Journal of MultiDisciplinary Evaluation
Volume 16, Issue 34, 2020

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Method for Using Rubric Ratings on Fishbone Diagrams to Compare Case Studies

Melinda E. Davey
Jacobs

Jonathan A. Morell
4.669...Evaluation and Planning

Background: In multi-case study program evaluations, the large amount of qualitative data that are generated from interviews can be difficult to utilize. This is particularly so when inference must be made as to why some cases succeed and some fail.

Purpose: This paper shows a method for comparing multiple evaluation sites by using a rubric to define ratings for relevant factors, and an Ishikawa fishbone diagram as a model to show relationships among those factors. We show how this technique identified reasons for differences in outcomes among the sites.

Setting: The evaluation setting was a large-scale safety innovation in the U.S. railroad industry. Four cases were considered–two passenger railroads and two freight railroads.

Intervention: The Confidential Close Call Reporting System (C^3RS) program allowed railroad workers to confidentially submit "close calls" which were reviewed by a team made up of labor, management, and the Federal Railroad Administration (FRA) to determine ways to improve safety.

Research design: Multiple comparative case study, Ishikawa root and contributing cause modeling, evaluative rubric scoring, and data visualization techniques.

Data collection & analysis: Interview data were collected from four pilot railroad sites, each of which participated in a five-year test of C^3RS. Testing periods overlapped, with the entire evaluation lasting about 12 years.

Findings: The method of using Ishikawa fishbone diagrams with ratings from an evaluative rubric was an effective method to summarize, analyze, and present large quantities of qualitative data. The approach succeeded in explaining degrees of success and failure across the sites. The sponsor and industry stakeholders were able to understand the analysis and the findings, and to develop deep insight into how to promote successful implementation.

Keywords: *Multiple comparative case studies; qualitative methodology; qualitative coding; data visualization; fishbone diagrams; Ishikawa diagrams; evaluative rubrics; close calls; near miss; data visualization.*

Introduction

This paper is about how we compared success and failure cases during the evaluation of a pilot safety program called the Confidential Close Call Reporting System (C^{3}RS) which was implemented at four railroad pilot sites. All stakeholders felt a strong need to understand reasons why some pilot sites succeeded more than others, and in particular, what drove failure. Here we report on the methods related to this topic. Our methods combined Ishakawa root cause modeling from Industrial Engineering, multiple comparative case study methodology, data visualization techniques, and evaluative rubric scoring.

This was an evaluation of a high-profile innovation, and hence, was of considerable interest to a variety of stakeholders – railroad management, railroad labor, the Federal Railroad Administration (FRA) sponsoring the program, and the academic accident research community. As a condition of participation in the pilot, each site agreed to provide extensive, and multiple types, of quantitative and qualitative data, including surveys, interviews, corporate archival safety data, and summary C^{3}RS program data to the evaluation. Over the course of the evaluation, customized reports of findings were provided at multiple times to each of these stakeholders. Core members of the evaluation team were the authors of this paper and staff from the Volpe National Transportation Systems Center.

Evaluation Background

For the purposes of this paper, a brief summary of the C^{3}RS evaluation appears below. An extensive description of the whole evaluation's methods and results can be found our final report (Ranney et al, 2019). In the C^{3}RS demonstration pilot program, railroad employees could submit confidential reports about "close calls" to a Third-Party government agency. The Third-Party agency then removed identifying information and provided the reports to Peer Review Teams (PRT) at the railroad carriers made up of labor, management, and FRA representatives. The PRTs analyzed the reports and provided recommendations for corrective actions to their carriers, who reviewed and selected corrective actions to implement with the help of labor.

The FRA also decided to sponsor a summative and formative evaluation to determine: (1) What conditions are necessary to implement C^{3}RS? (2) What is the impact of C^{3}RS on safety and safety culture? (3) What factors help to sustain C^{3}RS long-term? The evaluation comprised four demonstration pilot sites at four different railroads, two freight and two passenger. Each site's demonstration lasted five years with start and stop times overlapping, but not coordinated, with about twelve consecutive years of data collection in total. Three rounds of data analysis and reporting took place at each site in baseline, midterm, and final phases. Our mixed methods design included qualitative and quantitative data, e.g. interviews, surveys, corporate safety metrics, and summary data from the safety program. Thus, the research design had three dimensions: multiple data sources, time, and multiple case studies as shown in Figure 1. During each site's demonstration, we provided periodic formative analyses and presentations to stakeholders. We also completed a summative evaluation at the end of each site's demonstration period. After the four sites completed their demonstration period, the final step was to provide cross-site findings to the railroad industry and stakeholders concerning implementation, impact, and sustainability.

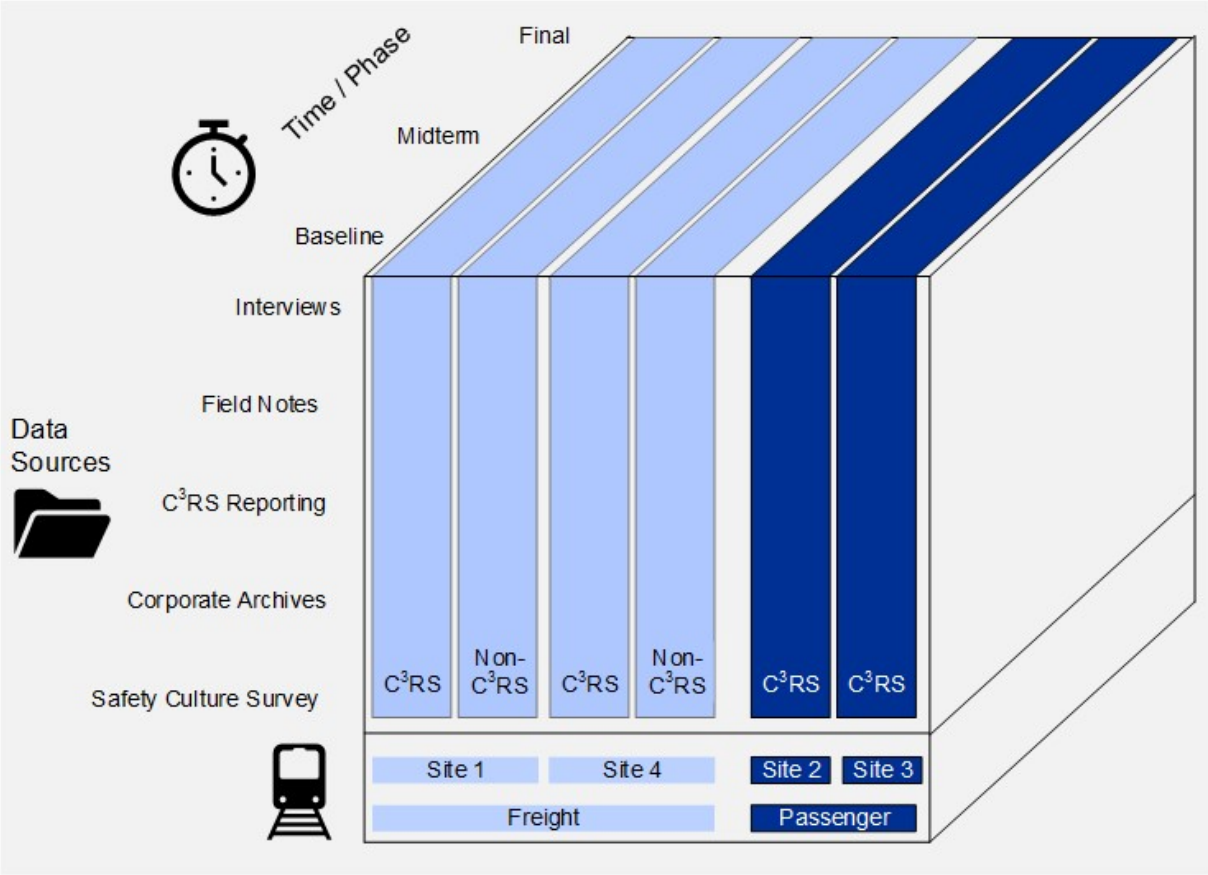

Figure 1. *Research design.*

Analytic Techniques

Case Focus

The unit of analysis for the C³RS evaluation was a site within a railroad, and we had four of them. Analysis of these cases drew on the approaches suggested by Yin's (2013) work on multiple comparative case studies and Brinkerhoff's (2005) Success Case Method (SCM). Yin articulates the logic of cross-case comparison, and provides an approach to identifying cases, methodology for analyzing each case, and procedures for making cross-site comparisons. In the Success Case Method, evaluators locate a success case, and then document the nature of the success. Brinkerhoff recognized that determining reasons for lack of success and comparing them with success cases could also be useful. SCM includes both storytelling and the identification of factors that enhance or impede a program (Medina et al., 2015).

Ishikawa Root Cause Modeling

Ishikawa diagrams, also known as fishbones, were originally used to in quality control and Industrial Engineering (Ishikawa, 1982). Since their inception, they have proved useful in many situations where the relationship between cause and effect needs to be known. In these models, an "effect" represents a desirable condition that is being sought. It is placed at the extreme right of the diagram. "Main factors" that contribute to that effect are placed before it, thus resembling "bones" on the fish. Then "detailed factors" for each main factor form the "sub-bones" (see Figure 2). As needed, the "sub-bones" can be further broken down into successively finer detail. This

process of driving toward more detail can be carried out until further detail does not contribute to understanding causal reasons for the effect. As with all model building, no matter how extensive the data input, judgement is needed, so discussions about the model can lead to productive conversation about assumptions and data interpretation. Also, as in all models, the approach has limitations. It does not provide insight as to the relative importance of inputs at the same factor level, and it does not account for interactions among elements. Still, the Ishikawa approach has proved useful in many settings over the decades since it was introduced.

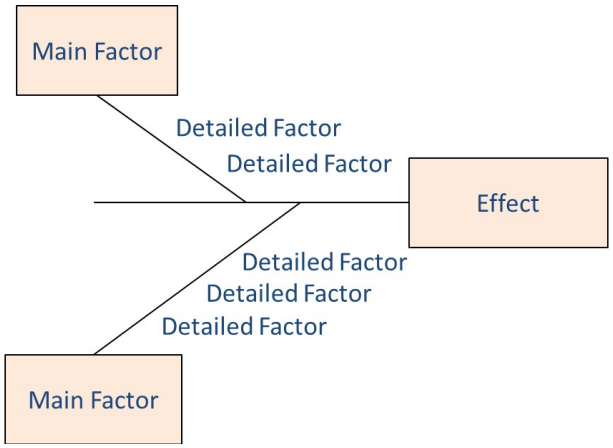

Figure 2. *Structure of fishbone diagram.*

Visual Display

Because the models would play a critical role in helping all the stakeholders understand success and failure, it was important to make sure that their graphical depiction conveyed as much information as possible. The need was to jointly maximize density of information and visual clarity. Without the former, relationships would not be revealed. Without the latter, relationships might be identified, but they would also be invisible. To obtain the best possible joint maximization of these design criteria, decisions about color, shape, and element grouping were paramount (Evergreen, 2014; Malamed, 2011; Jones et al., 2019).

Thinking about Failure

The original reason for our cause and effect modeling was to understand what happened at the one site where our data showed that C^{3}RS had failed. The stakeholders, including the program participants at the railroads and the sponsors at the FRA, wanted to know was what went wrong. This site's failure was particularly striking because the other sites were achieving success. However, it was obvious that to understand failure, it was also necessary to understand success and compare across the sites. Using traditional qualitative analysis techniques, we had previously performed content analysis on the interview data (Patton, 1987, 2015). We coded each interview for both themes from the initial evaluation logic model and themes that emerged during the content analysis activities. Then we summarized the data from each theme. But beyond a summary of interview themes, we needed to understand how all the individual challenges had contributed to the

overall failure. In Creswell's mixed method integration of "explaining", qualitative data can be used to "explain" the poor quantitative results in more depth (Creswell, 2014).

As a result of all our data collection and analysis, we had a strong sense of how and why C³RS's implementation at this site contributed to the failure. It remained to formalize our understanding and to ground it firmly in the data. To develop this understanding, we began by sketching an informal model explaining how negative aspects of the site's implementation contributed to other negative aspects and eventually the overall failure and decision to not continue the program after their demonstration period. Next, we began to think about how the other sites had generally performed differently in those areas and drew a fishbone diagram to explain success more generally (see Figure 3).

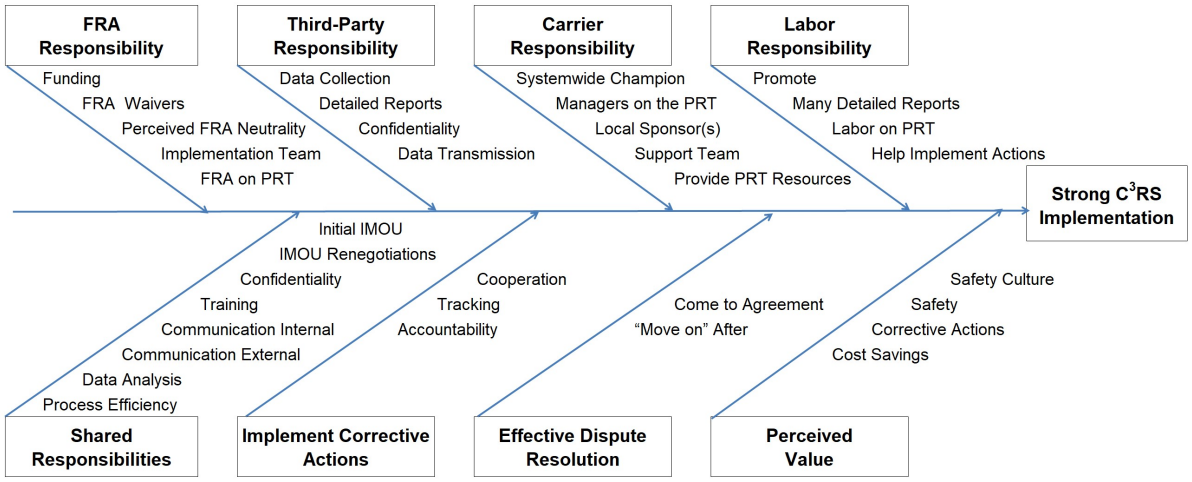

Figure 3. *Example fishbone diagram for implementation success.*

At the extreme right is the "effect", in this case successful implementation. At the top are the unique contributions of each stakeholder group. For instance, only the Third-Party can collect C³RS reports. This placement made it easier to understand the individual contribution of each stakeholder to implementation success. On the bottom at the extreme left are responsibilities that are shared by multiple stakeholders, for instance, communication between the stakeholders. The other elements at the bottom are critical C³RS activities (implementing corrective action and dispute resolution), and the perceived value of C³RS. (Detailed descriptions of each factor and the model are available in our final report [Ranney et al., 2019]).

Rating Each Site

Once the implementation fishbone was created, the next step was to rate how well each site had performed on each detailed factor. This required one hundred forty separate ratings. (35 elements per model x four railroads.) This work was carried out by a team of three people who had been with the evaluation since its early days. The process proceeded along the lines of "evaluative rubrics", as proposed by (Davidson, 2005). Davidson defines an evaluative rubric as a tool that describes different levels of performance and an evaluative description of what performance "looks like" at each level.

Our first step was to decide on operational definitions of the ideal performance for each implementation factor (i.e., for each bone on the fish). The next step was to decide on a

rating scale. Considerable thought went into deciding on a four-point scale. We made this decision because the team members agreed that it was the best possible compromise between the need for rating spread, and the precision at which we could rate factors. The rubric is presented in Table 1.

Table 1
Example Rubric for Rating Implementation Factors

Qualitative Rating	Quantitative Rating	Definition
Very Good	4	Execution of implementation factor is clearly exemplary, could not have been substantially better. Any gaps or weaknesses are not significant with respect to C^3RS operations and are managed effectively.
Good	3	Execution of implementation factor is functional and adequate. A few weaknesses may exist, but none are considered overly problematic.
Fair	2	Execution of implementation factor is inconsistent and/or has multiple weaknesses. It does not adequately support C^3RS operations but did not pose a major threat on its own.
Poor	1	Execution of implementation factor has numerous weaknesses, some of which pose a serious threat to C^3RS operations.

Once the rubric was in place, we assigned a rating to each detailed implementation factor in the fishbone model. The ratings were assigned in the "absolute" sense by comparing the site's implementation to the "ideal" operational definition, so we could identify factors that were challenging across all sites. A strictly regimented four-step process was applied to assigning the ratings:

1. Create a table with one row for each detailed factor, and a column for each evaluation case. (We had four columns, one for each site) (Table 2).
2. Looking at one detailed factor at a time, pick a rating for each site. Write a short explanation for each rating, referring to specific details from the data.
3. Review past presentations and data summaries to see if any important implementation details are missing. Add details to rating explanations as needed.
4. Review all ratings with additional analysts in a live discussion. When disagreements occur, return to the original coded interview data for more information. Repeat reviews as needed.

Our summative data analysis had revealed that three of the four sites achieved impact, but only two sites achieved long term sustainability after the demonstration pilot period ended. Because both impact and sustainability were important outcomes, we adapted the classic Ishikawa form by placing two sequential outcomes at the right-hand side of the model. In order to use differences among the models to reveal differences in outcomes across sites, we created standardized rating definitions for impact and sustainability, with possible values of "poor" or "good". In later analysis, we collected and analyzed data on reasons for lack of sustainability, which went beyond the causes shown in the Implementation Success Fishbone model, which is discussed in our final report (Ranney et al., 2019).

Table 2
Example Template for Implementation Factor Ratings

Implementation Factor	Site 1	Site 2	Site 3	Site 4
Factor Name: Operational definition of factor	Rating: ____ Explanation for rating using information from interviews.	Rating: ____ Explanation for rating using information from interviews.	Rating: ____ Explanation for rating using information from interviews.	Rating: ____ Explanation for rating using information from interviews.

Comparing Sites to Determine Final Findings

Once the ratings were complete, we created two complementary data visualizations to look for implementation findings related to the known impact and sustainability ratings. One version was organized by site/case (see Figure 4) and the second by implementation factor (see Figure 5). We cast the data in these two forms because each visualization provided different insight as to causes of implementation success.

In the data visualization organized by site/case, we created four total fishbone diagrams with color coding (or black and white symbols in other versions) to indicate the ratings (see Figure 4). In this figure, blues indicate "good" and oranges are "poor or fair". This allowed us to see which sites were more successful and show that implementation did appear to have an impact on success. In the figure, you can see that Site 4 has about twice as many areas of "poor or fair" implementation factors than the other three sites and also Site 4 had poor impact.

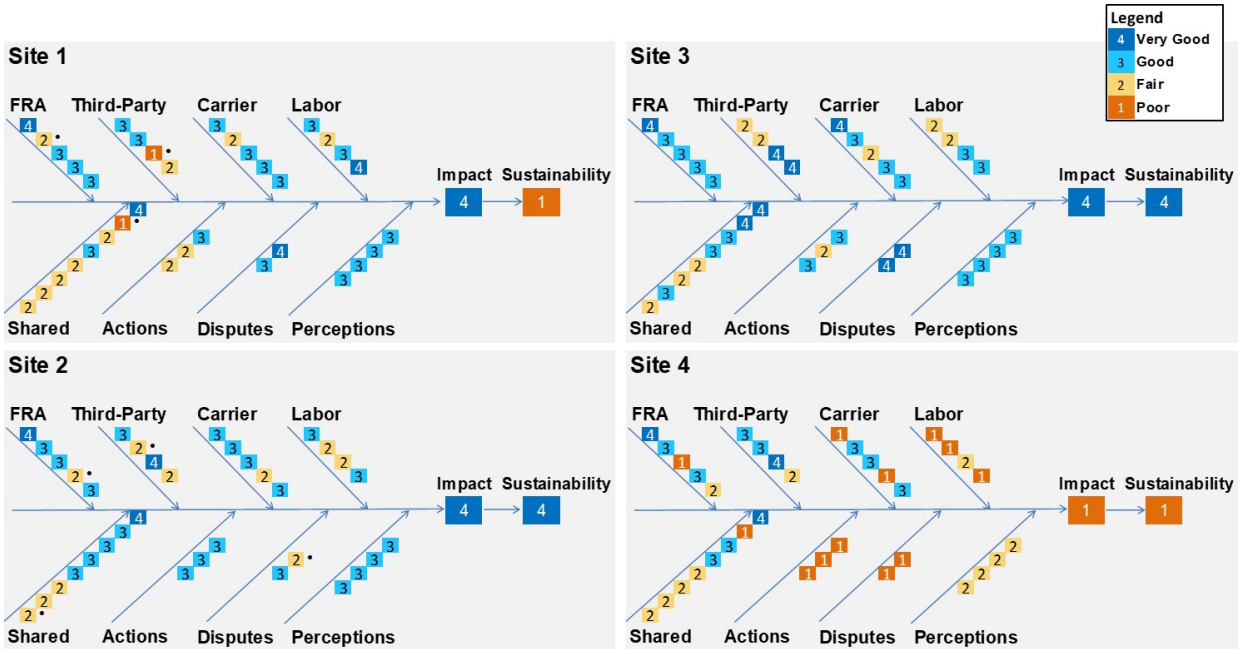

Figure 4. *Example fishbone diagrams with ratings organized by site.*

Figure 4 was designed to show differences across the railroads. Figure 5 was organized to focus attention on implementation factors. This organization showed which implementation factors were most difficult across sites, which factors seemed to be needed for success, and different methods that sites used to find success in the program. For example, under "Shared Responsibilities", you can see that "Communication" and "Process Efficiency" were challenging across the sites. Under "Labor Responsibilities", "Helping Implement Actions" was done well at all the three sites with positive impact but was performed poorly at the failed site.

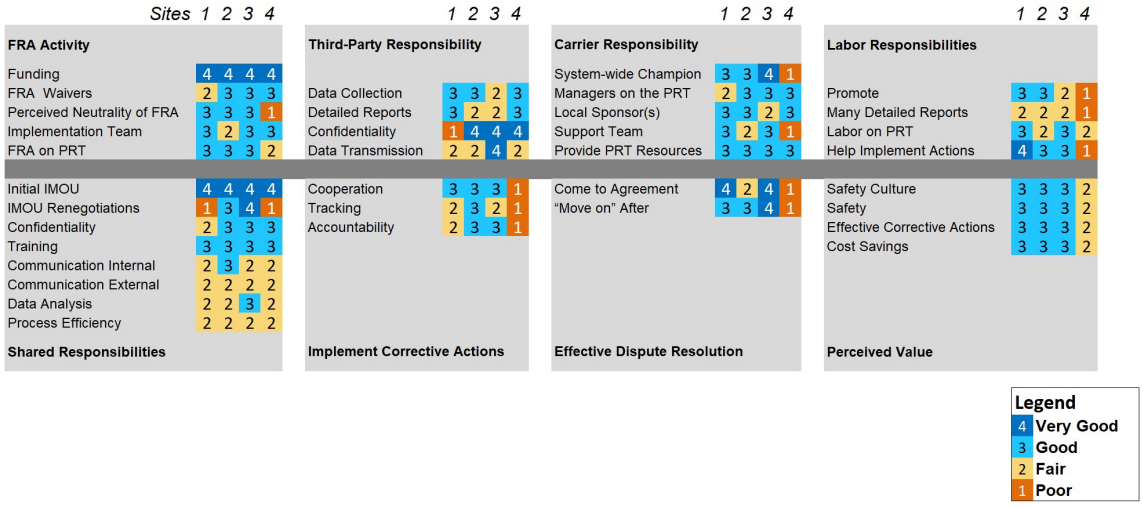

Figure 5. *Example rated fishbone diagrams organized by factor.*

Under "Carrier Responsibility", the distribution of scores was not the same across sites. The three successful sites had different ways they supported the program, some with local managers, some with a system-wide champion, and others with a local champion. This indicated that while some level of management support is needed for implementation success, it is possible to achieve it through different methods.

Sharing Findings with Stakeholders

The process described in the Comparing Sites to Determine Final Findings section above took place toward the end of the evaluation. Prior to that point, each stakeholder group was briefed many times on our emerging findings. Some of these briefings were to mixed groups of stakeholders, while some were customized for each group and for each individual railroad. Thus, by the time the information in this article came to light, all the stakeholders knew us personally, had heard our presentations many times, and had a pretty good idea of what we would say about each site and about C^{3}RS as a whole. We believe that these personal relationships, and familiarity with our findings, played an important role in acceptance of our end-of-project findings, not all of which were appealing or complementary.

Evaluation results were first presented to each site individually. Then we showed the set of all four diagrams to multiple groups of internal and external industry stakeholders. Finally, we wrote a technical report that contained all of the operational definitions of detailed factors, the scoring rubric, the explanation for each rating, figures, and a discussion of the findings (Ranney et al, 2019). These efforts provided us with consensual validation of our findings. The general sense from the stakeholders was that our assessment of their site's implementation was correct. These were polite but not shy

audiences, and they all agreed with our findings. That agreement was particularly satisfying because none of the presentations avoided reporting findings about what did not go well.

We are convinced that one of the reasons for our success was the effort we put into the quality of the diagram graphics. The reactions of our audiences indicated that we did indeed achieve a good combination of information density and visual clarity.

Conclusion and Future Work

We found that applying an evaluative rubric to score the elements in fishbone diagrams, combined with careful attention to color and layout, was an effective method to summarize, analyze, and present large quantities of qualitative data across multiple case studies. This methodology allowed us to satisfy the information needs of multiple stakeholders with respect to a close call reporting system that represented a departure from common practice in the railroad industry.

Another extension to this method could be to apply the rubric ratings to different styles of models, like the original logic model. In addition, determining ways to visually display the ratings for much larger numbers of case studies, bigger than our set of four, could also be explored.

Acknowledgements

This work was performed by the authors in collaboration with the Volpe National Transportation Systems Center. The project was sponsored by the Federal Railroad Administration.

References

Brinkerhoff, R. O. (2005). The success case method: A strategic evaluation approach to increasing the value and effect of training. *Advances in Developing Human Resources*, *7*(1), 86–101.

Creswell, J. W. (2015). *A concise introduction to mixed methods research*. Thousand Oaks, CA: Sage.

Davidson, J. E. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Thousand Oaks, CA: Sage.

Evergreen, S. D. H. (2014). *Presenting data effectively: Communicating your findings for maximum impact*. Thousand Oaks, CA: Sage.

Ishikawa, K. (1982). *Guide to quality control*. Asian Productivity Organization.

Jones, N. D., Azzam, T., Linnel Wanzer, D., Skousen, D., Knight, C., & Sabarre, N. (2019) Enhancing the effectiveness of logic models. *American Journal of Evaluation*, online first.

Malamed, C. (2011). *Visual language for designers: Principles for creating graphics that people understand*. Rockport Publishers.

Medina, L. Acosta-Perex, E., Velez, C., Martinex, G., Rivera, M., Sardinas, L., Pattatucci, A. (2015) Training and capacity building evaluation: Maximizing resources and results with the Success Case Method. *Evaluation and Program Planning*, *52*, 126-132.

Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (1987). *How to use qualitative methods in evaluation*. Thousand Oaks, CA: Sage.

Ranney, J. M., Davey, M., Morell, J., Zuschlag, M., & Kidda, S. (2019) *Confidential Close Call Reporting System (C^3RS) lessons learned evaluation—Final report*. Federal Railroad Administration. DOT/FRA/ORD-19/01.

Yin, R. K. (2013). *Case study research: Design and methods* (Applied Social Research Methods). (5th ed.). Thousand Oaks, CA: Sage.