Hindawi
Advances in Civil Engineering
Volume 2017, Article ID 7925193, 12 pages
https://doi.org/10.1155/2017/7925193

Research Article

Using Expert Opinion to Quantify Uncertainty in and Cost of Using Nondestructive Evaluation on Bridges

Alex A. Hesse,[1] Rebecca A. Atadero,[1] and Mehmet E. Ozbek[2]

[1]*Department of Civil and Environmental Engineering, Colorado State University, Fort Collins, CO 80253, USA*
[2]*Department of Construction Management, Colorado State University, Fort Collins, CO 80253, USA*

Correspondence should be addressed to Mehmet E. Ozbek; mehmet.ozbek@colostate.edu

Received 28 March 2017; Accepted 11 May 2017; Published 19 July 2017

Academic Editor: Deodato Tapete

A previous literature review indicated that there is little published experimental data that can be used to determine quantities such as bias, accuracy, reliability, and cost of common Nondestructive Evaluation (NDE) methods as far as their use on bridges is concerned. This study attempts to quantify these parameters for common bridge NDE methods through a four-round Delphi method survey with experts in the NDE bridge field. The survey results indicate that most commonly used bridge NDE methods tend to be underbiased and relatively reliable. Furthermore, the accuracy of commonly used bridge NDE methods tends to be relatively variable with the average test measuring a true response between 80% and 85% of the time. In general, it was shown by the participant responses that the more expensive the method was, the better the bias, accuracy, and reliability the method had, and vice versa. The information presented in this paper can serve as a starting point for characterizing different NDE methods for use in bridge management and inspection planning and identifies the type of information that is still needed. As such, this research has the potential to promote further research on this subject.

1. Introduction

The poor condition of the US bridge inventory requires an enhanced approach to bridge management. Bridge inspection is a key part of the management process, which is somewhat overlooked when it comes to strategic planning and optimization. The National Bridge Inspection Standards (23 CFR 650C) require inspection of all bridges (with a length of 20 feet or more) on public roads every 24 months [1]. In 2009, a joint American Society of Civil Engineers Structural Engineering Institute (ASCE/SEI) and the American Association of State Highway and Transportation Officials (AASHTO) Ad Hoc group was created to study how current bridge inspection practices could be improved for the future [2]. The group recommended that "*a more detailed inspection conducted less frequently may have a positive impact on the overall safety and maintenance of bridges in the U.S., allowing for broader application of Nondestructive Evaluation (NDE) technologies and a better understanding of the condition of individual bridges*" [2]. In order to move to a more realistic inspection model where the inspection of a given bridge is not necessarily conducted on a set 24-month cycle as suggested above, it is necessary to have a better understanding of the uncertainty in inspection results and an understanding of the costs of these more advanced NDE methods.

2. Problem Statement and Purpose

It was determined through an extensive literature review by Hesse [3] and Hesse et al. [4] that there is limited existing experimental work that can be used to quantify the level of bias, accuracy, reliability, and cost of common NDE methods as far as their use on bridges is concerned. The lack of such data is a significant obstacle to moving away from the set 24-month inspection cycle and to increasing the use of NDE more generally. The problem is that the experimental work necessary to find these values is difficult and would be extensive. Expert opinion could be used as a starting point, providing initial assessments of uncertainty that could then be updated based on agency specific findings that are collected as part of the normal inspection process rather than as separate experimental studies. This is similar to the approach

of the element deterioration models in AASHTOWare Bridge Management Software (formerly known as PONTIS) that can be updated as outlined by Hatami and Morcous [5]. This paper describes a study to collect expert opinions on the uncertainty and costs of commonly used NDE methods. The information presented in this paper can serve as a starting point for characterizing different NDE methods for use in bridge management and inspection planning and identifies the type of information that is still needed. As such, this research has the potential to promote further research on this subject.

3. Methodology

Expert opinion can be collected through surveying methods. However, there are limitations with a standard, single round survey because the results may not be as comprehensive and data can be skewed in the absence of a feedback loop. For these reasons, the Delphi method was chosen as an efficient and effective survey technique to gather this information. This survey aims to provide quantitative descriptions of uncertainty in NDE results in terms of statistics describing bias, accuracy, and reliability and a comprehensive comparison of the costs of various tests to provide information to researchers and practitioners working in the fields of bridge management and inspection.

The Delphi method was originally developed in the 1950s by Olaf Helmer and associates at the RAND Corporation [6]. The method is defined as "*a group process involving an interaction between the researcher and a group of identified experts on a specific topic, usually through a series of questionnaires*" [6]. The process is useful to gather opinions on complex topics when exact information is unavailable [7], making it a good tool to gather quantitative information of NDE methods based on experts' opinions.

The overall goal of the method is to reach a consensus within a group of experts [8]. This can be done by using a sequence of questionnaires to collect opinions from the group of experts. The process utilizes several iterations to provide feedback to the participants. This feedback allows the participants to reconsider their original opinion. Consequently, the results from previous iterations can change or be adapted by individual participants in later iterations based on the feedback of the group. With this feedback loop, the responses from the participants converge to the assumed true response. Furthermore, the feedback loop helps to eliminate noise in the data by allowing each participant to make additional conclusions and clarify the information from previous iterations based on these results [9].

The Delphi method is an iterative process until consensus of experts' opinions has been achieved. The method is based on an open-ended questionnaire that is used in the first round. The questions are often developed from literature reviews or past surveys. The purpose of this questionnaire is twofold: (1) to gather information about the type and level of expertise of the respondents and (2) to ask for specific information about the topics in question. After receiving the respondent's answers, the information is compiled and organized. The results for the specific information are then used to develop the second and subsequent rounds of the survey [9].

The questions in the second round are typically closed-ended questions that require the participants to rank and order specific responses developed by the surveyors. The responses are then compiled into a result sheet that summarizes the responses of all the respondents. The third and subsequent rounds are similar (or often the same) as round two except the result sheet is sent with the surveys. The participants are then asked to review the result sheet and answer the questions again based on their prior opinions and the results from all respondents. These rounds give the respondents an opportunity to revise their responses based on the overall responses of the group. They are also given the opportunity to specify reasons if they chose to remain outside the consensus. This process is then repeated with all of the respondents responses until it is determined that a consensus is reached [9].

Data analysis of the results from each round after the first round can be employed to determine the agreement and stability of the results for each question. Determining when the responses have converged to a single value and the responses are stable is key when implementing a Delphi study as it indicates when the study should be terminated.

English and Kernan [10] show that the coefficient of variation (COV) can be used to determine agreement by evaluating the COV of each question for each round in conjunction with a decision rule of predetermined ranges. Based on English and Kernan's range recommendation, all results with a COV lower than 0.5 are considered to be converging to the mean value. If a response was in a range from 0.5 to 0.8, the response was considered to be nearing convergence and should be analyzed in more to detail to understand the trend. Furthermore, a response with a COV greater than 0.8 is not considered to be converging to a mean value. All questions that are considered to be converging are then analyzed to determine the stability of the response.

Stability is a representation of how much the responses change from one round to the next. Kalaian and Kasim [11] present various parametric and nonparametric methods to determine stability. Based on the response rates and results of each round of the Delphi method implemented here it was determined that a nonparametric method should be employed for this study. Of the nonparametric options, Spearman's Rank Correlation Coefficient method, was most appropriate for the data and was used to calculate stability.

To determine stability using Spearman's Rank Correlation Coefficient method, Spearman's rho, r_S, must first be calculated using

$$r_S = 1 - \frac{6 \sum d_i^2}{n\left(n^2 - 1\right)}, \tag{1}$$

where d_i is the difference between ranks of the respondents for the ith question and n is the number of respondents. Note that when participants dropped out from round to round, n from the later survey was used. The rank correlation is then compared to a critical value determined from a table of critical values for Spearman's rho [12]. If the calculated value

is greater than the critical value, the response is determined to be stable. From this, the closer the value is to one, the more stable it is, and, conversely, a value close to zero indicates no stability. For this study a one-tailed test with a level of significance of $\alpha = 0.05$ was used based on Kalaian and Kasim's [11] recommendation. If the question is found to also be stable, the question can be removed from subsequent rounds.

4. Implementation of the Delphi Method for This Study

4.1. Previous Surveys about the Use of Bridge NDE Methods. Three previous surveys on the use of NDE methods on highway structures were discovered during the literature review. These surveys were only conducted over one round. Relevant findings from these surveys were used to form the framework of the Delphi method survey conducted for this study. The previous surveys included a study by Rens et al. [13] for the American Association of Railroads, a study by the California Department of Transportation (CALTRANS) [14], and a Federal Highway Administration (FHWA) survey [14]. The results from these surveys were used as a starting point in drafting the Delphi survey developed for this research.

4.2. Determining Participants. Prospective participants of the survey were determined through an Internet search of private companies in the United States that work in the NDE field and current department of transportation (DOT) employees that were involved in bridge evaluation. The directory of the AASHTO Subcommittee on Bridges and Structures was also used to determine DOT employees who have experience with NDE methods. These prospective participants were then contacted by mail and asked to participate. A total of 36 DOTs were contacted. A total of 27 private companies from around the country were also contacted. These companies all primarily work in the NDE field and have experience conducting tests on bridges and developing and selling NDE equipment. All surveys were sent directly to individuals who were deemed experienced with NDE methods.

As discussed earlier, there have been three previous surveys conducted on NDE use. The response rates from these surveys were used to determine the number of participants to be contacted. Based on those surveys' response rates and the recommended Delphi survey size suggested by Ludwig [15] as 15 to 20 participants, it was determined that about 60 possible participants should be contacted. Prior to implementing the survey, the procedure and a description of the possible participants were submitted for review to the Institutional Review Board (IRB) and approved by the IRB for implementation.

4.3. Round One Questionnaire. A total of 63 people were contacted and asked to participate in the survey. They were mailed a packet, which included the three-section questionnaire, a cover letter that explained the survey and acted as the release form, and a self-addressed and stamped return envelope to return the survey.

The first section of the questionnaire asked participants about their background and general experience with NDE. They were asked about their current education level, current NDE certification level, how long they have been working with NDE, the types of tasks they perform when working with NDE, and the number of bridges their organization and each participant individually evaluate in a given year. It should be noted that the private contractors were asked an additional question asking them in what geographic region they perform NDE. It was implied that the DOT personnel only perform NDE in their respective state.

Section two dealt with various NDE methods for steel bridges while section three dealt with NDE methods for concrete bridges. Participants were given a list of common NDE methods that are used and were also given space to write any test that the participants commonly used but was not listed. Respondents were asked to list the types of conditions their organization sought to identify with each technique. If their organization did not use a specific method, they were asked to leave the space blank. They were also asked to identify each method from the list their organization used at least once every month. If they did not use a specific technique at least once a month, they were asked to indicate which two methods they use the most. There were two purposes for the questions in these sections. The first purpose was to develop a list of the most widely used NDE methods based on the responses of the participants. The second purpose was to compile a list of common conditions that were tested for with each NDE method.

The participants were given about a month to complete and return the first questionnaire. In order to help the response rate, each survey was written to take an estimated 20 minutes to complete and a reminder was sent the week before the due date. Of the 63 people contacted, 14 people responded (11 DOTs and 3 contractors). While the response rate was lower than expected, the number of participants was deemed to be acceptable based on Ludwig's [15] recommendation.

4.4. Round Two Questionnaire. All 14 people who responded to the first questionnaire were contacted and asked to participate in the second questionnaire. This questionnaire was composed of two sections, which were similar as they both dealt with questions about specific NDE methods. The first section was for NDE methods on concrete while the second section was for NDE methods on steel. Each section contained five subsections. These subsections included

(i) bias: the tendency of a test to consistently measure a value either higher or lower than the actual or perceived value,

(ii) accuracy: the tendency of a test to measure true results,

(iii) precision: the reproducibility of a test in a controlled environment,

(iv) reliability: defined for this study as the reproducibility of a test in an uncontrolled environment,

(v) costs: including time spent running a test, time spent analyzing data, time to train an inspector, monetary

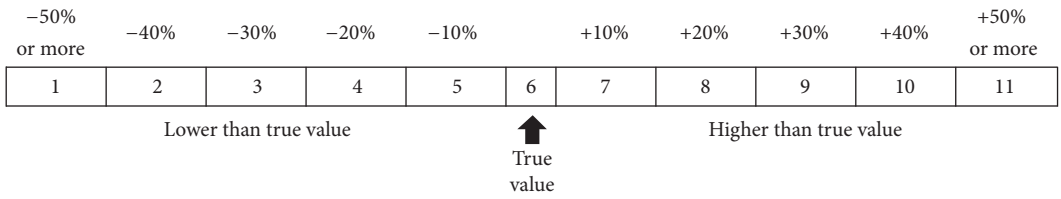

Figure 1: Response scale used for the bias subsection.

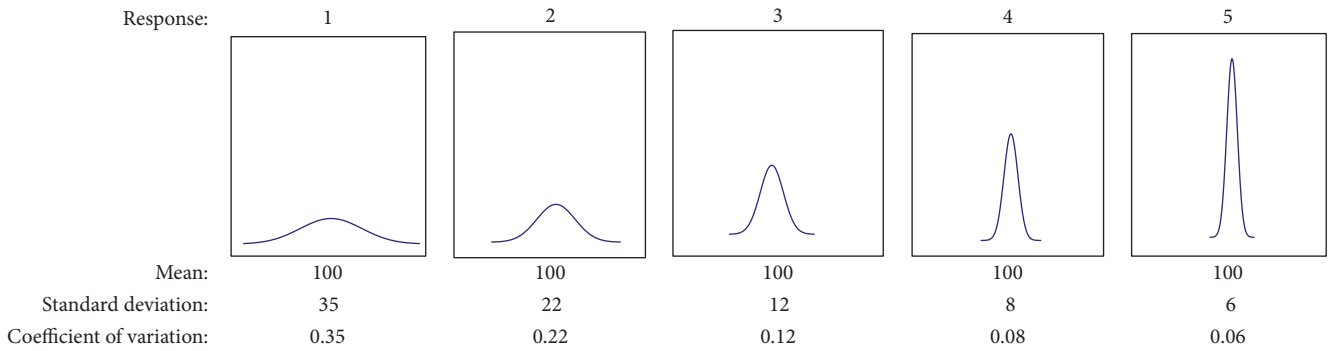

Figure 2: Representation of the scale used for the precision and reliability subsections.

cost for the equipment, and number of inspectors needed.

It should be noted that participants were asked to give a response for any method they had an educated opinion about for the second and subsequent surveys, not just methods they had extensive experience with.

The participants were once again given about a month to complete and return the second questionnaire, and brief reminder letters were sent a couple days before and after the deadline. Of the 11 DOTs contacted, eight responded to the second survey. Moreover, of the three contractors contacted, there were two responses.

4.4.1. Subsection One: Bias. The first subsection was used to determine the amount of bias present in methods that provide a quantitative result, for example, the length of cracking or the depth to reinforcement. It should be noted that if a method was determined not to give a quantitative result, that method was not included in this section. The methods that were included were cover meters/pachometers, impact echo, radar, ultrasonic testing, and visual inspection for concrete members; and acoustic emission, radiography, ultrasonic testing, and visual inspection for steel members. Ultrasonic testing for steel members was broken into three sections: crack detection, pin inspection, and weld inspection. This was done because, unlike other methods, this method is consistently used to identify all three of these defects, and the test for each defect could have a different bias. Participants were given a bias scale from one to eleven to use when answering the questions. This scale is provided in Figure 1.

For this scale, a response of six represented the true value. Any incremental responses lower than six indicated an extra 10% bias from the true value (i.e., a response of five would have a bias of 10% lower than the true value, a four would have a bias of 20% below to the true value, and so on). This same relationship also applied to the responses larger than six.

4.4.2. Subsection Two: Accuracy. Accuracy is defined as the tendency of a test to measure true results. For quantitative measurements, accuracy is affected by bias, as questioned in the previous section. For qualitative tests, accuracy is concerned with the correct identification of element condition. In this section, participants were given three options: false positive, false negative, and true response. False negative was defined as a test that indicates no damage even though damage is present, while false positive was defined as a test that indicates damage when there is no damage. A true response was defined as a test that measures damage when there is damage or a test that measures no damage when there is no damage. Participants were asked to estimate the percentage of time each test would have for each result. They were told their percentages should add to 100%. Similar to the "Bias" subsection, this subsection only included certain methods.

4.4.3. Subsections Three and Four: Precision and Reliability. The third and fourth subsections were used to determine the precision (repeatability in a controlled environment) and reliability (repeatability in the field) of each method and included all of the methods in question. Participants were given a scale based on hypothetical means and standard deviations. They were also given COVs corresponding to these numbers and a graphical representation of the corresponding distribution, which was assumed to be a normal distribution. The same scale was used for both precision and reliability and can be seen in Figure 2. Participants were asked to indicate either the reliability or precision for each method based on the scale provided.

	Very low	Low	Moderate	High	Very high
Scale	0–4 hours	4–8 hours	8–10 hours	10–15 hours	15+ hours

Figure 3: Example of a personal scale made by one of the participants for time spent running a test.

4.4.4. Subsection Five: Costs. The costs considered for each method were time spent running a test, time spent analyzing data, time to train an inspector, monetary cost for the equipment, and number of inspectors needed. For each type of cost, the participants were asked to develop a personal scale of five ranges by filling in the costs (in their respective units such as hours, $, or #) they considered to be very low, low, moderate, and so on. An example of this scale can be seen in Figure 3.

The participants were then asked to categorize each method based on the user-developed scale. The user-developed scale was implemented because it was predicted that two participants might have substantially differing opinions on what constitutes a "very low" or "very high" cost. In analyzing the data for the second round it was observed that some responses were vastly different than others. To facilitate further rounds of the survey, the data for each different cost scale was compiled and compared. A standard scale was then developed for each cost based on the participant-developed scales. These scales were developed to include as many responses as possible while staying relatively close to the average response for each range. The scales were then implemented for questionnaires three and four.

4.5. Round Three Questionnaire. All 10 people who responded to the second questionnaire were contacted and asked to participate in the third questionnaire. They were also sent a results packet that contained the individual's response (a unique response packet was used for each participant) along with the average group response for the bias, accuracy, and reliability subsections. This questionnaire was very similar to the second questionnaire but this time participants were asked to complete the questionnaire in conjunction with the results packet. The goal was for the participant to iterate their response based on their prior response and the average group response.

While the survey was nearly identical to the previous survey, there were a few minor changes. The first change was the removal of the precision subsection. Based on the results of the second questionnaire, it was shown that precision and reliability were nearly identical. To shorten the survey and in an effort to help preserve the response rate, the precision subsection was removed. The second change was the inclusion of predetermined scales based on prior responses for the costs subsection. The participants were once again given about a month to complete and return the third questionnaire. Also, a reminder letter was sent to participants who had not responded about a week before the deadline and again a few days after the deadline. Of the eight DOTs contacted, seven responded to the third survey. Moreover, both contractors responded to the survey. It should be noted that the participant who discontinued her/his participation was not a significant outlier relative to the average group response.

4.6. Round Four Questionnaire. All nine people who responded to the third questionnaire were contacted and asked to participate in the fourth questionnaire. They were also sent a results packet from the third questionnaire. This questionnaire was nearly identical to the second and third questionnaires; however, the bias and reliability subsections were removed from this round because it was shown that the responses from these subsections had converged and become stable. The participants were once again given about a month to complete and return the fourth questionnaire. Also, a reminder letter was sent to participants who had not responded three days before the deadline and again a few days after the deadline. All nine participants responded to the survey.

5. Results and Discussion

5.1. Results for the General Participant Information. The first questionnaire was used as a foundation for the second and subsequent questionnaires. The first section of this survey gave valuable information about the experience and certification level of all participants, which was sought to ensure that the participants could be considered knowledgeable in bridge NDE methods. According to the 14 original respondents, the average experience level of the participants with bridge NDE was 17.8 years with a maximum of 40 years and a minimum of 5 years. Most of these people were managers, but also assisted in data analysis, bridge inspection, and report writing. The most common education level was a 4-year degree. Two respondents had a Master's degree and two respondents had a high school diploma. One person did not respond. The certification level of the participants varied much more than the experience. Of the 14 original respondents, 73% of them possessed at least a Professional Engineering license. Along with this, three participants had an American Society for Nondestructive Testing (ASNT) NDT level II certification for at least one NDE method. Based on these results, it was determined that all participants could be considered knowledgeable about bridge NDE methods. Many respondents were managers; this was considered a natural result of their average experience level of nearly 18 years and was deemed desirable as it would put them in a position to have seen results of many different bridge evaluations. Also, as at least 73% percent of respondents had an engineering degree; the respondents were deemed well versed in quantitative thinking and capable of responding to the statistical scales presented in the remainder of the survey. The second and third sections of the round one questionnaire were used to determine the common NDE methods that are

Table 1: Number of respondents indicating experience with and types of damage identified with each NDE method.

	Frequency	Used to determine
Concrete NDE methods		
Visual	12	General flaws
Mechanical sounding	10	Delamination
Cover meters/pachometer	8	Located rebar, cover
Rebound hammer	6	Test compressive strength
Thermal	5	Delamination
Impact echo	4	Thickness, delamination
Radar	4	Located rebar, thickness
Ultrasonic	4	Delamination
Acoustic emission	*3*	*Monitor stay cables*
Electrical potential	3	Detect corrosion
Vibration	*2*	*Force measurement*
*Chloride samples**	*1*	*No response*
Radiography	*0*	—
Steel NDE methods		
Liquid penetrant	12	Weld imperfection, crack detection
Visual	12	General flaws
Ultrasonic	12	Weld imperfection, crack detection, corrosion detection, thickness measurement, pin inspection
Magnetic particle	10	Weld imperfection, crack detection
Radiography	7	Weld imperfection, crack detection
Thermal	*2*	*Deck inspection*
Acoustic emission	1	Monitor stay cables
Eddy current	*1*	*No response*
Vibration analysis	*1*	*Force measurement*
*Strain gauges**	*1*	*No response*

Notes. Rows in italic represent NDE methods that were deemed to be used infrequently. ∗ means write in response.

currently in use and the types of bridge conditions they are used to identify and measure. The results for both concrete and steel members are presented in Table 1.

Those methods with less than four experienced respondents (as indicated with rows in italic in Table 1) were removed from subsequent surveys. Electrical potential for concrete members and acoustic emission for steel members were included in the subsequent surveys because they were shown by the past surveys to be used more frequently than other methods that had similar responses from respondents. The reader is referred to Hesse [3] and Hesse et al. [4] for a detailed description of each of these NDE methods.

5.2. Results for the Bias of NDE Methods. After round three, all responses in the bias subsection both had converged and became stable. The results from round three (9 respondents) were considered the final values, and the bias subsection was removed from subsequent rounds. The response average and standard deviation are presented in Table 2.

Based on the results, respondents felt that all of the methods for concrete and steel bridges were slightly underbiased with an average response in the third round of all methods being between a response of 5 and 6 (5 being about 10% underbiased and 6 being the true value).

Table 2: Concrete and steel NDE methods response descriptive statistics for the bias.

	Average	Standard deviation
Concrete NDE methods		
Cover meters/pachometer	6.00	0.926
Impact echo	5.40	0.894
Radar	5.57	1.272
Ultrasonic testing	5.71	0.488
Visual inspection	5.89	0.782
Steel NDE methods		
Acoustic emission	5.50	1.000
Radiography	5.71	0.488
Ultrasonic testing-crack detection	5.88	0.835
Ultrasonic testing-pin inspection	5.89	0.782
Ultrasonic testing-weld inspection	5.86	0.900
Visual inspection	5.67	0.707

The literature review by Hesse [3] showed that there is limited data available for comparison; but, based on the data available, the survey results seem to show reasonable agreement to experimental findings. It was shown by Phares et al.

Table 3: Concrete and steel NDE method bias factor determined from the participant responses.

	Bias factor
Concrete NDE methods	
Cover meters/pachometer	1.000
Impact echo	1.060
Radar	1.043
Ultrasonic testing	1.029
Visual inspection	1.011
Steel NDE methods	
Acoustic emission	1.050
Radiography	1.029
Ultrasonic testing-crack detection	1.013
Ultrasonic testing-pin inspection	1.011
Ultrasonic testing-weld inspection	1.014
Visual inspection	1.033

[16] that, during a routine inspection, visual inspection of the superstructure, substructure, and deck had an overall bias of +3%, −5%, and +5%, respectively. Note that a positive bias means that the inspectors determined the bridge element was in better condition than it actually was. These numbers are close to (but slightly higher than) the numbers determined by the respondents. Barnes and Trottier [17] showed that radar tends to be slightly underbiased, but no specific numbers were given.

Since the responses had reasonable agreement, the data could be used to produce a bias factor. Table 3 shows how the response of the participants correlates to a bias factor for each method. The bias factor can be defined as the true value divided by the measured value. The bias factor was determined by fitting a trend line to the response and the bias representation. This factor could be used with an individual method's nominal value to give the inspector a more accurate representation of the true value. Therefore, multiplying the bias factor by the measured value would potentially yield a more valid result.

5.3. Results for the Accuracy of NDE Methods. As previously mentioned, participants were given three options for each method in the accuracy subsection: false positive, false negative, and true response. Participants were asked to estimate the percentage of times each test would have each result. Since the data provided by respondents was open-ended and the sum of the averages could result in a total percentage of more than 100%, for analysis, each participant's responses were normalized based on the group average to 100%. The normalized responses were then used to calculate the normalized average, normalized standard deviation, and normalized COV. Based on the convergence and stability data from round three, the questions were asked again for round four.

After round four, all responses except two (false negative response for electrical potential and thermal imaging) had a COV of less than 0.5 indicating convergence. For the two responses that were above this threshold, both COVs had dropped significantly and were now in the lower portion of the less than satisfactory range (between 0.5 and 0.8 COV) and were considered to have reached a consensus. The stability was then analyzed, and based on the results from round two to round three and round three to round four, it was seen that the responses became progressively more stable. By the end of round four, most of the false positive and false negative results could be considered stable, but most of the true response results had not reached stability (see Table 4). These issues in stability arose in part because of the large scale used to identify accuracy. If a respondent changed her/his answer by a seemingly small 5%, this change is drastically increased due to the exponential nature of the stability equation. As the rounds progressed, it was observed that the participants were becoming more reluctant to change their answers during the iteration process. Furthermore, it makes sense that if the false positive and false negative responses were becoming stable, the true response should trend towards stability, as well. Since the responses were considered to be converging, it was determined that there would be little change if another round was implemented, and the responses would be considered stable if they were asked in a subsequent round. Thus, the results from the fourth questionnaire were considered to be the final results. The normalized response average can be seen in Table 5.

Table 4: True response stability results from the accuracy subsection from round two through round four.

	Rounds 2 to 3 stability	Rounds 3 to 4 stability
Concrete NDE methods		
Electrical potential	−4.93	0.04
Mechanical sounding	−2.83	−1.69
Thermal	−3.23	−0.14
Visual inspection	−11.98	−1.77
Steel NDE methods		
Acoustic emission	1.00	−0.41
Liquid penetrant	−4.93	−2.75
Magnetic particle-crack	−1.18	0.10
Magnetic particle-weld	−0.95	−0.03
Radiography	−3.45	−0.42
Ultrasonic-crack	−8.46	1.00
Ultrasonic-pin	−6.58	−0.17
Ultrasonic-weld	0.22	1.00
Visual inspection	−4.15	−5.38

Note. Critical Value = 0.6 (values above critical value are considered stable).

There were very few comparative studies that provided information about the accuracy of bridge NDE methods. The studies that did provide information tended to agree with the results. Gucunski et al. [18] gave relative accuracy ratings for various concrete methods. It was shown that impact echo and ultrasonic and electrical potential tended to have more accurate measurements, with ground penetrating radar, infrared, and chain drag being slightly less accurate. The

Table 5: Concrete and steel NDE method normalized averages for accuracy determined from the participant responses.

	Normalized average		
	False positive	False negative	True response
Concrete methods			
Electrical potential	8.40%	11.60%	80.00%
Mechanical sounding	10.41%	12.22%	77.37%
Thermal	6.67%	33.33%	60.00%
Visual inspection	9.76%	11.95%	78.29%
Steel methods			
Acoustic emission	—	—	100.00%
Liquid penetrant	8.40%	8.16%	83.44%
Magnetic particle-crack	10.92%	12.26%	76.81%
Magnetic particle-weld	10.35%	10.51%	79.14%
Radiography	6.43%	9.07%	84.50%
Ultrasonic-crack	8.00%	8.00%	84.00%
Ultrasonic-pin	8.92%	8.92%	82.16%
Ultrasonic-weld	6.89%	8.23%	84.88%
Visual inspection	14.54%	14.83%	70.63%

relative scales of these ratings tend to agree with the responses from the participants.

5.4. Results for the Reliability of NDE Methods. After round three, all responses in the reliability subsection both had converged and became stable and were thus removed from subsequent rounds. The response average and standard deviation from the third and final round are presented in Table 6. Table 6 also presents the COV indicated by the participants' responses for each NDE method. These COVs were determined by fitting a trend line to the response and the COV representation (i.e., a response of 4 meant the test had a COV of 0.08).

Based on the results, most methods had an average response between 3 and 4. Two methods (rebound hammer and thermal) were significantly outside this range. Respondents felt these methods were less reliable than most other methods. The thermal method can be very dependent on both sun exposure and depth of flaw. Yehia et al. [19] showed that both of these factors could produce weak readings causing a decrease in surface area detected or no detection. Furthermore, Rens et al. [20] showed that the rebound hammer method did a "poor" job at detecting deterioration while Wood and Rens [21] showed that the method can be highly variable.

Phares et al. [16] showed that during routine visual inspections the inspectors provided values that were statistically different. Based on the inspector's average standard deviations and average reference rating, the COVs for the superstructure, substructure, and deck responses were 0.14, 0.12, and 0.16, respectively. Furthermore, Barnes and Trottier [17] showed that radar, chain drag (mechanical sounding), and electrical potential have a COV of 0.258, 0.183, and 0.536, respectively. While these values are not an exact match relative to the survey results (participant responses indicate a lower COV for each method), the relative reliability of the different methods based on these studies and the respondents do agree.

Gucunski et al. [18] conducted a study on various NDE methods used on concrete bridge decks and showed that all NDE methods that were tested and were conducive to data analysis had an average COV of less than 0.25. These results are again slightly higher than the participant's responses. Furthermore, Gucunski et al. [18] produced a relative repeatability grade based on graphical representation of the results for each method. Based on this, it was shown that impact echo, ultrasonic, radar, electrical potential, and mechanical sounding all had similar reliability and were relatively more repeatable when compared to thermal. While the COVs could not be compared, the relative scales of each method tend to agree with the results from the survey.

Using only the limited data available for comparison, it can be shown that the survey results for COV were slightly lower than what the previous experimental studies indicate. However, the relative scales tend to agree.

5.5. Results for the Costs Subsection of the Survey. Since the cost questions were changed from round two to round three, the convergence and stability values could not be computed until after round four. After round four, it was determined that 93% of the questions had a COV less than 0.5. The remaining questions were on the lower portion of the less than satisfactory range (between 0.5 and 0.8 COV). All of the COVs that were in this range had also dropped significantly from round three. Based on this, it was determined that all responses had reached a consensus. Furthermore, it was determined that all responses had become stable. Based on these factors, the responses from questionnaire four were considered to be the final values.

These final values were used to correlate the responses to the response range. The mean value of these responses could not be used to determine the exact cost using a trend line and interpolation (similar to the bias and reliability subsections) because the respondents chose a range (as opposed to a single value) given the use of categorical variables as options in the survey (see Figure 3). Based on this, the median response was used to correlate to these cost scales. In other words, the response range represents the range of costs based on the participant's median response and the cost scales provided during the third and fourth rounds of the survey. The response ranges as indicated by the participants are presented in Table 7.

Gucunski et al. [18] provided a comparison of speeds for each NDE method. For the speed category, the Time to Run a Test and the Time to Analyze Data parameters from the survey were used to compare with the speed results from the Gucunski et al. [18] study. It was shown by Gucunski et al. [18] that radar, electrical potential, thermal, and mechanical sounding all tended to be relatively quick with impact echo and ultrasonic being slower. Based on the comparison of the relative scales of the study and the survey, it can be seen that the data tends to agree with only two inconsistencies. One inconsistency is in the case of thermal. Gucunski et al. [18] determined that this method was relatively quick to use while the respondents indicated it took a relatively long time to

Table 6: Concrete and steel NDE methods response descriptive statistics for the reliability and the COVs indicated by participant responses.

	Average	Standard deviation	COVs indicated by participant responses
Concrete NDE methods			
Cover meters/pachometer	4.13	0.83	0.078
Electrical potential	3.50	0.84	0.093
Impact echo	3.40	0.55	0.097
Mechanical sounding	4.11	0.78	0.078
Radar	3.71	1.11	0.086
Rebound hammer	2.67	0.82	0.147
Thermal	2.60	0.89	0.153
Ultrasonic testing	4.14	0.90	0.078
Visual inspection	4.00	0.87	0.080
Steel NDE methods			
Acoustic emission	4.00	—	0.080
Liquid penetrant testing	3.67	1.21	0.088
Magnetic particle testing-crack detection	4.00	0.71	0.080
Magnetic particle testing-weld inspection	3.75	0.50	0.085
Radiography	4.25	0.50	0.076
Ultrasonic testing-crack detection	3.80	0.45	0.084
Ultrasonic testing-pin inspection	3.83	0.75	0.083
Ultrasonic testing-weld Inspection	4.00	0.82	0.080
Visual inspection	3.67	1.03	0.088

Table 7: Concrete and Steel NDE method cost ranges as indicated by participant responses.

	Time to run a test (hours)	Time to analyze data (hours)	Time to train an inspector (days)	Monetary cost for equipment ($)	Number of inspectors needed (#)
Concrete NDE methods					
Cover meters/pachometer	8–10 hours	4–8 hours	2–14 days	$1500–$3000	2 inspectors
Electrical potential	10–15 hours	4–8 hours	7–14 days	$1500–$3000	2-3 inspectors
Impact echo	10–15 hours	8–12 hours	21+ days	$6000+	3 inspectors
Mechanical sounding	4–8 hours	2–4 hours	2–7 days	0–$500	2 inspectors
Radar	8–10 hours	8–12 hours	21+ days	$6000+	3 inspectors
Rebound hammer	4–8 hours	2–4 hours	7–14 days	$500–$3000	2 inspectors
Thermal	10–15 hours	8–12 hours	14–21 days	$6000+	3 inspectors
Ultrasonic testing	8–10 hours	4–8 hours	21+ days	$6000+	2 inspectors
Visual inspection	4–8 hours	2–4 hours	7–14 days	0–$500	2 inspectors
Steel NDE methods					
Acoustic emission	8–10 hours	4–12 hours	14–21+ days	$6000+	2 inspectors
Liquid penetrant testing	8–10 hours	2–4 hours	2–7 days	0–$1500	2 inspectors
Magnetic particle testing-crack detection	8–10 hours	2–4 hours	7–14 days	$500–$1500	2 inspectors
Magnetic particle testing-weld inspection	8–10 hours	2–4 hours	7–14 days	$500–$1500	2 inspectors
Radiography	10–15 hours	4–8 hours	21+ days	$6000+	3 inspectors
Ultrasonic testing-crack detection	8–10 hours	2–4 hours	21+ days	$6000+	2 inspectors
Ultrasonic testing-pin inspection	8–10 hours	2–8 hours	21+ days	$3000–$6000+	2 inspectors
Ultrasonic testing-weld inspection	8–10 Hours	2–4 hours	21+ days	$6000+	2 inspectors
Visual inspection	4–8 Hours	2–4 hours	2–7 days	0–$500	1-2 inspectors

Table 8: Comparison of NDE methods using rankings indicated by participant responses.

	Bias	Accuracy	Reliability	Average cost
Concrete NDE methods				
Cover meters/pachometer	1	—	1	4
Electrical potential	—	1	6	5
Impact echo	2	—	7	9
Mechanical sounding	—	3	1	1
Radar	5	—	5	7
Rebound hammer	—	—	8	3
Thermal	—	4	9	7
Ultrasonic testing	4	—	1	6
Visual inspection	3	2	4	2
Steel NDE methods				
Acoustic emission	1	1	2	8
Liquid penetrant testing	—	5	8	2
Magnetic particle testing-crack	—	8	2	3
Magnetic particle testing-weld	—	7	7	3
Radiography	5	3	1	9
Ultrasonic testing-crack detection	3	4	6	5
Ultrasonic testing-pin inspection	2	6	5	5
Ultrasonic testing-weld inspection	4	2	2	5
Visual inspection	6	9	8	1

perform. Furthermore, ultrasonic testing was shown by the study to be very time intensive, but the respondents indicated that the test was near the midpoint in terms of time used relative to the other methods.

5.6. Comparison of NDE Methods. After all the data was collected, a preliminary comparison of the NDE methods was made. This comparison can be seen in Table 8. Each method was given a relative rank in each of the four categories measured: bias, accuracy, reliability, and cost. Note that there are a total of nine tests for both concrete and steel methods. These ranks were based on the most desirable outcome for each category. For bias, the lower the rank number, the less biased the test. Similarly, the lower the rank number for the cost category, the cheaper the average cost of the method. Since the cost subsection measured five different categories of costs, an average in terms of ranking for all the costs for each method was used for this comparison. In other words, a ranking was done for each category and an average was taken across all five categories. For reliability and accuracy, the lower the rank number, the more reliable or more accurate the test was, respectively.

By comparing the rankings in each of the four categories that were examined for each NDE method, it is possible to understand the relative differences between tests. Furthermore, an approximate correlation of the costs of a method to the bias, accuracy, and reliability can be made. In general, it was shown by the participants that the more expensive the method was, the better bias, accuracy, and reliability the method had, and vice versa. There were a few exceptions to this rule, however. Both thermal and radar tended to be relatively expensive. Thermal tended to be relatively inaccurate and very unreliable, and radar was relatively biased and fairly unreliable. By evaluating these comparisons it can be seen that inspection planning choices should consider the quality of information a test provides as well as the costs (in terms of time and money). It is important to note that while Table 8 can be used as an initial guide to select preferred NDE inspection method based on different criteria or a combination thereof, one needs to recognize the fact that when two NDE methods measure different things on a bridge, they are complementary, not necessarily interchangeable.

6. Summary and Conclusions

Given the limited amount of research done to quantify the uncertainty in common NDE methods or to compare the costs of various tests to one another (as far as their use on bridges is concerned), a Delphi method survey to gather expert opinion was identified as a means to gather the desired information. This survey aimed to provide quantitative descriptions of bias, accuracy, reliability, and costs to provide information to researchers and practitioners working in the fields of bridge management and inspection.

A total of four Delphi method rounds were conducted in order to determine quantitatively the bias, accuracy, reliability, and various costs of common NDE methods. The first survey was employed to determine background information of the participants and common NDE methods for bridges. The second and subsequent surveys were used to quantify the uncertainty.

The results of these surveys were used to develop quantitative information for each method. Based on these results the following conclusions can be drawn:

(1) Most commonly used bridge NDE methods tend to be underbiased, meaning the majority of the measured results are slightly less than the true value. All of these biases were shown to be less than 10%, however. These values tended to agree with the data from previous experimental studies.

(2) The accuracy of commonly used bridge NDE methods tends to be relatively variable. For concrete testing, most tests had a true response percentage of about 80%. The exception to this was thermal with a true response percentage of 60%. Furthermore, most steel tests had a true response percentage of about 85%.

(3) According to the respondents, most commonly used bridge NDE methods tend to be relatively reliable. However, comparing survey results to physical testing indicates that while inspection personnel seem to have a relative understanding of the variability in different tests, they tend not to have an understanding of the absolute scale of the variability. In an experimental setting Barnes and Trottier [17] showed that radar, chain drag (mechanical sounding), and electrical potential have a COV of 0.258, 0.183, and 0.536, respectively, while the participants responses indicated they thought the COV for radar, mechanical sounding, and electrical potential were 0.086, 0.078, and 0.093, respectively. This indicates that more experimental data of this nature is needed to educate bridge inspectors and managers.

(4) The various costs associated with the NDE methods examined tended to be very variable making this measure difficult to evaluate. However, there was a small trend that indicated that tests that were cheaper in terms of equipment also tended to be easier and faster to perform.

(5) By comparing the rankings of each of the four categories that were examined for each NDE method, it is possible to correlate the cost of a method to the bias, accuracy, and reliability. In general, it was shown by the participant responses that the more expensive the method was, the better bias, accuracy, and reliability the method had and vice versa.

Based on these findings, advancements to bridge inspection planning need to carefully consider the level of information needed about bridge condition and the costs of obtaining that information. Various approaches might be adopted to use the uncertainty in inspection findings and costs to improve inspection practice. A decision-making tool assigning various weights to the different ranking criteria indicated in Table 8 could be used to select preferred NDE inspection methods, while recognizing the fact that when two NDE methods measure different things on a bridge, they are complementary, not necessarily interchangeable. Additionally, the findings of this study could be incorporated into a larger risk-based framework for bridge inspection and management planning.

It is important to note that using a survey to determine statistical parameters is complicated. In order to help the respondents, the scales for bias and reliability included the graphical representations shown in Figures 1 and 2. Alternative means of collecting this information might have asked respondents to consider variability in specific scenarios, but this method might have been more difficult for respondents, depending on whether they had prior experience with the scenario. Future research can investigate other means of collecting information for bias and reliability.

It is also important to note that while the relative low response rate in this study (even though not too far from the recommended Delphi survey size as suggested by Ludwig [15]) is a potential limitation of this study, the information presented in this paper can serve as a starting point for characterizing different NDE methods for use in bridge management and inspection planning and identifies the type of information that is still needed.

Disclosure

The opinions and findings are those of the authors and do not necessarily represent the views of MPC.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The research work described in this paper has been partially funded by the Mountain-Plains Consortium (MPC), a competitively selected university program sponsored by the U.S. Department of Transportation through its Research and Innovative Technology Administration. The authors would like to thank all the survey respondents.

References

[1] USGPO, *Code of Federal Regulations (annual edition) 23 CFR 650 - Bridges, Structures, and Hydraulics*, 2009.

[2] "White Paper on Bridge Inspection and Rating," *Journal of Bridge Engineering*, vol. 14, no. 1, pp. 1–5, 2009, ASCE/SEI-AASHTO Ad-Hoc Group On Bridge Inspection, Rating, Rehabilitation, and Replacement.

[3] A. Hesse, *Using Expert Opinion to Quantify Accuracy and Reliability of Nondestructive Evaluation on Bridges*, Colorado State University, Fort Collins, Colo, USA, 2013.

[4] A. A. Hesse, R. A. Atadero, and M. E. Ozbek, "Uncertainty in common NDE techniques for use in risk-based bridge inspection planning: Existing data," *Journal of Bridge Engineering*, vol. 20, no. 11, Article ID 04015004, pp. 1–8, 2015.

[5] A. Hatami and G. Morcous, "Developing Deterioration Models for Nebraska Bridges," 2011, Nebraska Department of Roads Final Report.

[6] M. I. Yousuf, "Using experts' opinions through Delphi technique," *Practical Assessment, Research and Evaluation*, vol. 12, no. 4, 2007.

[7] H. A. Linstone and M. Turoff, *The Delphi Method: Techniques and Applications*, Addison-Wesley Publishing Co., Advanced Book Program, Boston, Mass, USA, 2002.

[8] C. Okoli and S. D. Pawlowski, "The Delphi method as a research tool: an example, design considerations and applications," *Information and Management*, vol. 42, no. 1, pp. 15–29, 2004.

[9] C.-C. Hsu and B. A. Sandford, "The Delphi technique: making sense of consensus," *Practical Assessment, Research and Evaluation*, vol. 12, no. 10, pp. 1–8, 2007.

[10] J. M. English and G. L. Kernan, "The prediction of air travel and aircraft technology to the year 2000 using the Delphi method," *Transportation Research*, vol. 10, no. 1, pp. 1–8, 1976.

[11] S. A. Kalaian and R. M. Kasim, "Termination sequential Delphi survey data collection," *Practical Assessment, Research and Evaluation*, vol. 17, no. 5, pp. 1–10, 2012.

[12] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Fla, USA, 3rd edition, 2004.

[13] K. L. Rens, T. J. Wipf, and F. W. Klaiber, "Review of nondestructive evaluation techniques of civil infrastructure," *Journal of Performance of Constructed Facilities*, vol. 11, no. 4, pp. 152–160, 1997.

[14] M. Moore, D. Rolander, B. Graybeal, B. Phares, and G. Washer, "Highway Bridge Inspection: State-of-the-Practice Study," 2001, Prepared by the United States Department of Transportation Federal Highway Administration. Final Report.

[15] B. Ludwig, "Predicting the future: Have you considered using the Delphi methodology?" *Journal of Extension*, vol. 35, no. 5, pp. 93–96, 1997.

[16] B. M. Phares, D. D. Rolander, B. A. Graybeal, and G. A. Washer, "Reliability of visual bridge inspection," *Public Roads (Publication of the Federal Highway Administration)*, vol. 64, no. 5, 2001.

[17] C. L. Barnes and J.-F. Trottier, "Ground-penetrating radar for network-level concrete deck repair management," *Journal of Transportation Engineering*, vol. 126, no. 3, pp. 257–262, 2000.

[18] N. Gucunski, A. Imani, F. Romero et al., *Nondestructive Testing to Identify Concrete Bridge Deck Deterioration*, Transportation Research Board, Washington, DC, USA, 2012.

[19] S. Yehia, O. Abudayyeh, S. Nabulsi, and I. Abdelqader, "Detection of common defects in concrete bridge decks using nondestructive evaluation techniques," *Journal of Bridge Engineering*, vol. 12, no. 2, pp. 215–225, 2007.

[20] K. L. Rens, C. L. Nogueira, and D. J. Transue, "Bridge management and nondestructive evaluation," *Journal of Performance of Constructed Facilities*, vol. 19, no. 1, pp. 3–16, 2005.

[21] J. C. Wood and K. L. Rens, "Nondestructive testing of the Lawrence street bridge," *Structures*, pp. 1–15, 2006.