviews, because the responses in focus groups are provided by the participant, the information reported may not be completely accurate due to social desirability bias. In a group setting, this can be particularly challenging, especially if one person reports good behavior; the remaining participants may feel uncomfortable providing honest responses about undesirable behavior (Grimm, 2010). Careful design of a focus group outline can reduce social desirability bias, as described in Section 3.3.3.

3.2.4 Observations

Observational data is collected by the researcher observing and documenting the target population's actual behavior. Because actual behavior is being documented (as opposed to self-reporting with surveys), the data is typically more accurate (Kimberlin & Winterstein, 2008; Grove & Fisk, 1992). Observations should be naturalistic and recorded as discretely as possible to be sure authentic behavior is being observed about an individual or a group. Otherwise, if the individual is aware they are being watched, they may act differently than they would if they did not know that what they were doing was being recorded (Kumar, 2011). Prior to collecting observational data, the researcher will develop a checklist that is used to record the anticipated behavior and describe relevant observations. This instrument is well suited for documenting behavior that can be observed in a public place, but it may not reveal as much about the target population's understanding of a targeted behavior. Changes in behavior are measured by comparing the data collected before and after a behavior campaign is implemented. This method works well with any sample size; however, the larger the sample size, the more expensive and time-consuming the data collection and analysis process becomes.

An indirect way to observe and understand changes in the target audience's behavior is to measure outputs. **For example**, if the campaign focused on pet waste disposal and included adding bag dispensers at parks, the number of bags removed from the dispenser could be counted. Because of uncertainty regarding why bags were taken from the dispensers, and not actually observing whether they were used to pick up pet waste, this type of information is typically supplementary to other data that is being collected. An example would be a project with a limited budget for collecting observational data may use results from counting outputs to further support their observational data results. Or these measurements could be collected over a long term to assess whether there is an increase in the demand for bags, which may indicate an increase in understanding and changes in behavior. If this method is used, it will be important to also consider the community growth where the bags are located to determine whether the demand is just increasing with the increase in population size.

3.2.5 Photos

Photographs may also be used to collect data about behavior change. A camera is used to observe and record any changes in behavior before and after the behavior change campaign takes place. The behavior change is measured by comparing the before and after photos side by side. Photographs typically work best with inanimate objects, such as dumpsters, and provide proof of behaviors such as dumpster management. Generally, it is best to ask permission before taking photos of private property. Notifying the property owner that the picture will be taken may introduce issues with the property owners/workers attempting to change their behavior before the researcher arrives to take the photo, so it is recommended that permission be sought at the time of the site visit and right before the photo is taken. Try to take a photo in the exact spot for both the before and after photo so the same views are compared. Typically, a checklist is developed to document and compare the observations.

Example: A campaign focused on dumpster management with a target audience of automobile repair shops within the city. Photos were taken to understand the businesses' dumpster management practices twice at each business during the evaluation: once before the campaign was implemented to collect baseline data, and again after the campaign was implemented to collect follow-up data. Photos were taken of the dumpsters and oil containers, as well as the locations of the nearest storm drains. The site visits took place at unannounced times, photos were taken from the same locations, and then the checklists were completed in the office. **Figure 3-1** provides an example of a before and after photo.

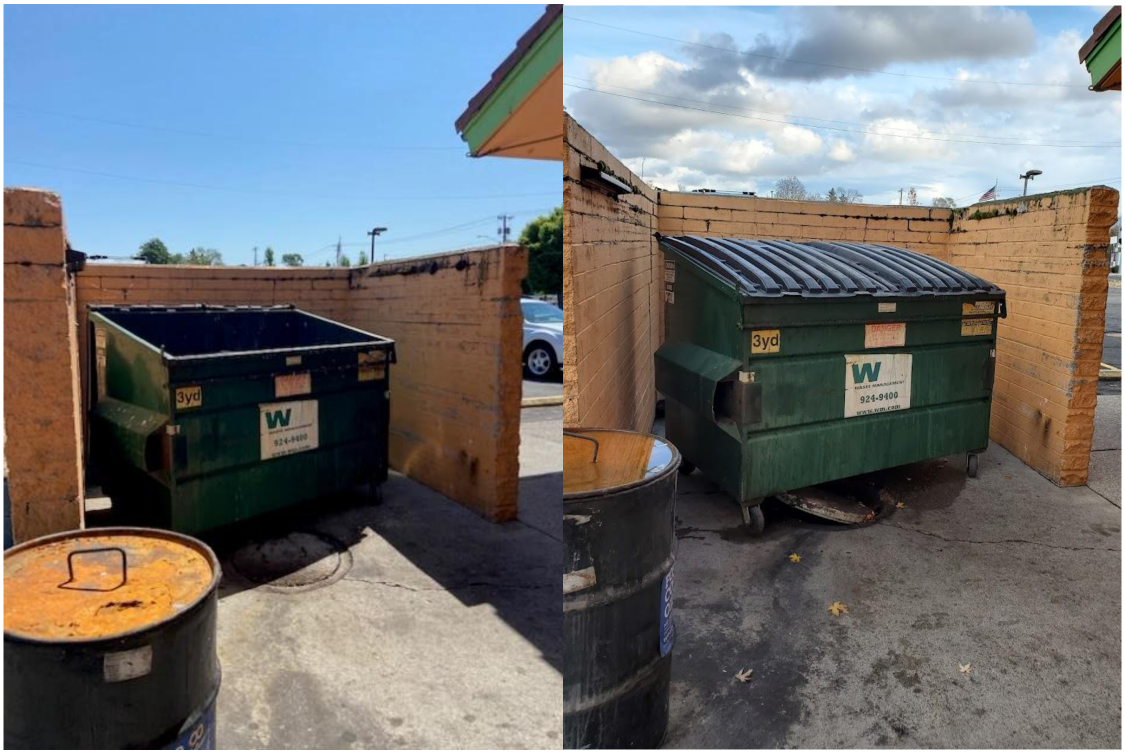


Figure 3-1 Dumpster Photo Before (left) and After (right) E&O Campaign

3.2.6 Drawings

Drawings can be used to evaluate changes in the target population's understanding, perceptions, and adoption of behaviors. This method is best suited for younger school-age children (K-6) and can be time-consuming to prepare for as well as collect and analyze data. Typically, the target population is asked to develop a drawing related to a specific topic (baseline data). Then they participate in an educational program and afterwards they are asked to draw the same thing (follow-up data). The two drawings are then compared to assess changes as a result of participating in the educational program. A checklist of relevant items in the drawings is then developed and used to identify what items are present in the drawings before and after the educational program (Xu, Read, Sim, & McManus, 2009; Miele, 2014). Like items are then grouped and coded into themes (more discussion on coding qualitative data is in Section 4.5), and differences in the two drawings related to each theme are calculated to measure change.

3.3 Considerations for Designing Instruments

This section focuses on things to consider when designing and developing instruments. When selecting an instrument for a particular evaluation, a critical consideration is whether the instrument is reliable and valid.

- A reliable instrument will collect similar data made on the same participants if the study is repeated.
- A valid instrument will measure what it is intended to measure (Takona, 2002).

Reliable and valid instruments are developed by selecting an instrument that is appropriate for a particular study, carefully designing the instrument to answer the study questions, and validating the instrument before it is used. Sections 3.2 and 3.4 provide considerations for selecting and validating instruments.

When designing instruments, it is important to collect data that will help answer the specific questions needed to conduct an evaluation (McKenzie-Mohr, 2011). Instruments should be designed to measure both the understanding and the behaviors of the target population. Instruments can easily become lengthy and complex while under development. Keeping the contents clear and brief is important for answering the question that the evaluation is intended to answer and prevents the instrument from managing data that is not needed. It may be helpful to assess the instruments after they are developed and remove questions that provide data that is interesting to know but will not drive your decision making relative to the evaluation goals.

3.3.1 Survey and Interview Questions

Considerations for developing survey and interview questions:

- When developing questions, only collect the required details needed to measure changes in behavior and understanding.
- Participants should be clear on what the questions are asking. Consider shaping questions in terms that a person with no stormwater background knowledge would understand. For public surveys, using a 5th grade reading level is suggested.
- Each survey should include instructions regarding the purpose for the survey, the jurisdiction(s) involved, and instructions for completing the survey. Instructions should be listed clearly and noticeably just before the first questions is asked.
- Most surveys should be designed to have participants complete the survey in 10 minutes or less. Any more time than this and the participants' attention-span declines and they are less likely to complete the survey.
- Questions may be either closed-ended or open-ended. Some key points and differences for each type of question are as follows:
    - Closed-ended questions offer limited options for responses, such as questions that have multiple-choice or yes/no response options. Closed-ended questions also include response options on the Likert Scale (see Section 4.4 for more details). Closed-ended questions are easier to analyze and can minimize misinterpretation of participants' responses, which can be an issue with open-ended questions.
    - Open-ended questions give participants the opportunity to answer in their own words and are typically designed to elicit more information that can be provided with closed-ended questions. However, it is time-consuming to analyze open-ended responses (see Section 4.5) and it typically takes the participant more time to answer the questions compared to closed-ended questions.
- Designing good survey or interview questions involves selecting the questions needed to meet the evaluation goals and evaluating the questions to make sure they are clear and answer the questions intended. The following four questions can be used to evaluate survey questions (Fowler, Jr., 1984):
    1. Is this a question that can be asked exactly as written?
    2. Is this a question that will mean the same thing to everyone?
    3. Is this a question that people can answer?
    4. Is this a question that people will be willing to answer?

- After the survey or interview questions have been developed, follow the suggestions in Section 3.4 to validate the instrument, which will improve the quality of the instrument and the data collected using the instrument.

3.3.2 Survey Design Resources

A TAC member provided the following information, which includes free resources that may also be helpful for survey design.

- **NOAA Coastal Management – Introduction to Survey Design and Delivery**
  - https://coast.noaa.gov/digitalcoast/training/survey-design.html
  - NOAA also offers training on survey design, and the pdf from its course is available for free download.

- **Survey Monkey™ – 10 Best Practices for Creating Effective Survey**
  - https://www.surveymonkey.com/mp/survey-guidelines/
  - Survey Monkey™ provides a simple starting place for survey design.

- **Harvard Questionnaire Design Tip Sheet**
  - https://psr.iq.harvard.edu/book/questionnaire-design-tip-sheet

- **Survey Fundamentals – A Guide to Designing and Implementing Surveys**
  - https://osteopathic-medicine.uiw.edu/_docs/getting-started-research/survey-fundamentals.pdf

3.3.3 Social Desirability

Social desirability bias is a type of response bias where survey respondents answer questions in a manner they believe will be viewed favorably by others. It can take the form of over-reporting "good behavior" or under-reporting "bad" or undesirable behavior (Grimm, 2010). This section provides suggestions for surveys, interviews, or focus groups that can minimize or reduce the effects of social desirability bias (Ipsos, M. O. R. I., Autumn 2012).

- Carefully consider the following:
  - **How research is introduced.** Avoid priming participants to respond in a more socially acceptable manner, or let them know it is okay to admit undesirable behavior.
  - **How questions are worded can encourage respondents to answer truthfully.** This might include presenting statements other people have made during an interview and then asking the respondent to provide a response that is closest to their own views. This sends a message that there are a range of ”acceptable” responses, rather than one ”right” one. Include statements in the survey or interview instructions indicating that ”there are no right or wrong answers,” to help reduce concerns participants might have about being judged for their responses.
  - **Ask participants what they do (or would do), not just what they think.** Research has indicated there is a disparity between self-reported opinions vs self-reported actions or a willingness to change actions. Asking participants what they do or would do is typically less affected by social desirability bias.
- Use multiple types of instruments and sources to collect and cross-check data to assist with understanding and interpreting the responses. For example, if a survey is released to the public

about car wash wastewater management, consider also collecting observational data, and even trends in washing cars at commercial car washes, and using the combination of results to support the survey findings.
- Ask the same question with the response option in reverse order. For example, if a question asks, "how likely are you to adopt a behavior" and the response options are "extremely likely, likely, neutral, unlikely, extremely unlikely," ask the same question at the end of the survey but put the response options in reverse order: "extremely unlikely, unlikely, neutral, likely extremely likely." If the participant answers, "extremely unlikely" and "extremely likely" to the same question, this may be an indication that their responses are not valid, and consideration should be given to excluding the data from their survey in the final dataset (Hopper, 2013). This same approach can also be done with reverse-wording the question but leaving the response options in the same order for both questions. For more information on this topic, consult the following resource: https://www.formpl.us/blog/how-to-get-the-truth-on-surveys-why-respondents-lie

3.3.4 Target Audience Research

Both social marketing and community-based social marketing (CBSM) recommend conducting target audience research to better understand existing behaviors and barriers that inhibit individuals from engaging in preferred behaviors. This information is also important for developing study instruments that will be used as part of the evaluation. Two ways to collect information about the target audience are noted below. Additional social marketing and CBSM resources are located in Section 1.4.

- Conduct a literature review to identify the target audience and/or determine what is known about the habits and demographics of a target audience relevant to a behavior change campaign. The CBSM website, https://cbsm.com/, is an option for reviewing environmentally-related case studies (Mckenzie-Mohr and Associates, 2005-2022). Web search engines or Google Scholar are also good resources for looking up published reports or more scholarly articles. Many of these options are free and readily available.
- Conduct surveys or focus groups to better understand the target audiences' current behaviors and barriers to behavior change. This information is commonly used to develop behavior change campaigns, and the information collected can also be used to develop survey or interview questions.

3.3.5 Use an Existing Instrument

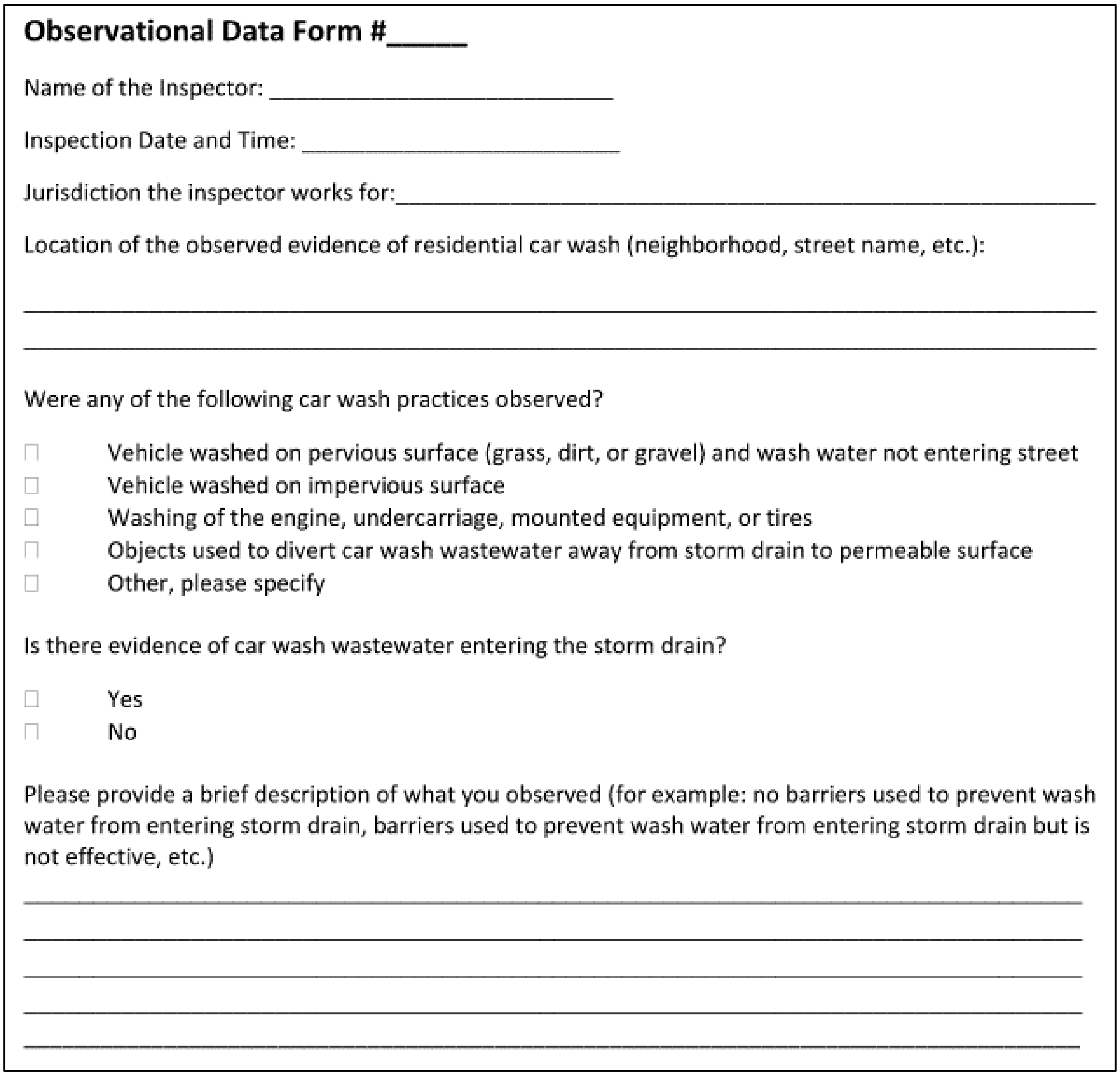
Instruments developed from other studies can be reused, which may eliminate the need to develop a new instrument. However, each study is unique, and an existing instrument will likely need to be adapted to the new study. Reference the waterbehaviorchange.org website for articles that may contain examples of instruments that have already been developed.

3.3.6 Multiple Instruments

It can be beneficial to use more than one instrument for data collection. One reason to do so is that it may be necessary to answer all the evaluation questions. For example, observational data may be used to understand whether changes in behavior occurred, but additional data may need to be collected (e.g., using a survey or interview questions) to determine whether the target audience's understanding of the behavior has changed or why they changed their behavior. When considering using more than one instrument for data collection, assign each instrument a purpose for what data will be collected, and identify how that data will answer the evaluation questions to confirm the additional data is needed. An additional benefit of using more than one instrument for data collection is it can improve the validity of the results, particularly if the results are similar from each instrument. The disadvantage of using more than one instrument is it may cost more time, money, and resources than using only one instrument.

3.3.7 Checklists for Collecting Observational Data

The purpose of the checklist is to reduce the time needed to record the observations and analyze the data. Checklists are typically developed by conducting research (literature reviews, focus groups, etc.) to understand the anticipated behavior and barriers of the target population and then pilot testing (see Section 3.4) to validate the checklist before it is used in a study. In addition, developing a checklist into a standardized form before starting to collect data will minimize errors in the process of collecting, recording, and analyzing errors (Radhakirishna, 2012). **Figure 3-2** provides an example of a checklist.



Observational Data Form #_____

Name of the Inspector: ___________________________

Inspection Date and Time: _________________________

Jurisdiction the inspector works for:____

Location of the observed evidence of residential car wash (neighborhood, street name, etc.):

____

____

Were any of the following car wash practices observed?

□ Vehicle washed on pervious surface (grass, dirt, or gravel) and wash water not entering street
□ Vehicle washed on impervious surface
□ Washing of the engine, undercarriage, mounted equipment, or tires
□ Objects used to divert car wash wastewater away from storm drain to permeable surface
□ Other, please specify

Is there evidence of car wash wastewater entering the storm drain?

□ Yes
□ No

Please provide a brief description of what you observed (for example: no barriers used to prevent wash water from entering storm drain, barriers used to prevent wash water from entering storm drain but is not effective, etc.)

____

____

____

____

____

Figure 3-2 Checklist from a Car Wash Wastewater Management Evaluation

3.4 Validating Instruments

After an instrument has been developed, the next step is to validate the instrument. Validation is a process used to verify that the instrument measures what it was intended to measure and produces stable results (Guba, 1981). Three common methods for validating instruments include:

- Peer Debriefing – Distribute the instrument to a group of your peers and have each of them review/use the instrument. Then have the group meet to debrief on their assessment of the instrument. This will include discussion regarding whether (a) there is more than one way to interpret a question or instructions, (b) the terminology seems clear for a diverse audience, or the terms should be revised because they may not be understood by the general public, or (c) whether the questions or instructions should be revised to improve clarity. The instruments are then revised until the group mutually agrees.

- Field testing instruments before broad implementation – This may include using focus groups or pilot testing the instruments before they are implemented for a study. (Focus groups are described in more detail in Section 3.2.3.) Pilot testing would include implementing the instruments with small subgroups of the target population or separate control groups. Data collected from pilot testing is then used to update the instruments before they are used as part of the evaluation. Data collected from pilot testing would not be included as part of data collected from the actual evaluation.

- Use established instruments from similar studies that have already been validated. Refer to Section 3.3.5 for more information on this topic.

4.0 Data Types

4.1 Chapter Overview

Once data has been collected using the instruments listed in Chapter 3, data will need to be prepared for analysis. Depending on the data collection instrument used, either qualitative or quantitative data will be produced. Qualitative values consist of descriptions, whereas quantitative data can typically be measured or counted and has numerical values. This chapter focuses on data management, providing an overview of qualitative and quantitative data types, and guidance for coding qualitative data as well as converting qualitative data to quantitative data.

4.2 Data Management

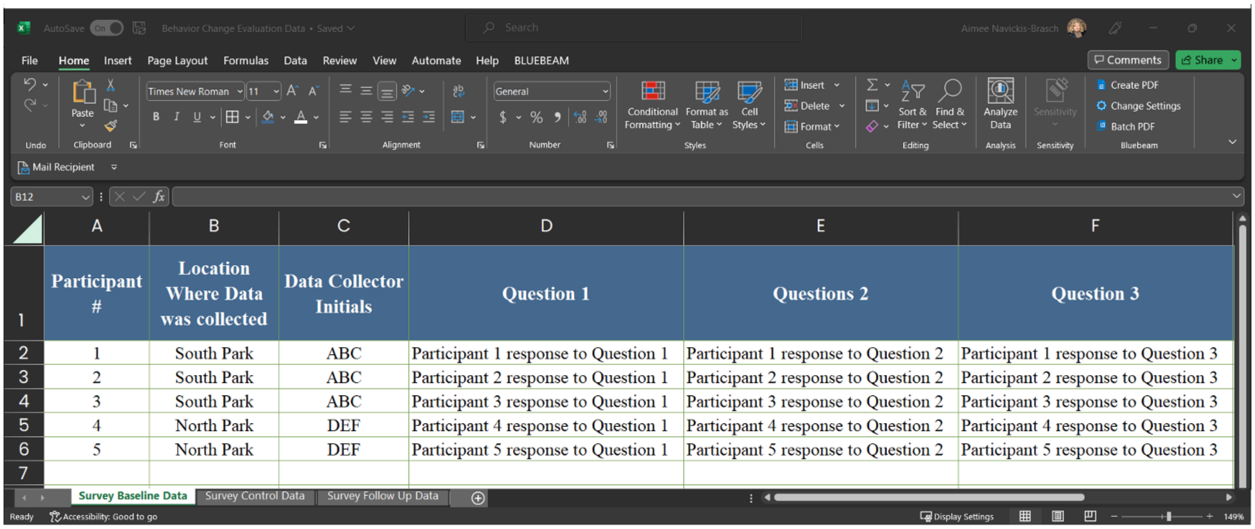
Data management is the organization, storage, and preservation (or archiving) of data collected during the evaluation. It is the everyday management of the data during the data collection and analysis phases of a project and is an important step to reduce the potential for errors (Radhakirishna, 2012). Proper data management also ensures that, should an unanticipated change in key team members take place, the project can be more easily continued by the new team member. It is generally recommended that a plan be developed prior to data collection that outlines how data will be managed.

The remainder of this section focuses on data organization. To accurately measure if there was a change in behavior, data will need to be collected both before (baseline) and after (post or follow-up data) a campaign is implemented. Baseline data provides information about the target population before they are exposed to the campaign, and follow-up data provides information about the target population after they are exposed to a campaign. Baseline and follow-up data are then compared to evaluate changes in the target population's understanding and adoption of a behavior. Both forms of data should be collected in the exact same manner so that, ideally, the only changing variable in the study was that the campaign took place. If it is not possible to collect baseline data, data collected from a control group may be used instead. Control groups are not exposed to the campaign materials and should have characteristics similar to the target population. The same instruments used on the target population should be used to collect data from the control group. In addition, Sugiarto and Cook (2022), recommend collecting a combination of baseline and follow-up data as well as data from a control group because it is considered the gold standard in high-quality evaluations: it eliminates many of the factors which could lead you to mistaken conclusions. However, collecting additional data from control groups uses limited resources and may not be feasible in many cases (Sugiarto & Cook, 2022).

Suggestions for organizing data are as follows and illustrated in **Figure 4-1**:

- Excel© or a similar program is recommended for organizing data.
- Organize data first by the evaluation instrument used to collect the data and then separate data into baseline, follow-up, or controlled responses. Depending on the amount of data collected, it may be easiest to separate each set of data (e.g., baseline or controlled and follow-up) into different worksheets.
- Put each question or item from the evaluation instrument into the column header and arrange responses from each participant into separate rows in the same column. This will allow for an easier comparison of data.

- If data is coded into themes (Section 4.5) or responses are converted to numerical values (Section 4.6), then themes or values can easily be added to the adjacent column in the same row.



	A	B	C	D	E	F
1	Participant #	Location Where Data was collected	Data Collector Initials	Question 1	Questions 2	Question 3
2	1	South Park	ABC	Participant 1 response to Question 1	Participant 1 response to Question 2	Participant 1 response to Question 3
3	2	South Park	ABC	Participant 2 response to Question 1	Participant 2 response to Question 2	Participant 2 response to Question 3
4	3	South Park	ABC	Participant 3 response to Question 1	Participant 3 response to Question 2	Participant 3 response to Question 3
5	4	North Park	DEF	Participant 4 response to Question 1	Participant 4 response to Question 2	Participant 4 response to Question 3
6	5	North Park	DEF	Participant 5 response to Question 1	Participant 5 response to Question 2	Participant 5 response to Question 3
7						

Figure 4-1 Example of Survey Data Organization

4.3 Qualitative

Qualitative data is descriptive data (non-numerical) that can be placed into categories (Creswell, 2013). Qualitative data generally refers to text, such as open-ended responses to survey, interview, or focus group questions, but also includes data collected from observations, photos, and pictures. Quantitative and qualitative data provide different information and are often used together to develop a better understanding of the target population (Austalian Bureau of Statistics, n.d.). Collecting and analyzing qualitative data can provide insights into quantitative results. **For Example**:

- **Quantitative Data:** Observational data collected of people walking their dogs found that 40% do not pick up their dogs' poop.
- **Qualitative Data:** During interviews with dog owners, the top reason why they do not pick up poop is because they forgot to bring a bag with them.

Nominal and *ordinal* data are two forms of qualitative data. Nominal data groups variables into categories that are purely descriptive and do not have any numerical value. Some examples of nominal data include sex, religion, or race. Nominal data may be collected through asking questions such as open-ended questions or answering questions that have a given list of multiple-choice or yes/no response options. Observations recorded from pictures, site visits, or drawings are also considered nominal data. These observations should be recorded through text to express the description of the resulting behavior, whereas ordinal data groups variables into ordered categories, which has a natural order or rank based on some hierarchal scale such as high to low (Kumar, 2011). An example of ordinal data is a survey question that asks how much a person agrees with a statement and the response options include statements such as "Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree."

Qualitative data can be collected using several different evaluation instruments, as described in Chapter 3, including questionnaires, interviews, observations, pictures, or drawings. After collecting qualitative data, that data is commonly coded (Section 4.5) and/or converted to quantitative data (Section 4.6) to make data analysis simpler. Additional discussion about data analysis is included in Chapter 5.

4.4 Quantitative

Quantitative data has a numerical value that expresses a certain quantity, amount, or range. **For example**, if a survey question asks how often a person washes their car each year, the response would be a numerical value and is considered quantitative data. Numerical data can be represented in many ways, including percentages, proportions, or rates of change. Interval and ratio data are forms of quantitative data that represent positions along continuous number lines rather than categories like qualitative data. Quantitative data is also amenable to statistical analysis (Kumar, 2011).

Intervals represent values that have a defined numerical scale where the order of the variables is known as well as the difference between the variables; however, the zero point is arbitrary. Examples of interval data include credit scores and SAT scores. In both cases, it is not possible to get a zero score. Likert Scales are another example of interval data that is often used to give quantitative value to qualitative data. A Likert Scale is similar to ordinal data in which a survey question asks how much a person agrees with a statement and the response options include statements such as “Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree.” The difference is that after data is collected, the response options are converted to a numerical value for data analysis (Kumar, 2011). For example, “Strongly Agree” responses are converted to a 5, “Agree” responses are converted to a 4, and so on. With a Likert Scale, zero has no real meaning.

Ratio data is like interval data in that the numerical distance between points is consistent and can be measured. The difference is that the zero point reflects an absolute zero, unlike interval data in which zero is arbitrary (Kumar, 2011). Examples of ratio data include weight, height, length, area, length of time, or duration. For each of these, zero is possible (e.g., zero duration or weight).

Quantitative data can be used to conduct statistical analyses, including calculating the average response and hypothesis testing, as described in Chapter 5. Analyses such as averages or percent change use quantified values to calculate behavior change and change in understanding. The number of data points, whether that be a quantitative response or the number of times an observation was recorded, can be analyzed as quantitative data. Qualitative data can also be converted to quantitative to simplify data analysis, as described in Section 4.6.

4.5 Coding Qualitative Data

The possibilities of open-ended responses and observations are limitless, with some responses similar but not exactly the same. Similar responses can be coded and grouped together into themes (Gibbs, 2008). Coding is the process of labeling and organizing qualitative data to identify different themes that make it easier to interpret the evaluation findings. The process begins with assigning labels to words or phrases from the target population’s responses. These labels represent the important themes in the response, and labeling data makes it easier to group similar responses (Insights, n.d.). **For example**, a survey question asks the target population why they have not adopted a particular behavior. **Table 4-1** provides an example of responses that were labeled (e.g., the text highlighted in yellow below) and then similar codes were grouped into themes.

Table 4-1 Example Themes Identified from Responses

Example Responses	Theme – Definition
• It is easier to do what I do now • It would take too much time	Convenience – it is more convenient or takes less time to keep their current behavior
• It costs more than what I do now • The recommended products are expensive	Cost – it costs less to keep the same behavior, or the recommended behavior is perceived as expensive
• I was not aware that what I am doing has a negative impact on water quality • There is no stormwater in neighborhood	Unaware – unaware of the impact of their behavior or of stormwater

Things to consider when coding responses:

- A response code may fit into more than one theme, which is acceptable. Record all themes that code responses belong to. All codes will count as one response toward the overall total number of responses for a theme. **For example**, a survey question asks the target population why they have not adopted a particular behavior and they provide the following response: *it is easier to do what I do now, and the recommended products are too expensive*. Based on the themes shown in Table 4-1, there are two themes in this response: convenience (easier to do what I do now) and cost (recommended products are too expensive). Once the data coding is complete, then the number of responses for each theme would be counted and the response in this example would count toward both convenience and cost.
- Clearly define each theme (shown in the second column of Table 4-1) and use the definition consistently in coding all responses. These definitions will also make it easier to code if more than one person is reviewing the data.
- Themes should be the same between baseline and follow-up data review. This will make it easier to compare the differences between the data sets and determine whether there is a measurable difference.

Determining which responses (labels) belong to which theme is based on the interpretation of the person reviewing the data. To confirm the validity of the coding, consider conducting a peer review to backcheck coding methods. This may include a peer who works in the same field and is familiar with the subject of the study. After all the data has been coded by the initial data reviewer, the peer will review a portion of the data to determine whether they agree with the themes identified by the initial data review for different responses. Then the initial data reviewer will meet with their peer to discuss and compare their results until they mutually agree on the interpretation of the coding, which may include changing how some responses were coded or adding additional codes if needed.

4.6 Converting Qualitative Data to Quantitative Data

Qualitative data can be converted to quantitative data to simplify data analysis. **For example**, counting the number of responses to each multiple-choice response option for each question. This information could then be used to calculate the percentage of responses to each multiple-choice option, which makes it easier to compare baseline and follow-up data. Alternatively, multiple-choice or yes/no responses can be converted to a numerical scale similar to using the Likert Scale, as described in Section 4.4. The response options would be assigned using an even incremental range of values. If applicable, the options would be scored considering the relativity to the desired answer, with the desired response receiving the highest value and the most undesirable the lowest value. The scoring could be assigned respectively as 5, 4, 3, 2, and 1. It could also be scored respectively as 2, 1, 0, -1, -2. The zero score works well for neutral responses; then, when the average is calculated, it is easier to compare baseline and follow-up results as more positive or negative responses. The actual values of these numbers do not matter as long as the difference between the scores is consistently equal. Converting responses to a numerical scale is particularly important if hypothesis testing is used to demonstrate whether there is a statistical difference between two data sets. Hypothesis testing is discussed in Section 5.3.2.

Example: A campaign focused on disposal of F.O.G and mop wash water with a target audience of restaurants within the city. Survey responses were collected as baseline and follow-up data, both before and after the implementation of the campaign. Researchers took the survey responses and grouped them together into themes, assigning a numeric qualitative value to each of these themes. Because each question posed different responses, the codes vary for each question. A small sample of the questions and their paired coded responses are shown in **Table 4-2**. After the data has been coded, the results can be analyzed following the methods in Section 5.2. If hypothesis testing is conducted, the numerical scales for each response code would be compared for each question to determine whether there is a statistically significant difference between the baseline and follow-up data. An example of hypothesis testing using this data is included in Section 5.3.2.

Table 4-2 Example Converting Qualitative Data to Quantitative Data

Question	Numerical Scale and Response Codes
What are the impacts of F.O.G./wash water if they reach the storm system?	1 – Does not understand the harm 2 – Knows it is bad but unsure why 3 – Fully understands the impacts of F.O.G. and wash water if they reach the storm system
How are employees educated on F.O.G./wash water disposal?	1 – Trained on proper disposal when hired 2 – Video training once when hired 3 – None/some employees are educated on this topic
Are specific employees trained to inspect and clean the grease traps/interceptors?	1 – Managers 2 – Cooks 3 – No specific employees 4 – Majority of employees 5 – Maintenance or external company

5.0 Data Analysis

5.1 Chapter Overview

After data has been organized and qualitative data has been coded and/or converted to quantitative data (Chapter 4), the next step is to analyze and compare the data. Section 5.2 provides guidance for using basic statistics to calculate and describe the central tendency and variance in data sets. Section 5.3 provides guidance for comparing the data sets to assist with determining whether there is a difference between the baseline and follow-up data. Examples for applying the different methods are also included in this chapter. Section 5.4 provides a list of software options that can be used to perform the data analysis in this section.

5.2 Descriptive Statistics

Descriptive statistics summarize information about a data set that can be broken down into measures of the central tendency or variability. *Central tendency* describes the center position of a data set, and measures of central tendency include the mean, median, and mode. *Variability* describes the spread of a data set, and measures of variability include standard deviation and range. This section provides guidance and examples for calculating central tendency and spread.

5.2.1 Mean

The mean reports the average value of a given data set. This is the most used measurement of central tendency. The mean is calculated by summing all the variables in a data set and dividing by the total number of variables in the data set, as shown in Equation 2. The mean is typically used for normally distributed data, which typically has a low number of outliners. Examples for calculating the mean are included at the end of this subsection. *If the mean is calculated using a program such as Excel™, the formula is =Average(cell1, cell2,…).*

$$\bar{x} = \frac{\sum x_i}{n} \times 100\% \qquad \textbf{Equation 2}$$

Where:

$\bar{x}$ = average or mean

$\sum x$ = sum of the variables in the data set

n = number of variables in the data set

EXAMPLE: A multiple-choice question was used on a survey and 100 people (n=100) responded to the question. There were four response options labeled as A, B, C, and D. The total number of responses to each option was as follows: A (10), B (25), C (15), and D (50). The average percentage of responses to each option can be calculated using Equation 2. An example calculation for option A is as follows. If the analysis is repeated for each response option, the results would be: 10% responded to option A, 25% to option B, 15% to option C, and 50% to option D.

$$\bar{x} = \frac{10}{100} = 0.1 * 100\% = 10\%$$

EXAMPLE: A survey question asked people to indicate how much they agree with a statement, and the response options included "Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree." 100 people responded to the question and the number of responses provided for each response option is shown in **Table 5-1**. The data can be analyzed by determining the average percentage of responses to each response option using the method described in the previous example or by determining the average overall score, as shown in Table 5-1. For the average score, the response options are first converted to a Likert Scale, as described in Section 5.6. Then the number of responses to each option are multiplied by the corresponding Likert Scale value to determine a score for each response, and the total response score is summed. Finally, the average score is determined using Equation 2. The results can then be described as follows: the responses indicate an average score of 3.7, meaning that on average the target population's response is between *agree* and *neutral* to the survey question statement. This type of analysis makes it easier to compare results between baseline and follow-up data.

Table 5-1 Mean Example Data for Likert Scale

Response Option	Number of Responses	Likert Scale	Response Score
Strongly Agree	25	5	125
Agree	40	4	160
Neutral	20	3	60
Disagree	10	2	20
Strongly Disagree	5	1	5
		Total (Σx_i) =	**370**

$$\bar{x} = \frac{370}{100} = 3.70$$

5.2.2 Median

The median is the middle value of an ordered data set and is not affected by outliers. The median would work best for responses to open-ended questions that have been sorted into a range of numerical values (Section 4.6). The median is determined by listing all numbers in ascending order and then locating the middle number. *If the median is calculated using a program such as Excel™, the formula is =median(cell1, cell2,…).*

EXAMPLE: A survey asked how often the target population participated in a specific behavior, such as washing their car each year. Nine people responded to the question and their responses in ascending order were: 0, 1, 2, 3, **4**, 4, 6, 6, 24. The median, or middle number, in this data set is 4. The average value could also be reported for this data set. However, the disadvantage of using the average is, if the data has large outliers (such as 24 times per year), it will strongly influence the average, making the median a better representation of the middle value.

5.2.3 Mode

Mode is the value that appears most frequently in a data set. A data set may have one mode, more than one mode, or no mode at all. Mode is most useful when describing categorical data such as qualitative data that have coded into themes, as described in Section 4.5, and the themes have been a numerical value (Section 4.6). *If the mode is calculated using a program such as Excel™, the formula is =mode(cell1, cell2,…).*

EXAMPLE: **Table 4-1** provided an example for coding responses from open-ended questions. After the data was coded into themes, it was found that the responses included 10 about convenience, 5 about cost, and 2 about being unaware. The mode, or most frequently reported response, would be convenience.

5.2.4 Standard Deviation

The standard deviation is a measure of the average distance of the individual data points from the mean. It is the most used method for describing the variability or spread of a data set. Data sets that have multiple values similar to the mean will have a lower standard deviation, whereas data sets with multiple values that are spread out (i.e., much larger or smaller than the average value) will have a larger standard deviation. Equation 3 is used to calculate standard deviation, and an example calculation is included at the end of this subsection. *If the standard deviation is calculated using a program such as Excel™, the formula is =stdev(cell1, cell2,…).*

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$ **Equation 3**

Where:

s = standard deviation which is also denoted as σ

EXAMPLE: If the data from the median example was used to calculate the average number of times the target population washed their car each year, it would be helpful to also report the standard deviation to indicate how spread out the responses were. **Table 5-2** provides an example for a sample size of nine respondents (n=9). For this example, the average number of times the target population reported washing their cars was 5.56 times per year, with a standard deviation of 7.21 times per year. The results are typically reported as 5.56 ± 7.21, and these results mean that the target population provided a wide range of responses. If the standard deviation were smaller, such as ± 0.21, that would mean the target population provided similar responses.

Table 5-2 Example Standard Deviation Calculation

Reported Frequency	$x - \bar{x}$	$(x - \bar{x})^2$
0	-5.56	30.86
1	-4.56	20.75
2	-3.56	12.64
3	-2.56	6.53
4	-1.56	2.42
4	-1.56	2.42
6	0.44	0.20
6	0.44	0.20
24	18.44	340.20
$\bar{x}$ = 5.56		$\sum(x-\bar{x})^2 = 416.22$

$$s = \sqrt{\frac{416.22}{9-1}} = 7.21$$

5.2.5 Range

Range is the simplest technique for reporting. Range is the differences between the highest and lowest values in a data set. Extreme data points (or outliers) will increase the range. Removing outliers from a data set will decrease the range.

EXAMPLE: Using the data from **Table 5-2**, the range would be the highest reported car washing frequency (24) minus the lowest reported frequency (0), which equals 24. The results are typically reported with the average, such as: the average number times the target population reported washing their cars was 5.56 times per year and the range of responses was 24 times per year.

5.3 Evaluating Results

Once the data from each data set has been analyzed as described in Section 5.2, the results from the baseline and follow-up (or control group) will need to be compared to determine whether there is a change in the understanding and adoption of a targeted behavior. This section describes two methods for comparing results. Section 5.3.1 describes a simplified method and Section 5.3.2 describes methods for conducting hypothesis testing. The advantages and disadvantages of both methods are also described.

5.3.1 Comparing Results

The simplest method to evaluate the results is to compare the baseline results (collected before a campaign was implemented) to the follow-up results (collected after a campaign was implemented) and note the differences (changes) in the results. If it is not possible to collect baseline data, data collected from a control group may be used instead (Section 4.2). This method works best when the same instruments are used to collect baseline and follow-up data and when there is a limited amount of data to work with. The advantage of this method is that it is simple and provides a fast evaluation of the two data sets. The disadvantage of this method is that it does not consider the confidence level of the result regarding whether the result is due to chance or a factor of interest, as described in Section 5.3.2.

EXAMPLE: A campaign focused on F.O.G. and wash water management was evaluated using interview questions. One of the questions the target population was asked was, "*What are the impacts of F.O.G./wash water if they reach the storm system*?" A total of 20 fast-food restaurant managers were interviewed before the campaign was implemented (baseline data) and again after the campaign was implemented (follow-up data). **Table 5-3** provides a summary of the number of responses received to each response code along with the change or difference between the baseline and follow-up data. The results indicate that there are fewer managers that do not understand the harm (four fewer managers) or know it is bad but are unsure why (one fewer managers), and there are more managers that fully understand the impacts of F.O.G. and wash water if they reach the storm system (five more managers). These results indicate that after the campaign there was an increased understanding of the impacts of F.O.G. and wash water.

Table 5-3 Example of Simple Method for Comparing Results from Coded Responses

Numerical Scale and Response Codes	Number of Baseline Responses	Number of Follow-up Responses	Change
1 – Does not understand the harm	8	4	-4
2 – Knows it is bad but unsure why	7	6	-1
3 – Fully understands the impacts of F.O.G. and wash water if they reach the storm system	5	10	+5

EXAMPLE: The same survey was used to collect baseline and follow-up data about a target population's willingness to change their behavior, and the responses were converted to a Likert Scale. The average response to each question is reported in **Table 5-4,** along with the change in the average response from the baseline survey to the follow-up survey. For the Likert Scale, a response of 5 indicates they are very likely to change their behavior, and a response of 1 means they are not likely to change their behavior. Based on the change in the results to all the survey questions, there is an overall increase in the target audience's willingness to change their behavior after being exposed to the campaign.

Table 5-4 Example of Simple Method for Comparing Results for a Likert Scale

Survey Question #	Baseline Results Average	Follow-up Results Average	Change
1	4.0	4.5	+0.5
2	3.5	3.6	+0.1
3	3.75	3.5	-0.25
4	3.25	3.5	+0.25
5	3.0	3.5	+0.5

5.3.2 Hypothesis Testing

The purpose of hypothesis testing is to determine whether there is a statistically significant difference between the two data sets (i.e., baseline data and follow-up data) based on assumptions (see null and alternative hypothesis below). The advantage of using hypothesis testing is that it provides a confidence level regarding whether the result is likely due to chance or the factor of interest. In the context of this document, a factor of interest would be a variable such as the campaign strategy. Hypothesis testing is not required by the MS4 Permits to evaluate differences in data sets; however, it is nearly always a feature of high-quality evaluations. Hypothesis testing works best with large data sets with a minimum of twelve (n=12) samples in each data set needed to conduct testing.

Hypothesis testing starts with defining a null hypothesis and an alternative hypothesis, which are assumptions about the results. Both terms are described below along with relevant hypothesis examples.

- **Null Hypothesis** (H_0) – there is no significant difference between the two data sets. **For example**, a null hypothesis would mean there is no change in understanding or adoption of the target behaviors between the baseline data and the follow-up data.
- **Alternative Hypothesis** (H_A) – there is a significant difference between the two data sets. **For example**, an alternative hypothesis would mean there is a change in understanding or adoption of targeted behaviors.

There are several different methods used to test the hypothesis, and the appropriate method is based on the type of data. Two common methods used in educational research are: (1) Mann-Whitney or Wilcoxon Rank Sum Test, which is for non-normally distributed data and compares the medians between two sets of data, and (2) Paired T-Test, which is for normally distributed data and compares the averages/means and standard deviations of two data sets[2]. Statistical software is commonly used to determine whether data is normally or non-normally distributed and to conduct hypothesis testing. For a Paired T-Test, Excel®™ can be used to conduct the analysis. Reference Section 5.4 for additional software options.

Hypothesis testing is used to confirm or reject the null hypothesis. The confidence interval is selected by the researcher and is used to describe the likelihood that the true value lies within the data set, meaning that results accurately represent the target population's response. A typical confidence interval is $\alpha = 0.05$, meaning there is a 95% confidence level that the result is real instead of being due to chance or error. Conversely, there is a 5% chance of concluding that a relationship exists, even though no relationship exists in the target population (Olejnik, 2016; Israel, 1992b; Israel, 1992a). A researcher may choose to adjust their confidence interval or level to describe the results. For example, results where $\alpha = 0.05$ or less could be considered statistically significant, and α between 0.051 and 0.10 could be considered moderately significant $\alpha > 0.10$ considered insignificant. Most researchers agree that less than a 90% confidence interval is not a robust statistical association.

EXAMPLE: A survey was implemented to determine how likely the target population is to change their behavior. The response options were "Likely, Neutral, Unlikely." The responses were converted to the following Likert Scale: 3 – Likely, 2 – Neutral, and 1 – Unlikely. A summary of the baseline and follow-up responses is shown in **Table 5-5**. The data was then input into a statistical software program called MiniTabs™. A 95% confidence level and a confidence interval of $\alpha = 0.05$ were selected along with the null hypothesis noted above (no difference between data sets) and input into the software. Next, the data was evaluated to determine whether the data is normally distributed using a Normality Test, and the results are shown in **Figure 5-1**. Since the p-value reported from normality testing is greater than our selected $\alpha = 0.05$, and the data points do not follow a straight line, the data is considered non-normally distributed, so the Mann-Whitney or Wilcoxon Rank Sum Test was used to conduct the hypothesis

[2] Paired t-tests depend on the assumption that the data is normally-distributed. In other words, that if one plotted the data, they would resemble a bell curve.

testing. The results of the analysis are shown in **Figure 5-2**. Since the reported p-value is less than 0.05, which indicates that there is a statistically significant difference between the base line data. Based on these results we would reject the null hypothesis and accept the alternative hypothesis (there is a significant difference between the two data sets). Next, the baseline and follow-up data need to be compared to assess whether the differences between the data sets indicate that the target population is more or less likely to change their behavior. This can be done by summing and comparing the response codes. Considering that the sum of values in **Table 5-5** increased from the baseline to follow-up data and that Likert Scale responses with higher scores show an increase in the willingness to change behavior, these results suggest that, as a result of the campaign being implemented, the target population is more likely to change their behavior.

Table 5-5 Summary of Survey Responses

Participant #	Baseline Response	Follow-up Response
1	1	2
2	1	2
3	2	3
4	1	2
5	2	2
6	1	2
7	3	3
8	2	2
9	1	1
10	1	2
11	3	3
12	1	1
13	2	3
14	1	2
15	1	2
16	3	3
17	1	2
18	2	2
Sum =	**29**	**39**
Average =	**1.61**	**2.17**

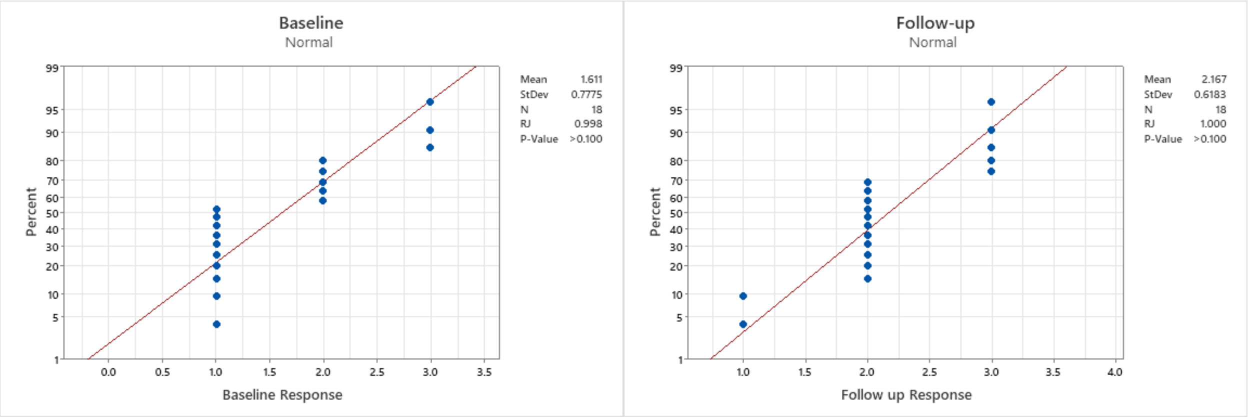
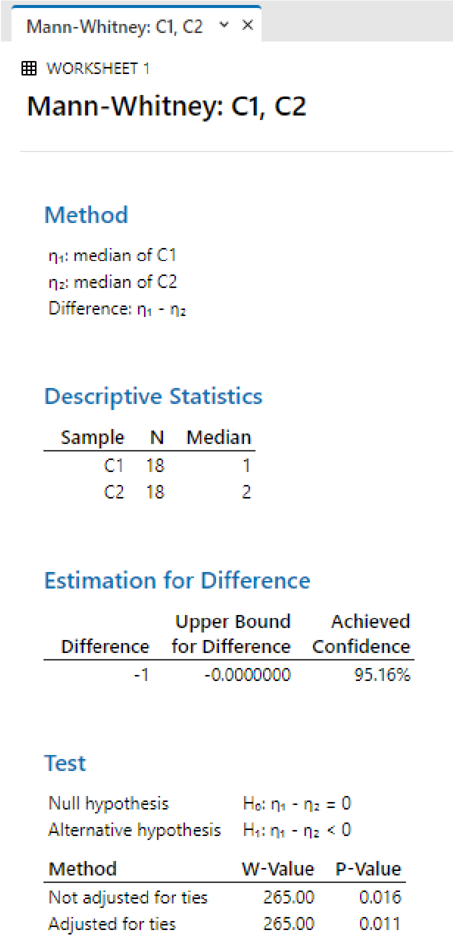


Figure 5-1 Results from Normality Testing



Mann-Whitney: C1, C2

Method

η_1: median of C1
η_2: median of C2
Difference: $\eta_1 - \eta_2$

Descriptive Statistics

Sample	N	Median
C1	18	1
C2	18	2

Estimation for Difference

Difference	Upper Bound for Difference	Achieved Confidence
-1	-0.0000000	95.16%

Test

Null hypothesis H_0: $\eta_1 - \eta_2 = 0$
Alternative hypothesis H_1: $\eta_1 - \eta_2 < 0$

Method	W-Value	P-Value
Not adjusted for ties	265.00	0.016
Adjusted for ties	265.00	0.011

Figure 5-2 Summary of Hypothesis Testing Results

5.4 Software Options

Data analysis is typically done with some type of software. For the descriptive analysis methods, software such as Excel™ is typically sufficient to perform the calculations. For hypothesis testing, more robust software may be needed, and **Table 5-4** provides a summary of options.

Table 5-4 Summary of Statistical Analysis Software Options

Software	Types of data	Additional Description	Cost	Current Version Released	Manufactured/Developed By	Link to Website
Excel			Free version; $139.99 without Microsoft Office 365	Microsoft Excel 2019	Microsoft	
Minitab	t tests; one and two proportions; normality test; chi-square; equivalence tests	Offers government pricing on implementation, training, & maintenance	Starts at ($1,400/user)/yr	20.1.3 (January 2021)	Minitab Inc.	Data Analysis, Statistical & Process Improvement Tools | Minitab
Statistical Package for the Social Sciences (SPSS)	t-tests, ANOVA, z-tests, confidence intervals, proportions, non-parametric tests, etc.		Starts at ($99.00/user)/month	27.0.1.0 (November 2020)	IBM corporation	SPSS Statistics - Overview | IBM
Statistical Analysis System (SAS)		Advertises to benefit a number of industries (public sector being one)	Must contact for pricing	9.4M7 (August 2020)	SAS Institute North Carolina, USA	Data Management Software | SAS
R	ANOVA; t-tests; "linear and generalized linear models, nonlinear regression models, time series analysis, classical parametric and nonparametric tests, clustering and smoothing"	A programming language used for statistical computing and graphics (charts, graphs, etc.); Base for Rstudio software	Free	4.0.4 (February 2021)	Ross Ihaka & Robert Gentleman from R core team	R: The R Project for Statistical Computing (r-project.org)
Rstudio	vectors; lists; matricies; arrays; factors; data frames	Uses the R language to develop statistical programs; Provides further functionality for R	Starts at $995/yr	Rstudio 1.4 (January 2021)	Founded by J.J. Allaire	RStudio | Open source & professional software for data science teams - RStudio
Stata			Starts at ($765/user)/yr	Stata 16.1 (February 2020)	StataCorp	Stata: Software for Statistics and Data Science
Web: **G-Power**	t tests; F tests; x^2 tests; z tests; ANOVA (one-way & multi-way); chi-square tests; some exact tests	Compute data and graphics	Free	3.1.9.7 for Windows (March 2020); 3.1.9.6 for Mac (February 2020)	Heinrich-Heine-Universität Dusseldorf (HHU) – German company	Universität Düsseldorf: gpower (hhu.de)
Web: **Sample Power**	t tests; ANOVA; McNemar's Z test; Cox; test odds	Web-based calculator	Free		SPSS	Power and Sample Size Calculators | HyLown
Web: **StatPages.net**		Statistical search engine	Free			

6.0 References

(n.d.).

Austalian Bureau of Statistics. (n.d.). *Statistical Language - Quantitative and Qualitative Data*. Retrieved from What are quantitative and qualitative data?: https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/Statistical+Language+-+quantitative+and+qualitative+data#:~:text=What%20are%20quantitative%20and%20qualitative,symbol%2C%20or%20a%20number%20code.

Biddix, P. J. (2022, December 1). *Instrument, Validity, Reliability. Uncomplicated Reviews of Educational Research Methods.* Retrieved from Research Rundowns: https://researchrundowns.com/quantitative-methods/instrument-validity-reliability/

Creswell, J. (2013). *Qualitative Inquiry and Research Design: Choosing Amoung Five Approaches.* Thousand Oaks, California: SAGE.

Fowler, Jr., F. J. (1984). *Survey Research Methods.* Beverly Hills, CA: Sage.

Giacalone, K., Mobley, C., Sawyer, C., Witte, J., & Eidson, G. (2010). Survey Says: Implications of a Public Perception Survey on Stormwater Education Programming. *Journal of Contemporary Water Research & Education*, 92-102.

Gibbs, G. R. (2008). Analysing Qualitative Data. *Sage Publications Inc*.

Grimm, P. (2010). Social Desirability Bias. *Wiley International Encyclopedia of Marketing*.

Grove, S. J., & Fisk, R. P. (1992). Observational data collection methods for services marketing: An overview. *Journal of the Academy of Marketing Science*, 217-224.

Guba, E. G. (1981). Criteria for Assessing the Turstworthiness of Naturalistic Inquires. *ECTJ*, 75-91.

Hopper, J. (2013, January 30). *Versta Research*. Retrieved from Tips on “Reverse Wording” Survey Questions: https://verstaresearch.com/blog/tips-on-reverse-wording-survey-questions/

Insights. (n.d.). *Learn to code qualitative data*. Retrieved from Coding Qualitative Data: How to Code Qualitative Research: https://getthematic.com/insights/coding-qualitative-data/

Ipsos, M. O. R. I. (Autumn 2012). Four ways to get the truth out of respondents. . *The social research newsleTTer from ipsos mori scoTland*.

Israel, G. D. (1992a, November). Determining Sample Size. *University of Florida Cooperative Extension Service*.

Israel, G. D. (1992b, October). Sampling the Evidence of Extension Program Impact. *University of Florida Cooperative Extension Service*.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, 2276-2284.

Kumar, R. (2011). *Research Methodology: a step by step guide for beginners.* Thousand Oaks, California: Sage Publications.

Lee, N. R., & Kotler, P. (2011). *Social Marketing: Influencing Behaviors for Good.* SAGE Publications.

Mckenzie-Mohr and Associates. (2005-2022). *Community Based Social Marketing*. Retrieved from Community Based Social Marketing: CBSM.org

McKenzie-Mohr, D. (2011). Fostering Sustainable Behavior. *Fostering Sustainable Behavior-Community Based Social Marketing.*

Miele, E. (2014). Using the draw-a-scientist test for inquiry and evaluation. *Journal of College Science Teaching, 43(4)*, 36-40.

Nehe, J. (2021). Survey Versus Interviews: Comparing Data Collection Tools for Exploratory Research. *The Qualitative Report*, 541-554.

Olejnik, S. F. (2016, October 28). Planning Educational Research: Determining the Necessary Sample Size. *The Journal of Experimental Education*, 40-48.

Radhakirishna, R. (2012). Ensuring Data Quality in Extension Research and Evaluation Studies. *Journal of Extension 50(3), 3*.

Social Marketing Services, Inc. (2008). *Social Marketing Services, Inc.*

Sugiarto, W., & Cook, J. (2022). *A Synthesis and Annotated Bibliography on Stormwater Behavior Change Campaigns* . Pullman, WA: Washington State University Stormwater Center.

Takona, J. (2002). *Educational Research: Principles and Practice.* Lincoln, Nebraska: Writers Club Press.

Wilbur, J. (2006). *Getting Your Feet Wet with Social Marketing: A Social Marketing Guide For Watershed Programs*. Retrieved from Extension University of Wisconin-Madison: https://fyi.extension.wisc.edu/wateroutreach/files/2015/12/GettingYourFeetWet1.pdf

Xu, D., Read, J., Sim, G., & McManus, B. (2009). Experience it, draw it, rate it: capture children's experiences with their drawings. *Proceedings of the 8th international conference on interaction design and children*, 226-270.