Practitioner’s Guide to the Analysis of In-Service Performance Evaluation Data

Christine E. Carrigan
Phone: 207 513 6057
e-mail: christine@roadsafellc.com

Malcolm H. Ray
Phone: 207 514 5474
e-mail: mac@roadsafellc.com

RoadSafe, LLC
Box 312
12 Main Street
Canton, Maine 04221

Submitted
July 25, 2015

Resubmitted
November 4, 2015

Word count
Text = 5,606
Figures & Tables: 4 @ 250 words each = 1,000
Total number of words= 6,606

Paper prepared for consideration for presentation and publication at the 95th Annual Meeting of the Transportation Research Board, January 2016
Practitioner’s Guide to the Analysis of In-Service Performance Evaluation Data
Christine E. Carrigan and Malcolm H. Ray

ABSTRACT

In-service performance evaluations (ISPEs) have been acknowledged as a fundamental component of the roadside hardware design evaluation process and have been recommended for the last 30 years in all editions of crash test and evaluation guidelines. This paper consolidates the accepted analytical methods used for analyzing and establishing the confidence interval as well as interpreting the results obtained in an ISPE of roadside hardware. This paper demonstrates there is no uncertainty about how to conduct ISPEs, but a lack of institutionalization for performing them and a need to change the mindset of the greater roadside safety community. ISPEs are needed not only to better understand the performance of roadside hardware, but also to improve the crash testing specifications which are used in the design and acceptance of new hardware for use on the nation’s highways as well as to improve design guidelines used for warranting and placing roadside hardware in the field.
INTRODUCTION

An in-service performance evaluation (ISPE) of roadside hardware assesses hardware performance under actual field conditions, thereby making the leap from a handful of specific impact speeds, angles, and vehicle types assessed during a crash test to the wide array of actual impact locations and traffic conditions observed in the field. The performance of roadside hardware in the field is more important than its performance in a few controlled crash tests performed under ideal circumstances using carefully constructed test installations because it is the effectiveness of roadside hardware under real-world conditions that matters most to the traveling public.

Crash test and evaluation criteria have been updated regularly over the last 30 years. One recurring theme in each re-writing is the recommendation to conduct ISPEs. Michie et al. recommended ISPEs in the crash testing procedures documented under the National Cooperative Highway Research Project (NCHRP) Report 230, published in 1981. The importance and need for ISPEs has been reiterated by Ross et al. in NCHRP Report 350 as well as by AASHTO in the latest crash testing procedures, the Manual for Assessing Safety Hardware (MASH). (1; 2; 3)

MASH likens crash tests to experiments when encouraging ISPEs. For example, the first paragraph of MASH Chapter 7 reads: “…in-service performance evaluation (ISPE) is a very important step in the assessment of the impact performance of a new or extensively modified safety feature. …The in-service performance evaluation remains an important follow-up to the crash test experiments described in previous chapters. Testing and analysis only partially assess the efficacy of a feature and a more thorough and in-depth knowledge of the feature is important to its proper implementation.”

The Federal Highway Administration (FHWA) says “ideally, all highway agencies should know precisely what has been incorporated into its roadway/roadside infrastructure and be able to monitor the performance of individual components of its highway system. Asset management has become a primary means of accomplishing this goal in many states. However, there remains one area where in-service evaluation or performance monitoring seems to be minimal at best, and that is the area of roadside safety features.” (4)

In-service performance evaluation procedures were published over a decade ago in NCHRP Report 490, In-Service Performance of Traffic Barriers which presented a detailed procedure for conducting several types of ISPEs of roadside hardware. (5) According to Report 490, an ISPE includes three major tasks (i.e., planning/preparation, data collection and analysis). Some of the key aspects of a Report 490 style ISPE are:

- Choosing the type of ISPE study (i.e., retrospective or prospective),
- Defining the data collection area, time period and sampling plan,
- Defining data filtering procedures to identify the cases of interest, and
- Analyzing the data and establishing the confidence interval for the results.

This paper consolidates the accepted analytical methods used for analyzing collected data and establishing the confidence interval as well as interpreting the results obtained. This paper demonstrates there is not a question of how to conduct ISPEs, but a lack of institutionalization for performing them and a need to change the mindset of the greater roadside safety community such that decision-making is driven by the result of evidence observed under field conditions. ISPEs are needed not only to better understand the performance of roadside hardware, but also to improve the crash testing specifications which are used in the design and acceptance of new...
hardware for use on the nation’s highways as well as to improve design guidelines used for 
w warranting and placing roadside hardware in the field.

**BACKGROUND**

**Crash Testing Progression**

NCHRP Report 153 (6) was published in 1974 to provide uniform barrier testing procedures and 
more detailed guidelines for performing and evaluating full-scale vehicle crash tests. Report 239 
included four service levels for bridge railings and attempted to establish the service levels based 
on the capacity of the bridge railings based on the Report 230 supplemental tests. The AASHTO 
Guide Specification (7) introduced the concept of multiple performance levels for bridge 
railings. NCHRP Report 350 (2) was published in 1993 and expanded the concept of 
performance levels, specifying six different test levels for roadside hardware.

The Manual for Assessing Safety Hardware (MASH), the latest version of test and 
evaluation guidelines, was published in 2009. (3) Changes in vehicle fleet characteristics 
prompted NCHRP Project 22-14(02), "Improved Procedures for Safety-Performance Evaluation 
of Roadside Features." MASH includes essentially the same test level approach with some 
changes to vehicle characteristics. The selection of test vehicles for test levels 1-3 (passenger 
vehicle tests) were established by sales data, which represented the major change between Report 
350 and MASH.

**Test Impact Conditions**

The practical worst case impact conditions have “been defined as the combination of the 5th 
percentile lightest and heaviest passenger vehicles impacting a safety feature at the 85th 
percentile highest speed and 85th percentile highest angle. This combination of nearly worst case 
weight, speed, and angle is believed to produce a rare impact event. This definition of the 
practical worst case impact conditions was originally implemented for large passenger vehicles 
with the first set of evaluation guidelines presented in Highway Research Board Circular 482. (8) 
The precedent established with the first set of guidelines for full-scale crash-testing has been 
extended through to the present.

Vehicle masses are normally selected to be equal to the 95th and 5th percentile values 
from the passenger car fleet. MASH reduced the light truck vehicle mass to the 90th percentile 
and the small car mass to the 2nd percentile of the 2002 new vehicle fleet. MASH notes this was 
done in recognition of the recent increase in the size of passenger vehicles and the expectation 
that higher gasoline prices might push vehicle masses down. (3)

Changes were also made to impact speed between NCHRP Report 350 and MASH. 
These changes, however, were motivated by the desire to distinguish between TL-4 and TL-3 not 
by field observation. The impact speed for the single-unit truck test was increased from 80 km/h 
to 90 km/h to provide this distinction.

Changes were also made to the impact angle for some tests when updating NCHRP 
Report 350 to MASH. These changes were motivated by the desire to provide consistency 
across the test matrix, not by field observations. For example, the small car impact angle was 
increased from 20 to 25 degrees to match the impact angle used with light truck. The impact 
angle for length-of-need testing of terminals and crash cushions was increased from 20 to 25 
degrees to match the impact angle used for longitudinal barriers.
Side-Impact Testing
Appendix G of NCHRP Report 350 contains “recommended test and evaluation procedures for side impact testing.” This appendix was included in NCHPR Report 350 for informational purposes and was suggested for use until those or other guidelines were nationally accepted. Those side-impact testing and evaluation procedures were developed under FHWA-RD-92-062 and were applicable to “roadside structures like luminaire supports, guardrail terminals, and utility poles.” There are no side impact crash tests recommended in MASH for guardrail terminals, crash cushions or support structures. Citing considerable literature on the subjects of side impacts and non-tracking impacts with roadside features, MASH concedes on page 54 that crash data analyses shows “…half of all run-off-the road crashes involve non-tracking vehicles in a yawing or sideslip motion at the time of impact. Furthermore, crash data studies also appear to indicate that the impact performance of roadside features can be adversely affected by non-tracking behavior.” While the importance of side impacts has been recognized in MASH, recommendations for side impact testing and evaluation procedures were removed.

Summary
The test and evaluation guidelines are presumed to represent changes in the vehicle fleet and the latest advances in scientific knowledge. In the authors’ opinion, as a result of the lack of ISPEs, the only substantial changes to crash test and evaluation guidelines over the last 30 years have been changes to the test vehicles characteristics to reflect changes in sales data and some minor variations in the speed and angle to provide consistency within the testing matrix. Verifying that correctly installed and maintained roadside hardware is functioning as intended in the field is necessary to evaluate the relevance of crash test and evaluation guidelines. Institutionalizing ISPEs, such that ISPEs of hardware are as commonplace as the crash testing of hardware, will provide data needed to improve the crash testing standards to a point where crash testing provides a stronger representation of impact conditions observed in the field, thereby an improved evaluation procedures for new roadside hardware.

IN-SERVICE PERFORMANCE EVALUATIONS
Performance, in the context of roadside hardware, is measured by achieving the Roadside Design Guide’s goal of reducing crash severity. Typical questions which may be answered in an in-service evaluation of roadside hardware performance include:

- What is the risk of severe or fatal injury for the hardware studied?
- Is the risk substantially different for different vehicles?
- Is the risk substantially different for different speeds?
- Is the risk substantially different for different crash types (i.e., end-on, length-of-need, side impact, angled on the end, etc.)?
- Is the risk substantially different between manufacturers of the same or similar systems?
- Is the risk substantially different if the hardware is not properly installed or maintained?
- What are the common installation/repair mistakes and do they affect the risk?
- Is any type of hardware more likely to be installed or repaired incorrectly?
- What proportions of crashes are not reported to the police (i.e., roadside crash successes)?

The primary objective of any ISPE should be to determine how well the hardware is performing in the field under field conditions. The secondary objectives should be to determine
why issues with performance are observed (e.g., maintenance, installation, hardware design, etc.). The gathered data may include crash, maintenance, repair and construction records. It may also include an inventory and condition assessment of the existing hardware. One may also simply gather crash records to determine if there is an unacceptable level of risk within a certain group of hardware (e.g., guardrail, guardrail ends, bridge piers, etc.) and if further analysis is warranted.

**ISPE Study Design**

There are two common approaches to ISPE study design: (1) a retrospective study that uses information from past crashes to assess performance and (2) a prospective study that collects new information as new crashes occur. Retrospective ISPEs use data that is already available and has already been collected (e.g., police reported crash data, maintenance records, traffic data, etc.). Prospective studies collect data on new crashes as they occur and supplement the data available in retrospective studies with detailed inventories and condition assessments of the hardware, vehicle, and crash scene at the time of the collision. Generally, retrospective studies are quicker but less detailed (i.e., the characteristics and condition of the hardware at the time of the crash are often not known) and prospective studies require longer data collection periods and larger data collection areas, but much more detailed information can be obtained. Table 1 lists the primary advantages and disadvantages of both types of studies.

When gathering and analyzing ISPE data, one should be mindful that the primary objective of any ISPE is to determine how well the hardware is performing in the field and the secondary objectives are to determine why issues with performance are observed. The typical questions discussed above and these two study designs have been tabulated to demonstrate which questions can typically be answered by which study design and the available data or example data to address these questions (see Table 2).

The objective of a comprehensive ISPE is to cover many facets of the performance of a system which are not assessed through traditional crash testing. It should be expected that new revelations about the performance of the systems studied will be made and techniques to improve crash outcomes will be discovered. The new understandings of these systems which will result from ISPE research may be used to update National and State policies for the design, installation, maintenance, and repair of the studied category of roadside safety device. It may also be used by hardware designers to improve their designs and by AASHTO to develop new crash testing procedures which are more relevant to field experience.
Table 1. Advantages and Disadvantages of Retrospective and Prospective ISPEs.

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Retrospective Study</strong></td>
<td></td>
</tr>
<tr>
<td>- Large number of crashes already available.</td>
<td>- Detailed characteristics of the condition of the hardware prior to and after the crash are often not known.</td>
</tr>
<tr>
<td>- Results can be obtained quickly.</td>
<td>- Inventory of the specific hardware may not be available.</td>
</tr>
<tr>
<td></td>
<td>- Events are sometimes miscoded in police reports so validation is necessary.</td>
</tr>
<tr>
<td><strong>Prospective Study</strong></td>
<td></td>
</tr>
<tr>
<td>- Detailed information about the condition and type of hardware in-place can be obtained.</td>
<td>- Collecting sufficient cases for statistically meaningful results may require considerable time or a very large geographic area.</td>
</tr>
<tr>
<td>- Detailed information about the crash sequence and event can be obtained as crashes occur.</td>
<td>- Successful data collection depends on many parties cooperating (e.g., police, maintenance, traffic records, etc.).</td>
</tr>
<tr>
<td>- Post-event hardware and vehicles can be inspected.</td>
<td></td>
</tr>
</tbody>
</table>

Data Collection and Analysis

Subsequent to determining the study approach (i.e., retrospective or prospective), well designed ISPEs demonstrate care in the collection and processing of the study sample to avoid the introduction of bias. The analysis of the collected sample and inferences from the analysis to the population are relatively straight-forward. NCHRP 490 provided a comprehensive review of ISPEs which were available at that time. These summarized ISPEs and most ISPEs conducted after NCHRP 490 was published use the absolute risk of severe or fatal crashes as a measure of hardware performance. (5; 9-14) This section is intended to pull together the available literature on data analysis to allow for those conducting an ISPE to have a reference for these tools in one place, thereby hopefully helping to institutionalize ISPEs.

Experiment Design

Dixon and Massey explain that “[t]he method of choosing a sample is called the ‘design of the experiment.’” Dixon and Massey stress the importance of proper experiment design by stating “[v]ery misleading results can occur if samples are not taken correctly.” (15) Sampling for the successful determination of probabilities relies on random sampling. “In a random sample, every individual in the population has an equal and independent chance of being chosen for the sample.” Further, “if a random sample was desired, but some individual [sic] in the universe are more likely to be chosen then [sic] others, the sample is said to be biased.” (15)
Table 2. Example Questions Addressed by Retrospective and Prospective ISPEs.

<table>
<thead>
<tr>
<th>Example Objectives</th>
<th>Retrospective Study</th>
<th>Prospective Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the risk of severe or fatal injury for the hardware studied?</td>
<td>• Conduct “quick” study without inventory with any State geocoded crash data.</td>
<td>• Collect inventory data then collect crash data. No assumptions needed and any state crash data where specific hardware is a crash code can be used.</td>
</tr>
<tr>
<td>Is the risk substantially different between manufacturers of the same system?</td>
<td>• Use Google Earth, photologs or inventory data to determine the likely hardware at time of crash.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Use maintenance/ construction records to make assumptions about inventory. Arizona, Washington, Iowa, are good examples.</td>
<td></td>
</tr>
<tr>
<td>Is the risk substantially different if the hardware is not properly installed or maintained?</td>
<td>• Not possible with this data.</td>
<td>• Only possible if a field inventory and condition assessment is made before the crashes are observed.</td>
</tr>
<tr>
<td>What are the common installation/repair mistakes and do they affect the risk?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is any hardware more likely to be installed or repaired incorrectly?</td>
<td>• A field review of existing practices can be done in little time.</td>
<td></td>
</tr>
<tr>
<td>Is the risk substantially different for different vehicles?</td>
<td>• Best approach may be with state-wide crash data for non-manufacturer specific analysis. Datasets such as the HSIS data may be used for this.</td>
<td></td>
</tr>
<tr>
<td>Is the risk substantially different for different speeds?</td>
<td>• An evaluation of crash narratives and police report sketches are needed.</td>
<td></td>
</tr>
<tr>
<td>Is the risk substantially different for different crash types (i.e., end-on, length-of-need, side impact, angled on the end, etc.)?</td>
<td>• A review of crash records and maintenance records are necessary. Arizona and Washington are good examples of data.</td>
<td>• This study design would facilitate discovery by periodic surveys of the data collection area.</td>
</tr>
<tr>
<td>What proportions of crashes are not reported to the police?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Random sampling produces a representative sample by eliminating *undercoverage* bias and *voluntary response* bias. Voluntary response bias is hardly a concern when studying crash data as it can be safely assumed that none of the crash victims volunteered to strike roadside hardware. Undercoverage bias, however, is a type of selection bias which is often a serious problem with convenience samples. A convenience sample is a non-probability sampling method comprised of easy to identify cases (e.g., only fatal crashes with hardware). Non-probability sampling methods are often chosen because they offer convenience and cost savings, however, one cannot make inferences across the population from a convenience sample. Only probability sampling methods or sampling the whole population (i.e., a census) allow for conclusions to be applied across the population.

Probability sampling methods are complex. For simplicity, a census of data from any particular region and any particular period is often chosen to ensure an unbiased sample. Using a census, all the cases in a particular area and during a particular time are collected so there is no sampling bias. The results from the analysis of a census sample are not estimates but are the actual observations for that time and geographic location. The observations from the non-biased census can then be used as an estimate of the performance at other times and in other places.

**Data Filtering**

Case filtering is often a necessary feature of data analysis. When considering roadside hardware, filtering may involve removing cases that lie outside the design criteria. For example, it is common to eliminate cases involving non-passenger vehicles like motorcycles, heavy trucks and buses when studying test level three (TL3) devices because these vehicles are not covered by the test level three hardware crash testing guidelines.

A specific type of hardware (e.g., terminals, bridge ends, utility poles, etc.) can be filtered from the larger crash dataset in several ways. The analyst may choose cases where the first harmful event was an impact with hardware X; or the most harmful event was an impact with hardware X; or any harmful event in the crash sequence was an impact with hardware X; or hardware X was the first and only harmful event in the sequence. Each filter strategy examines a slightly different question – first and only harmful events are the best measures of the performance of the hardware itself but this filter strategy may eliminate cases where the hardware was not located on the site correctly (e.g., a case where a terminal is struck and the vehicle re-directed behind the terminal where it strikes a tree would be eliminated). Conversely, first harmful event filtering allows the analyst to examine the whole range of performance issues including site conditions but may include crash severity that is more appropriately attributed to another part of the crash sequence (e.g., a rollover or impact with another object). A hardware impact in any of the sequence of events likewise may confuse the harm done by collisions with other vehicles prior to the hardware impact with the harm done by collision with the hardware. Each filter strategy is a legitimate path of inquiry, but the analyst needs to determine exactly what question they intend to address.

**Absolute Risk**

The absolute risk of an incapacitating or fatal crash (%A+K) involving roadside hardware is simply the portion or percentage of all crashes involving that hardware type that result in fatal or severe injuries. The absolute risk when defined this way is also the probability of observing an A+K crash given all crash severities. The absolute risk of A+K can be found by summing the total number of A+K crashes in the data for that hardware and dividing by the total number of all crashes of all severities for that hardware, as shown here:
Absolute Risk = \frac{A + K}{KABCOU}

Where:

A+K  
Severe and fatal injury crashes

KABCOU  
Crashes of all reported severities including U for unknown severities. If unreported crashes have been studied, these unreported crashes should also be included in the denominator.

Statistical Significance

The next step in the analysis is to determine if the results are statistically significant. Statistical significance has a specific meaning in the context of statistical analysis. The calculation is discussed first and the use of the calculated results is discussed next.

The absolute risk represents a proportion or the probability of severe or fatal crashes (p). The absolute risk determined above is a point estimate calculated from a sample of the population of interest. Since the absolute risk (p) of the entire population is unknown, the estimated absolute risk (\hat{p}) from the sample is used and expressed with a confidence interval to allow inferences to be made on the larger population. (16; 17)

The 95% confidence interval, for example, can be interpreted to mean if you were to take 100 additional samples, 95 times out of 100, the absolute risk would be between x% and y%.

You can report the results this way: “based on our sample data, we are 95% confident that the ‘true’ absolute risk of an A+K crash with the hardware studied is between x% and y%.”

Specifically, the confidence interval corresponds to percentage of the area under the standard normal distribution curve which is assumed to be “bell-shaped” given that error is random and the events are independent. For example, a 95% confidence interval covers 95% of the area, therefore, the probability of observing a value outside of the area is less than 0.05 (i.e., 1-0.95=0.05). The normal distribution curve is symmetric, therefore for a 95% confidence interval, the area in each tail of the curve is equal to (1-0.95)/2= 0.025. This area can then be used to consult a published z-table (i.e., normal distribution table) to find the number of standard deviations away from the mean (z) equals 1.96. Common confidence levels and the corresponding z-values are shown in Table 3.

<table>
<thead>
<tr>
<th>Table 3. Published z values for Normal Distribution.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence Level</td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td>0.70</td>
</tr>
<tr>
<td>0.75</td>
</tr>
<tr>
<td>0.80</td>
</tr>
<tr>
<td>0.85</td>
</tr>
<tr>
<td>0.90</td>
</tr>
<tr>
<td>0.92</td>
</tr>
<tr>
<td>0.95</td>
</tr>
<tr>
<td>0.96</td>
</tr>
<tr>
<td>0.98</td>
</tr>
<tr>
<td>0.99</td>
</tr>
</tbody>
</table>
The confidence interval is calculated using the following equation:

\[ \hat{p} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]

Where:
- \( \hat{p} \) Absolute risk calculated from the sample.
- \( z \) Number of standard deviations away from the mean (see Table 3).
- \( n \) Sample size.

Recall the exemplar question for ISPEs discussed above, including:
- What is the risk of severe or fatal injury for the hardware studied?
- Is the risk substantially different for different vehicles?

Taking for example the first question, “what is the risk of severe or fatal injury for the hardware studied?” and assuming you have already calculated the risk (A+K%) and the 95% C.I. for hardware A, B, C, and D, as shown in Figure 1. Now you want to determine if your results are statistically significant.

To determine statistical significance, you will need a null hypothesis to evaluate. The null hypothesis generally takes the “devil’s advocate” position. Through hypothesis testing, an attempt is made to overturn the null hypothesis. The risk is a proportion expressed as a percent, therefore the value must be between 0-100. In this example the “devil’s advocate” null hypothesis would be something like “the risk for different vehicles does not fall between 0-100%.” The alternative hypothesis is that the risk does fall between 0-100%. Hypothesis testing is used to examine each hardware evaluated individually in this example (e.g., A, B, C, and D).

If the null hypothesis is rejected, then the alternative hypothesis is accepted. If the null hypothesis is not rejected, then the alternative hypothesis is not accepted. The null hypothesis is rejected for D, but accepted for A, B, and C. In other words, we have significant findings for A, B, and C.

Reusing Figure 1 to work through an example for the second question, “Is the risk substantially different for different vehicles?” and assuming you have already calculated the risk (A+K%) and the 95% C.I. for vehicles A, B, C, and D, as shown in Figure 1. The null hypothesis in this example would be “the risk is the same for each vehicle.” A cursory review of the figure shows the data points and error bars do not overlap for vehicles A, B, and C, however vehicle D overlaps all of the other data points. One could reject the null hypothesis and conclude the findings for vehicle D are not statistically significant, but the findings for A, B, and C are.

As a final note on significance, just because a null hypothesis is not rejected does not mean that the statement is true. The ability to obtain significant results should be interpreted in much the same way a jury verdict is interpreted in the United States. “Just because a person has been declared “not guilty”, it does not mean that he is innocent.” (18) Statistically non-significant results can be used to determine a trend in the data. One may find that hardware A is preforming better than all other hardware in many studies, but never with statistical significance. Consider that your data may support the thesis to overturn your null hypothesis (e.g., vehicle D is less safe than other vehicles) even though the effect is not statistically significant. You should have more confidence than before the research was conducted. However, the support is weak and
the data were inconclusive. A reasonable course of action would be to continue data collection. If the same results were found, you would have more confidence.

**Figure 1. Example Figure Summarizing Fictional Absolute Risk and 95% C.I.**

**Summary**

The collection and analysis of the data for an ISPE have been discussed. The literature for the analysis of the data has been summarized and it has been demonstrated that the analysis portion of an ISPE is the most elementary part of the study. It has been shown that complicated algorithms and fancy software are not needed to assess the performance of roadside hardware, but simple equations may be used by any person with access to the field collected data. It has become common in roadside safety to use the absolute risk of severe or fatal crashes as a measure of hardware performance. This review of available literature allows for the analysis of data collected during an ISPE using tools available to anybody, thereby removing additional obstacles to institutionalizing ISPEs.

**CONCLUSIONS AND DISCUSSION**

New hardware designs must meet some minimum standard before being deployed in the field. This need for an approval standard has resulted in the development of crash testing standards. (1; 2; 3; 6) Crash tests are the roadside safety equivalent of “experiments” which are conducted to an established standard to ensure repeatability of the tests and comparability between tests. The variability of vehicles, occupants, and impact conditions are controlled in the testing environment but cannot be controlled in-service. Field conditions, common installation/repair mistakes, or maintenance issues are not observed in crash tests. Variability in occupants are not observed in crash tests. Driver reactions and behaviors are not observable in crash tests. The range of vehicle size, impact speed, and angles are thought to be addressed by crash tests but this has not been verified. Each of these variable characteristics can be assessed through the in-service review of roadside hardware.
While these testing standards have reiterated a call to evaluate the hardware in the field following the initial crash test evaluation, ISPEs have not seen the same global institutionalization as crash testing. The lack of documented field experience leads to the authors of crash testing guidelines and hardware developers relying on crash testing experience alone for progress in hardware development. Field observations of issues related to side-impacts with hardware lead to the development of the side-impact testing and evaluation protocol introduced in NCHRP Report 350 Appendix G. One could argue that the optional side-impact testing provided in Report 350, but removed in MASH, was a step backward in the evolution toward testing standards which replicate field conditions. When evolution of these testing standards is grounded in the observations of field performance, not testing experience alone, side-impacts, and other field observations will not be ignored in the testing environment.

Institutionalizing ISPEs in the roadside safety design, test, and evaluation process will require the collaboration of hardware designers, the AASHTO Technical Committee on Roadside Safety (TCRS), the States, manufacturers, FHWA, and researchers. Simply stated, each member of the community has a vested interest in the performance of hardware on the roadside and each member can play a valuable role in the institutionalization of ISPEs. States may be best suited to collect the inventory, manufactures may provide a portion of funds from the sale of hardware, and researchers may conduct analyses when granted access to the data. The TCRS may provide the leadership to change the collective mindset from crash testing as the gold standard of safety evaluation to field performance as the gold standard. Funding agencies such as the FHWA or NCHRP may provide the catalyst to outline the cooperative action and conduct any initial setup. In any case, observing improved performance where it really matters (i.e., in the field) is not possible without ISPEs.
REFERENCES


<table>
<thead>
<tr>
<th>Reference</th>
<th>Year</th>
<th>Authors</th>
<th>Title</th>
<th>Source</th>
</tr>
</thead>
</table>