9-9-2010

A framework for assessing and improving quality of data from visual evaluation of asset conditions

Arturo Adrian Cordova

Arturo Cordova Alvidrez
Candidate

Civil Engineering
Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

_______________________________________, Chairperson

____

____

A FRAMEWORK FOR ASSESSING AND IMPROVING QUALITY OF DATA FROM VISUAL EVALUATION OF ASSET CONDITIONS

BY

ARTURO A. CORDOVA ALVIDREZ

B.S. UNIVERSIDAD AUTONOMA DE CHIHUAHUA
CHIHUAHUA, MEXICO
2007

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Civil Engineering

The University of New Mexico
Albuquerque, New Mexico

August 2010

DEDICATION

Quiero dedicarle esta que es, hasta la fecha, mi más grande obra y logro, a mi familia, Córdova Alvídrez. A lo largo de mi vida he contado con la fortuna de contar con el apoyo de la gente adecuada, en el momento oportuno. Sin embargo, han sido mis padres y mis dos hermanos quienes han estado presentes en todos y cada uno de los acontecimientos que han definido mi vida.

A mis padres, Ing. Arturo Córdova González y Sra. Margarita Alvídrez de Córdova, los autores intelectuales y materiales de mi existencia. Ellos construyeron los cimientos y definieron la estructura de mi persona, asegurando que tenga la oportunidad de seguir el camino que el Señor y la vida tienen preparado para mí. Espero que la presente obra sea fiel reflejo de los méritos que ellos han construido en mí y en mis hermanos.

A mis hermanos, Flor Margarita Córdova Alvídrez y Daniel David Córdova Alvídrez, que son las personas con quien más he convivido a lo largo de mi vida. Su particular forma de ver la vida dio colorido a aquellos momentos en mi andar que lo carecían. Espero que encuentren en esta obra un modelo a seguir en la propia.

A todas aquellas personas que siempre creyeron en mí a pesar de mis falencias, y que me es imposible mencionar en su totalidad por este conducto. A todos ellos: MUCHAS GRACIAS.

Ing. Arturo Adrián Córdova Alvídrez

ACKNOWLEDGMENTS

I would like to acknowledge my academic advisors, Dr. Susan Bogus-Halter and Dr. Giovanni Migliaccio, for their time and patience in the development of my degree. I deeply appreciate their efforts and the lessons learned with them in the classroom and in our weekly meetings. I am thankful with them for opening me the doors to grad school in the United States. I will particularly take with me the academic and the professional advice given by them throughout this time.

I also want to acknowledge my graduation committee, Dr. Susan Bogus-Halter, Dr. Giovanni Migliaccio, and Dr. James Brogan, for their advisement and support during the long thesis process. Their insight and input were of significance in the completion of this thesis. I also want to thank Mr. Robert Young and Mr. Tito Medina from the NMDOT for their vast support during the data collection that took place in the summers of 2008 and 2009.

Finally, I want to acknowledge the Department of Civil Engineering of the University of New Mexico for giving me the opportunity to prove my capacity in a competitive country, like the United States.

A FRAMEWORK FOR ASSESSING AND IMPROVING QUALITY OF DATA FROM VISUAL EVALUATION OF ASSET CONDITIONS

BY

ARTURO A. CORDOVA ALVIDREZ

ABSTRACT OF THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Civil Engineering

The University of New Mexico
Albuquerque, New Mexico

August 2010

A FRAMEWORK FOR ASSESSING AND IMPROVING QUALITY OF DATA FROM VISUAL EVALUATION OF ASSET CONDITIONS

By

Arturo A. Cordova Alvídrez

B.S., Civil Engineering, Universidad Autónoma de Chihuahua, 2007

M.S., Civil Engineering, University of New Mexico, 2010

ABSTRACT

Investments in transportation infrastructure assets are among the largest investments made by governmental agencies. Besides requiring a large investment for design and construction, transportation infrastructure also requires a significant amount of resources and effort for performing maintenance and/or rehabilitation activities. Along with other considerations, data on asset conditions are used to make decisions regarding the timing of maintenance activities, the type of treatment, and the resources employed. Some parameters under assessment, however, are evaluated through visual – or manual – assessments performed by evaluators on the site due to a lack of reliable, inexpensive automated methods to collect the data. While manual assessments for surface distresses are widely used, they still have the stigma that the results are based on subjective judgments by the individual evaluators. This thesis describes the Data Quality Assessment & Improvement Framework that has been developed to measure, and to improve, the performance of multiple pavement evaluators. This framework is based on a Continuous Quality Improvement cyclic process, where the main components include: a) assessment of the consistency over time – performed using linear regression analysis, b) assessment of the agreement between evaluators – performed using inter-rater agreement analysis, and c) management practices performed to improve the results shown by the assessments. When the Data Quality Assessment & Improvement Framework is applied to actual pavement distress data collected manually by different evaluators, the results show that it is an effective method for quickly identifying and solving data collection issues. The benefit of this framework is that the analyses employed provide performance data during the data collection process, thus minimizing the risk of subjectivity. The Data Quality Assessment & Improvement Framework can be used as part of an asset management program, or in any engineering program where the data collected are subjected to the judgment of the individuals performing the evaluation.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION **1**

1.1. Overview **1**

1.2. Study Objectives **6**

1.3. Research Methodology **10**

1.4. Reader's Guide to Thesis **11**

CHAPTER 2. BACKGROUND & LITERATURE REVIEW **12**

2. 1. Asset Management Systems **12**

2.2. Pavement Management Systems **14**
2.2.1. Broader Management Concerns. ***14***
2.2.2. Network Level. ***15***
2.2.3. Project Level. ***15***
2.2.3.1. Design. *16*
2.2.3.2. Construction. *16*
2.2.3.3. Maintenance and Rehabilitation (M&R). *16*
2.2.4. Research and Special Studies. ***17***
2.2.5. Data Base. ***17***

2.3. Pavement Condition Assessment **18**
2.3.1. Functions of Pavement Evaluation ***18***
2.3.2. Pavement Outputs ***19***
2.3.2.1. Surface Distress *19*
2.3.2.2. Structural Capacity *19*
2.3.2.4. Surface Friction or Skid Resistance *20*

2.4. Pavement Surface Distress Evaluation **21**
2.4.1. Overview ***21***
2.4.2. State of the Practice ***22***
2.4.2.1. Evaluation Procedures *22*
2.4.2.2. Sampling Procedures *33*
2.4.2.3. Rating Procedures *33*
2.4.2.4. Evaluation Methods *34*
2.4.3. Quality Management of Pavement Distress Surveys ***37***

2.5. Inter-rater Agreement –Adapted from Bogus, Migliaccio, and Cordova (2010a, 2010b) 42
2.5.1. Overview 42
2.5.2. IRA Indexes 43
2.5.2.1. James et al (r_{WG}) 43
2.5.2.2. Schmidt & Hunter (SD) 46
2.5.2.3. Lindell et al (r^*_{WG}) 47
2.5.2.4. Burke et al (AD) 48

2.6. Summary 50

CHAPTER 3. RESEARCH METHODOLOGY 51

3.1. Research Objectives 51

3.2. Research Design 55

3.3. Data Collection 58

3.4. Data Analysis 66
3.4.1. Inter-Rater Agreement 66
3.4.2. Linear Regression Analysis 68

CHAPTER 4. DATA QUALITY ASSESSMENT & IMPROVEMENT FRAMEWORK (DQAIF) 69

4.1. Overview 69

4.2. Conceptual Structure 69

4.3. DQAIF Process Flow 72
4.3.1. Data Collection 72
4.3.2. Agreement Between Evaluators (ABE) Assessment 74
4.3.3. Consistency Over Time (COT) Assessment 86
4.3.4. Improvement Assessment 93
4.3.5. Control Measures 97

4.4. Summary 97

CHAPTER 5. CASE STUDY: NORTHERN NEW MEXICO PAVEMENT EVALUATION PROGRAM 99

5.1. Overview 99

5.2. Data Analysis & Results 100
5.2.1. Assessments from Previous Years 100
5.2.1.1. 2007 Northern New Mexico Pavement Evaluation Program Analysis 101
5.2.1.2. 2008 Northern New Mexico Pavement Evaluation Program Analysis 104
5.2.2. 2009 Northern New Mexico Pavement Evaluation Program Results and Analysis 107
5.2.2.1. First Assessment Round 107
5.2.2.2. Second Assessment Round 110

5.3. Discussion & Recommendations 123

CHAPTER 6. CONCLUSIONS 125

6.1. Summary of Study 125

6.2. Research Questions Rationale & Findings 126

6.3. Opportunities for Future Research 128

REFERENCES 129

APPENDIX A: ESTIMATION PROCESSES FOR INTER-RATER AGREEMENT ALTERNATIVE MEASURES 135

LIST OF FIGURES

Figure 1. Types And Examples Of Infrastructure (Adapted From Moteff And Parfomak, 2004; Clough Et Al, 2004). 3
Figure 2. Generic Asset Management System Components (From Usdot, 1999) 13
Figure 3. Major Components Of A Pavement Management System (From Haas Et Al, 1994). 14
Figure 4. Longitudinal Cracks On Flexible Asphalt Pavement (From Miller & Bellinger, 2003). 24
Figure 5. Sealed Transverse Crack On Flexible Pavement (From Miller & Bellinger, 2003). 25
Figure 6. Block Cracks On Flexible Asphalt Pavement (From Miller & Bellinger, 2003). 25
Figure 7. Reflection Crack Overview On Flexible Asphalt Pavement (A), And Reflection Crack Closeup (B) (From Miller & Bellinger, 2003). 26
Figure 8. Pothole On Flexible Asphalt Pavement (From Miller & Bellinger, 2003). 27
Figure 9. Measure Of A Rut Depth On Flexible Asphalt Pavement (From Miller & Bellinger, 2003). 27
Figure 10. Bleeding On Flexible Asphalt Pavement (From Miller & Bellinger, 2003). .. 28
Figure 11. Raveling On Flexible Asphalt Pavement (From Miller & Bellinger, 2003). .. 29
Figure 12. Lane-Shoulder Drop-Off (From Miller & Bellinger, 2003). 29
Figure 13. Longitudinal Cracks On Rigid Concrete Pavement (From Miller & Bellinger, 2003). 30
Figure 14. Transverse Crack On Rigid Concrete Pavement (From Miller & Bellinger, 2003). 30
Figure 15. Durability Cracking On Rigid Concrete Pavement (From Miller & Bellinger, 2003). 31
Figure 16. Faulted Transverse Joints On Rigid Concrete Pavement (From Miller & Bellinger, 2003). 31
Figure 17. Severe Blowups On A Rigid Concrete Road (From Miller & Bellinger, 2003). 32

Figure 18. Durability Cracks On Continuously Reinforced Pavement (From Miller & Bellinger, 2003). 32
Figure 19. Pavement Distress Data Collection Methods (Adapted From Haas Et Al, 1994; Gramling, 1994; Mcghee, 2004). 35
Figure 20. Quality Management System Components (Adapted From Morian Et Al, 2002; Mcpherson And Bennett, 2005; Iso 9000:2000, 2000). 39
Figure 21. Constructs And Variables Of The Research Hypotheses. 53
Figure 22. Research Process 56
Figure 23. Field Operations In The Nmdot Pavement Evaluation Program (Unm, 2009). 59
Figure 24. Nmdot Severity And Extent Descriptions For Rutting And Shoving (Nmdot, 2004). 61
Figure 25. Tqm Circle Of The Northern New Mexico Pavement Evaluation Program. .. 64
Figure 26. Quality Control Levels Of The Northern New Mexico Pavement Evaluation Program. 64
Figure 27. Dqaif Conceptual Structure (From Bogus, Migliaccio, And Cordova, 2010a). 70
Figure 28. Spreadsheet Showing The Overall Process To Compute Burke Et Al (1999) Single- And Multiple-Item $Ad_{md.}$ 71
Figure 29. Overall Sequence Of The Dqaif Process. 73
Figure 30. Abe Assessment Process Framework. 75
Figure 31. Ira Analysis Spreadsheet Format. 78
Figure 32. Spreadsheet Format Of X_{kj} 80
Figure 33. Location Of The Evaluator And Alternatives Counts Within The Ira Spreadsheet. 80
Figure 34. Estimate Of The Median In The Ira Spreadsheet. 81
Figure 35. The Deviation Around The Median Matrix (Dm_{md}). 82
Figure 36. Estimation Of The Single-Item Ad_{md} Indexes In The Ira Spreadsheet. 82
Figure 37. Estimate Of The Multi-Item Ad_{md} In The Ira Spreadsheet. 83
Figure 38. Ratings Frequencies Counts Within The Ira Spreadsheet. 85
Figure 39. Ratings Frequencies Histogram Based On A 4-Point Scale Rating Protocol. 86

Figure 40. Cot Assessment Process Framework. 88
Figure 41. Lra Spreadsheet. 89
Figure 42. Rating Count Histogram Of Two Assessment Times For One Distress. 93
Figure 43. Radar Graph Used During The Improvement Assessment Stage. 94
Figure 44. Improvement Assessment Process Framework. 96
Figure 45. Ad_{md} Results Of The 2007 Season. 102
Figure 46. Ad_{md} Results Of The 2008 Season. 105
Figure 47. Ira Analysis Results For The First Assessment Round. 108
Figure 48. Ratings Frequencies Histograms Of Bleeding Severity In Four Items. 109
Figure 49. Ira Analysis Results For The Second Assessment Round. 111
Figure 50. Ratings Frequencies Histograms Of The Five Different Distresses. 112
Figure 51. Scatter Plots Comparing The Ratings At Different Times Of Assessment. .. 114
Figure 52. Example Of A Rating Count Histogram. 116
Figure 53. Ratings Count Histograms For Bleeding Severity And Extent. 117
Figure 54. Ratings Count Histogram For Edge Cracks Extent. 118
Figure 55. Radar Graph Of The $Ad_{md(J)}$ Results From The First And Second Assessment Rounds. 119

LIST OF TABLES

Table 1. Nmdot Distress Descriptions. 60
Table 2. Weighting Factors For Flexible Pavement Distresses (Nmdot, 2004). 62
Table 3. Roads Used For Case Study Data Collection (From Unm, 2009). 65
Table 4. Summary Of Ira Indexes. 67
Table 5. 2007 Cot Season Results 103
Table 6. 2008 Cot Season Results 106

LIST OF ACRONYMS

AADT. Average Annual Daily Traffic.

AASHTO. American Association of State Highway and Transportation Officials.

ABE. Agreement Between Evaluators.

ACP. Asphalt Concrete Pavement.

AD. Average Deviation.

COT. Consistency Over Time.

CQI. Continuous Quality Improvement.

CRCP. Continuously Reinforced Concrete Pavement.

DM. Deviation Matrix.

DQAIF. Data Quality Assessment & Improvement Framework.

DR. Distress Rate.

FHWA. Federal Highway Administration.

GIS. Geographic Information System.

HPMS. Highway Pavement Monitoring System.

IRA. Inter-rater Agreement.

IRI. International Roughness Index.

IRR. Inter-rater Reliability.

JPCP. Jointed Plain Concrete Pavement.

JRCP. Jointed Reinforced Concrete Pavement.

LRA. Linear Regression Analysis.

LTPM. Long Term Pavement Monitoring.

M&R. Maintenance & Rehabilitation.

NCHRP. National Cooperative Highway Research Program.

NMDOT. New Mexico Department of Transportation.

PCI. Pavement Condition Index.

PMS. Pavement Management System.

PSI. Pavement Serviceability Index.

QA/QC. Quality Assurance/Quality Control.

TQM. Total Quality Management.

UNM. University of New Mexico.

USDOT. United States Department of Transportation.

CHAPTER 1. INTRODUCTION

1.1. Overview

In economics, an asset is "anything -tangible or intangible- that is capable of being owned or controlled to produce value" (O'Sullivan & Sheffrin, 2003). In civil engineering, this concept is generally associated to the term *infrastructure*, which is defined as "The basic facilities, services, and installations needed for the functioning of a community or society." (The American Heritage Dictionary of the English Language, 2000). Then, as assets, infrastructure can be conceptualized as the set of tangibles owned by the society that can be managed to contribute to the development of the communities.

The latter has been the role of infrastructure within the society since the times of the ancient civilizations. Roads were built in England, India, and Middle East before Roman times (i.e. the Persian *Royal Road*). With the advent of the Roman Empire, the use of crushed stone and earth materials became common in the construction of roads (Lay & Vance, 1992). Canals and irrigation systems came along with the birth of civilization, during the rise of Mesopotamia, the Indus Valley Civilization, Egypt, and Ancient China; and they started to be built in Europe in the Middle Ages (Hadfield, 1986; Needham et.al., 1971; Rodda and Ubertini, 2004).

Moreover, infrastructure has not only walked along with civilization, it has been one of the drivers of its development. Energy infrastructure provides society with the energy and fuels necessary to run most of their daily activities. Water management infrastructure supplies communities with this liquid, considerably important to support life. Communications infrastructure facilitates the flow of data and information within and among communities. Transportation infrastructure facilitates the moving of goods and

people. It provides the society with the means to meet the demand of products and resources. People become closer to others, in a sense of taking less time and effort to move from one place to another. Additionally, infrastructure not only stimulates the flow of commerce, but it also induces the development of the different industries and economic sectors by supporting their activities between separate locations. Thus, infrastructure has been an important means for the development of human communities.

There are different types of infrastructure, the most important being the ones listed in Figure 1 (next page). However, all of them play a major role in the development of our communities. Moreover, taking the measures and efforts to effectively and timely deliver infrastructure is as important as the functions it has within the society, and just as challenging. How efficiently a community operates relies, considerably, on the capacity of its infrastructure. Moreover, performing the design, construction, and maintenance of these facilities involves the participation of several different groups of people to conduct a complex myriad of tasks, and the appropriate use of large amounts of resources – money, machinery, manpower, and time. Therefore, it is critical to deploy an organized and well-established system to manage these societies' assets.

A concept that emerged to address this issue is asset management, which is a systematic process of effectively administering the entire life-cycle of physical assets, by combining engineering principles, sound business practices, and economic theory (FHWA, 1997).

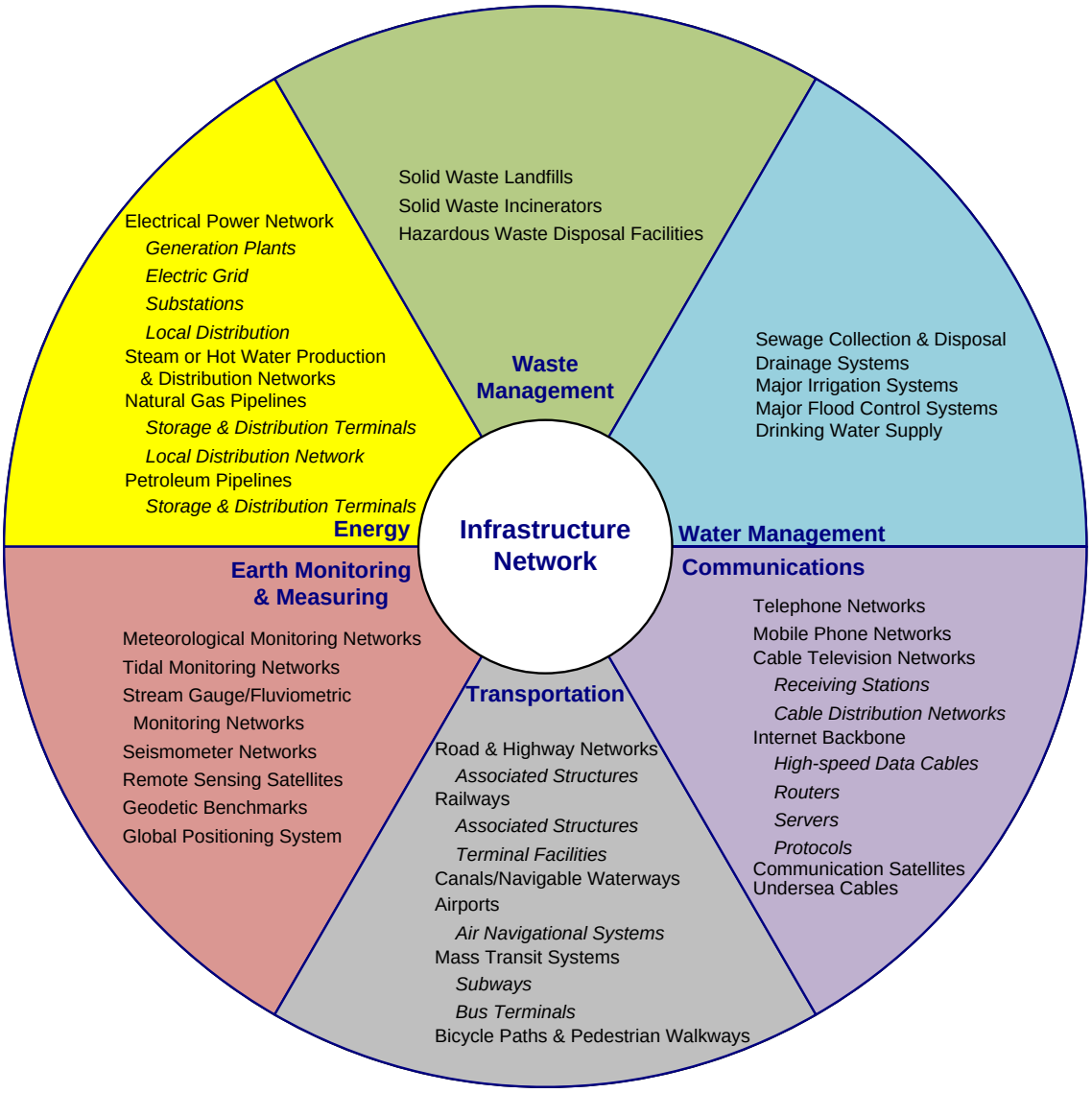


Figure 1. Types and examples of infrastructure (Adapted from Moteff and Parfomak, 2004; Clough et al, 2004).

Nevertheless, different management systems have been developed to address the particular needs of each type of asset, but these still include the earlier functions. For instance, within the transportation sector, particular interest has been focused on pavements, which play a major role in any transportation system, as was asserted by Haas et al (1994):

> [From today's transport systems,] only marine and pipeline transportation do not make use of pavements. Certainly, the major structural load-carrying elements of the highway system are the pavements. For air travel, pavements are required in the form of runways, taxiways, and parking aprons. Likewise, the railroads operate in a form of pavement historically made of rails, ties, and ballast, not dissimilar to a highway pavement design. In fact, modern design principles show that rails can easily be mounted on a properly designed continuous pavement (pg. 7).

The construction and maintenance of pavement systems entail considerable amounts of resources. Several crews of laborers and heavy construction equipment place and compact large amounts of earth materials, which constitute the pavement system –all this with the expenditure of large amounts of money. This is repeated along the lifespan of the pavement which, in most cases, extends for at least 10 years. This increases the importance of taking care of, or "managing", pavement.

With a similar approach to asset management, researchers in the 1960s and 1970s coined the term Pavement Management System (PMS) and its downstream concepts (Hudson et.

al., 1968; Hutchinson & Haas, 1968; Wilkins, 1968; Scrivner, 1968). A PMS is the framework of methodologies and processes applied for the activities of planning, designing, constructing, and maintaining pavements (Haas & Hutchinson, 1970). The main objective from this approach is to deliver and maintain pavements that meet the end users' expectations.

A critical feature in a management system is the *assessment –or evaluation- of current conditions*. In the case of pavements, this assessment involves the measurement and analysis of four main groups of outputs: a) serviceability, b) structural adequacy, c) surface distress, and d) safety.

Assessing current pavement conditions is a major task that has to be performed in any PMS. This assessment provides information on the current condition of the pavement, and by analyzing the data, the pavement management agency can determine a) whether the pavement is still in adequate condition to operate; b) whether the pavement provides the service it was meant to; and c) whether maintenance actions have to be implemented. Another major application of the pavement condition assessment is that, if performed continuously over the lifespan of the asset, the data can be used to model the pavement's overall performance; thus, forthcoming conditions can be predicted to identify future needs and a management plan for the asset can be developed ahead of time (Haas et al, 1994; Shahin, 2005).

1.2. Study Objectives

In the case of pavement condition assessments, there are two major methods for data collection: 1) manual condition assessment, where the severity and extent of pavement distresses are visually assessed on site by a pavement evaluator; and 2) automated techniques, which mainly consist of either the use of automated tools and devices to measure the distresses of the pavement onsite, or of image scanning onsite and data analysis offsite. Offsite data analysis may be performed either with imaging techniques or by a pavement evaluator (NMDOT, 2009). While it has been noted in the literature that automated pavement condition data collection is safer and faster, it has also been reported that the data gathered by onsite manual assessments (i.e. walking surveys) is more precise (Haas et al, 1994; Shahin, 2005). Additionally, analyses of different data collection practices have shown that manual surveys are more cost-effective than surveys using images and fully automatic methods (automated data collection and analysis procedures) (NMDOT, 2009).

With manual surveys, however, federal and state agencies may be concerned about the consistency of data. In fact, manually collected data may include variability due to the fact that manual collection methods involve multiple evaluators.

What happens in the case of manual evaluations is that, even though the PM agency develops a protocol that has to be followed in order to perform the visual assessments, most times this protocol leaves room to more than a single interpretation, which is based on the judgment an evaluator can have at the moment of performing the evaluation. Nevertheless, sometimes the protocol cannot be more specific to narrow down the possibilities of multiple interpretations, because more specific descriptions may not

consider situations that can be present on site, or just because there is not a specific way to measure the characteristics that are being assessed.

To illustrate this issue, let's consider the severity level descriptions for edge cracking from the Distress Identification Manual for the Long-Term Pavement Performance Project (Miller and Bellinger, 2003):

> Low [severity]
>
> Cracks with no breakup or loss of material.
>
> Moderate [severity]
>
> Cracks with some breakup and loss of material for up to 10 percent of the length of the affected portion of the pavement.
>
> High [severity]
>
> Cracks with considerable breakup and loss of material for more than 10 percent of the length of the affected portion of the pavement. (p. 7)

In this particular case, it is ultimately left to the evaluator to decide what the boundaries between "some" and "considerable" are for breakup and loss of material. Initial training can cover these concerns, but it is not possible to train for each particular case that may fall between the higher limit of what is considered a "moderate" level severity and the lower limit of a "high" level severity. It is here where the final output of the evaluation is left to the judgment (engineering-related or non-engineering-related) of the evaluator which, in various cases, differs among different evaluators –even between persons with

similar profiles and backgrounds. Thus, it can happen that two evaluators may rate differently the same pavement sample.

This variability concern does not only apply between multiple evaluators but also to the same evaluator between different evaluations. This is due to the fact that an evaluator may "change" or, more appropriately, develop his or her judgment with time. As a result, the same evaluator could rate differently the same pavement sample at different times.

The fact that the body of knowledge of the evaluators can differ and change over time is still a concern that most advocates of automated data collection and analysis techniques point at. Pavement and highway agencies are also concerned about the data collected through manual or visual assessments varying considerably to the point of affecting the way these agencies spend public resources, based on arguable evaluation outputs. Variability in manual data collection methods is still an issue that has not yet been resolved (Rada et al, 1997).

However, visual inspections –or visual conditions surveys, as they are called within the field- still cannot be entirely replaced by automated methods. In addition to the cost-effectiveness benefits aforementioned, manual inspections are still necessary to collect performance-related data, ever since the development of this concept (Carey and Irick, 1960). Thus, the improvement of variability of the data collected in manual condition assessments is still a concern that should be addressed in the pavement engineering and management fields, and from which pavement and highway agencies will benefit to better assure the delivery of assets that have the capacity to positively influence the development of the society.

The study presented here aimed to find a solution to issue of variability inherent to manual asset condition assessments, with the development of a Data Quality Assessment & Improvement Framework (DQAIF). This framework measures the variability of data among evaluators, and between evaluations performed at different times, by following a set of procedures as part of a Total Quality Management (TQM) system. The main research question is whether the variability of the data collected through methods influenced by subjectivity and judgment can be reduced by continuous efforts of assessment and training. These efforts measure and maintain the body of knowledge that is being used by the panel of evaluators, which consists on the protocol –assumed to remain the same throughout the entire evaluation project- and the evaluators judgment – component whose change will be monitored and controlled by the DQAIF. A second research question arises from the development of the DQAIF, as to whether variability of manual assessments can be measured in two "dimensions", that being: a) among evaluators, and b) across time; and, thus, controlling variability in these dimensions would help reduce overall variability of manual pavement condition surveys.

This research is focused on the development of the DQAIF to assess and improve the quality of the data collected in manual pavement condition assessments, in terms of reduced variability. The Data Quality Assessment & Improvement Framework can be used as part of an asset management program, or in any engineering program where the data collected are subjected to the judgment of the individuals performing the evaluation..

However, this framework is only intended to be used by the asset management agency, as a management tool within their asset management program. It is developed only to measure and control the quality of the data collected in manual assessments, as part of a

Quality Management Plan, but is not intended as the basis to establish quality control and quality assurance responsibilities in a contract between the pavement management agency and the pavement evaluation contractor since there are no grounds to support this use; thus, the DQAIF cannot serve as such –at least, not until new research supports this type of use.

1.3. Research Methodology

The present study strived to find a way to reduce variability in manual asset condition assessments. It was focused on pavement manual evaluations, but the efforts on this study were also directed to be applicable to the condition assessment of any type of asset while the assessment is based on the subjectivity, or expert opinion, of the evaluator. The principles and components of the DQAIF were developed based on research and review of previous efforts within different engineering fields, and the procedures were developed to address the scope's needs –in this case, to fit within an asset management system. The DQAIF was then tested in the case of the Northern New Mexico Pavement Evaluation Program, by collecting and analyzing data from their 2009 summer program. Data were collected from the same pavement sections, at different times. Each time, Inter-Rater Agreement (IRA) analyses were performed to assess the variability among pavement evaluators. Linear Regression Analysis (LRA) was performed to assess the variability of the data between different assessment times. After each assessment time, actions were taken to reduce variability in the subsequent assessments, which consisted mainly on additional training focused on the issues that needed to be addressed, according to the results of the assessment. At the end, conclusions were drawn from the analyses performed, and recommendations regarding the collection and analysis of data were

developed to help practitioners to implement the proposed framework in any asset management program.

1.4. Reader's Guide to Thesis.

This thesis discusses the variability of visual asset inspections. It contains six chapters and one appendix. Chapter 1 introduces the reader into the topic of the research, and the scope and limitations of the study. Chapter 2 further explains the concepts associated to the research topic, and presents a summary of previous research performed on the research topic. Chapter 3 presents the study's scope and process, as well as the methods employed in the research. Chapter 4 introduces the Data Quality Assessment & Improvement Framework (DQAIF), which represents the main product of this thesis; this chapter also explains the process flow and the methods employed in the DQAIF. Chapter 5 presents the results of a case study performed in order to prove the applicability of the DQAIF. Chapter 6 presents the conclusions regarding the study, the answers to the research questions introduced in Chapter 3, and opportunities for future research regarding the DQAIF concepts and its use, and visual asset inspections. Appendix A is a step-by-step explanation of the process to compute inter-rater agreement (IRA) measures, other than average deviation around the median (AD_{Md}), that were not used during the study, but still are alternatives that can be employed by the DQAIF user.

CHAPTER 2. BACKGROUND & LITERATURE REVIEW

2. 1. Asset Management Systems

The framework of an asset management system has to include, at least, the following functions: a) Setting up the system objectives; b) defining system needs; c) developing and implementing the system's program; and d) monitoring or revising the system. The flow of these functions would be similar to the depiction in Figure 2 (next page). First, goals and performance expectations are established; these should be consistent with the institution's goals, organizational policies, and within budget and time constraints. Second, inventory and performance information are collected and analyzed. This information provides input on future system requirements (also called *needs*). Third, production of budget and program strategies is carried out, with the help of analytical tools and reproducible procedures, in order to satisfy agency needs and user requirements, using performance expectations as critical inputs. Then, alternatives are evaluated and the ones that better satisfy long-range plans, policies, and goals are selected. The entire process is reevaluated annually through performance monitoring and systematic processes (USDOT, 1999).

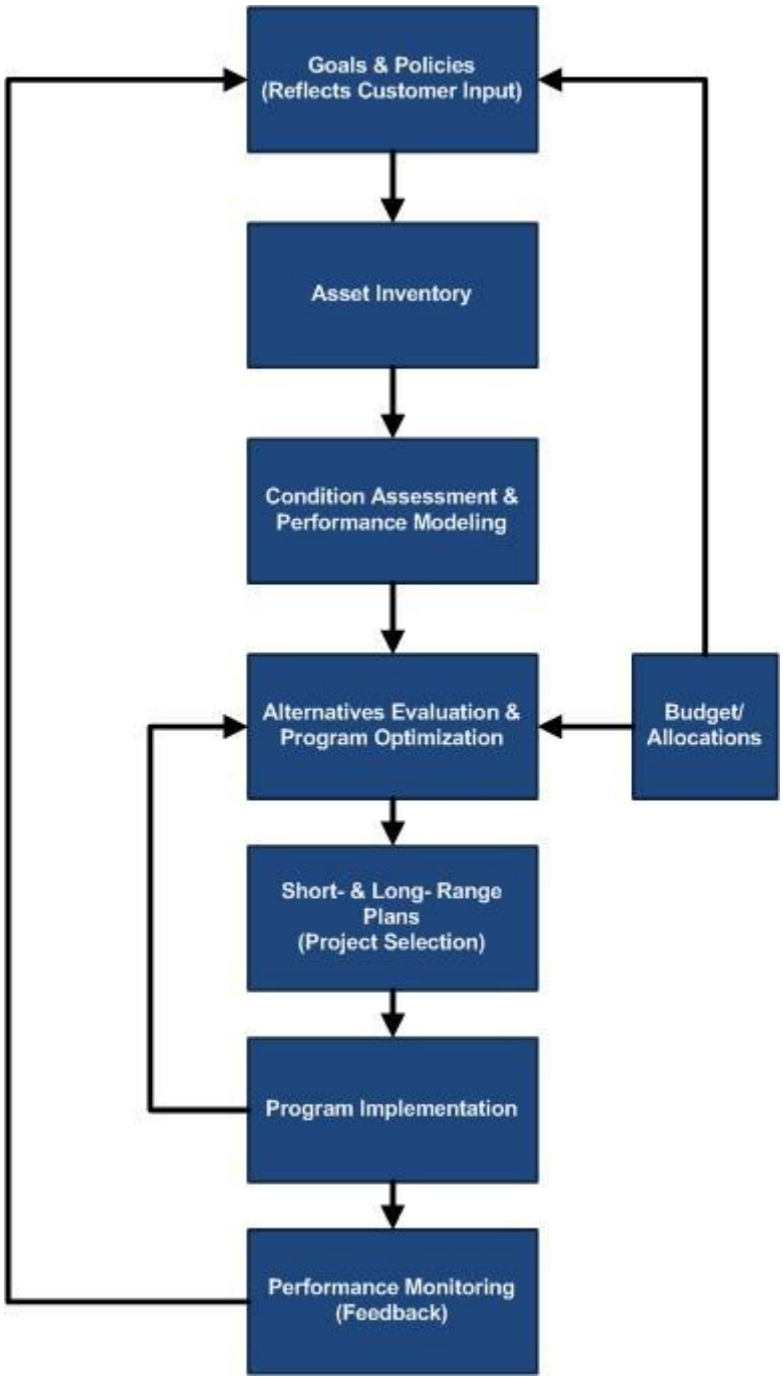


Figure 2. Generic asset management system components (From USDOT, 1999)

2.2. Pavement Management Systems

The overall structure of a PMS is comprised of the following main features -Figure 3 (Haas et al, 1994):

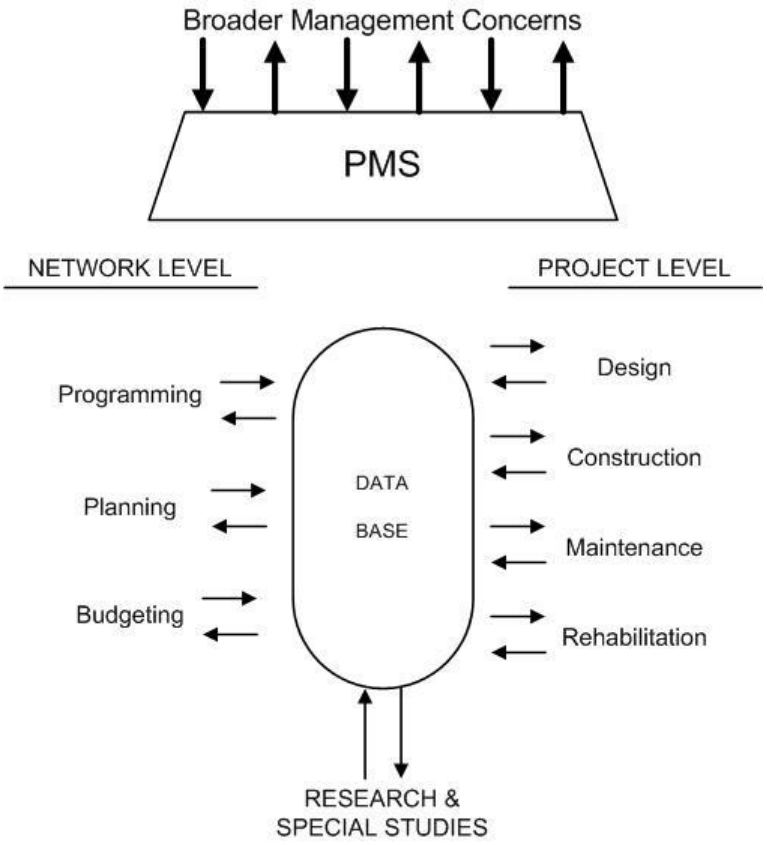


Figure 3. Major components of a pavement management system (from Haas et al, 1994).

2.2.1. Broader Management Concerns.

These are the issues and decisions made at levels higher than the network level –i.e. the overall highway administration of an entire region comprising several pavement networks, or the whole transportation management system of the network's region.

2.2.2. Network Level.

Managing at the network level has the mission of programming and scheduling maintenance and rehabilitation (M&R) or new construction work, within budget and broader management constraints. This level is further divided into –from bottom up: a) project selection level, which comprises funding decisions over certain projects, or groups of projects –the planning and budgeting subsystems are carried out at this level; and b) program level, where policy R&M decisions of the network, as a whole, are being made.

Since the limitations of the budget for the network represent the main constraint at this management level, the programming of M&R work is handled through a priority analysis with a "from a top down" flow –meaning that the results at lower levels (i.e. individual projects) are the result of the decisions made at the top (i.e. network level).

The network management level is primarily the responsibility of administrators who also work with technical input, even though this is more approximate than at the project level.

2.2.3. Project Level.

At this level, management deals with technical concerns –such as detailed design decisions- for an individual project. It represents the physical implementation of the network decisions. The activities performed at this level are just as important as the activities at the network level, since these serve the function of providing data "from the bottom up" to update the network level estimates. This pavement management level is further divided in the following subsystems:

2.2.3.1. Design.

This subsystem is the generation of alternatives concerning the assignation of the physical characteristics of a pavement system. There are several models that have been developed for this but, typically, their inputs include load and environmental factors, materials characteristics, subgrade properties, construction and maintenance variables, and costs. The outputs would be a set of design strategies that minimize total life-cycle costs –including construction, maintenance, and user costs- while satisfying user, physical, and administrative constraints.

2.2.3.2. Construction.

In this subsystem, the recommendations from design are turned into physical reality. Successful construction meets the planning and design objectives, within budget and time constraints. Some of the processes and activities associated with this subsystem are contract tendering and awarding, construction schedule, materials supply and processing, actual construction, and quality control.

2.2.3.3. Maintenance and Rehabilitation (M&R).

A complete PMS must include maintenance and rehabilitation tasks, since it's been recognized in the industry that how maintenance is carried out can significantly influence pavement performance and rehabilitation intervals –timing. Its definition may vary but, in a physical sense, maintenance consists of "a set of preventive activities directed toward limiting the rate of deterioration of a structure, or corrective activities directed toward keeping the structure in a serviceable state" (Haas et al, 1994). The separation of maintenance and rehabilitation has been vague throughout the industry –among pavement

and highway agencies- and has depended mainly on administrative policies. Thus, both type of actions are regarded within a single subsystem.

2.2.4. Research and Special Studies.

In general sense, research constitutes the tackling of problems to achieve new or better processes, materials, methods, procedures, decisions, or economy. The major elements of a long-term pavement research framework for state transportation agencies have been defined (Hudson & Haas, 1991).

2.2.5. Data Base.

A data base that includes all the aspects involved in pavement management is required to support the activities and features of a PMS. In addition, all the data should be readily accessible to any member of the pavement management staff. Thus, the data base is a central feature of a PMS that has interaction with all the other features. All the decisions regarding funding, programming, and constructing, as well as research, can be heavily supported by a comprehensive data base. The data contained in a data base include section description, performance related historic related, policy related, geometry related, environment related, and cost related data.

2.3. Pavement Condition Assessment

2.3.1. Functions of Pavement Evaluation

Pavement evaluation is the determination of the current conditions of the pavement structure by measuring and assessing its outputs (AASHTO, 2001; Haas et al, 1994; Shahin, 2005). These groups of outputs will be further explained in the forthcoming paragraphs. The function of pavement evaluation serves three main purposes within a PMS (AASHTO, 2001; Haas et al, 1994; Shahin, 2005):

a) To determine the current condition of the pavement network, in terms of the pavement outputs;

b) To project over time the future conditions of the pavement network, and so to identify when either of the outputs of the pavement will reach to a minimum or maximum level permissible; and

c) To provide with data to determine, plan, organize, and execute actions to maintain the pavement network within acceptable levels, in terms of the pavement outputs.

Even though condition assessment is not a subsystem by itself, it is a function of major relevance that supports all the elements of a PMS. It provides data of the present conditions of the entire pavement network that is stored in the agency's database. The data collected through evaluation can be used for research purposes, or in special studies, in order to improve any of the PMS' features, or the system as a whole. At a network-level management, the information on the present conditions of the pavement inventory is used to prioritize work and to update the network M & R program. At the project level,

the pavement evaluation data is used to update the design models, as well as to find opportunities for improvement in construction and M & R procedures.

2.3.2. Pavement Outputs

Among all the different types of outputs that are evaluated from pavement, four are of major importance. These four groups are (AASHTO, 1990; Peterson, 1987):

2.3.2.1. Surface Distress

Damage to the pavement surface. Distress surveys are performed to determine the type, severity, and quantity of surface distress. This information is often used to determine a pavement condition index (PCI), which can be used to compute a rate of deterioration and is often used to project future condition. Surface distress and the current or future PCI values are often used to help identify the timing of maintenance and rehabilitation as well as the fund needs in the PMS process. Distress is the measure most used by maintenance personnel to determine the type and timing of needed maintenance.

2.3.2.2. Structural Capacity

The maximum load and number of repetitions a pavement can carry. Structural analysis is normally conducted to determine the current pavement load-carrying capacity that can be compared to the capacity needed to accommodate projected traffic. Non-destructive deflection testing of the pavement is […] a reliable method to assist in making this evaluation; however, coring and component analysis techniques may be used as well.

2.3.2.3. Roughness (ride quality)

A measure of pavement surface distortion or an estimate of the ability of the pavement to provide a comfortable ride to the users. Roughness is often converted into an index such as the present serviceability index (PSI) or the international roughness index (IRI). Pavement roughness is considered the most important indicator of pavement condition by the using public, and it is especially important on pavements with higher speed limits […] It is also considered to calculate vehicle-operating costs.

2.3.2.4. Surface Friction or Skid Resistance

The ability of the pavement surface to provide sufficient friction to avoid skid-related safety problems, especially in wet weather. Skid resistance is of most importance for pavement where vehicles operate at higher speeds. It is generally considered a separate measure of the condition of the pavement surface, and it may be used to determine the need for remedial maintenance itself to address safety.

2.4. Pavement Surface Distress Evaluation

2.4.1. Overview

Physical distress is a measure of the road surface and subsurface, deterioration by traffic, environment, and aging (AASHTO, 1990). "The type, amount, and severity of distress occurring within a portion of roadway are used as indicators of how well that roadway is performing its intended function of transporting goods and people" (Gramling, 1994, p. 8).

> Most highway and airport agencies conduct periodic surface distress surveys of their pavements. They measure and evaluate various types of cracking, raveling, disintegration, deformation, and so on. Such surveys are directed in large part toward assessing the maintenance measures needed to prevent accelerated, future distress, or the rehabilitation measures needed to improve the pavement (Haas et al, 1994, p. 131).

> Distress surveys measure various types, severity, and density, or extent of distress. There is some degree of commonality between the different methods with respect to the components or factors that are usually measured. These often include the following general classes of factors:

1. Surface defects
2. Permanent deformation or distortion
3. Cracking
4. Patching

Several specific distress types exist within each of these classes.

> Pavement distress data has long been recognized by engineers as an important parameter for quantifying the quality of a pavement surface. It is important at both the network and project levels of pavement management systems, although the level of detail required for each application is considerably different. In both cases, the pavement distress information is useful in selecting appropriate treatments (Haas et al, 1994, p. 132).

For instance, at a network level management the concern would be the treatment overall strategy, program, and policies while, at the project level management, the focus would be on the specific treatment method and the extent of the repair.

2.4.2. State of the Practice

2.4.2.1. Evaluation Procedures

Among the four pavement outputs, surface distress evaluation is the one that has historically been characterized by a lack of uniformity in data collection practices, since there are currently no standards accepted by the entire transportation community (Haas et al, 1994; Gramling, 1999; Flintsch & McGhee, 2009). However, there have been important efforts to standardize distress types and severities definitions, as well as the procedures to measure these. Early attempts include the publication of a pavement condition rating report (Shahin et al, 1977a) and an *Airfield Pavement Distress Identification Manua*l (Shahin et al, 1977b) by the United States Air Force.

A few years later, the *Federal Highway Administration* (FHWA) published the *Highway Pavement Distress Manual for Highway Condition and Quality of Highway Construction Survey* (Smith et al, 1979), which provided distresses types and severities definitions, as well as practices to measure these in jointed plain concrete pavement (JPCP), jointed reinforced concrete pavement (JRCP), continuously reinforced concrete pavement (CRCP), and asphalt concrete surfaced pavement (ACP); later, that year, the United States Army Construction Engineering Laboratory published *Technical Report M-268:Development of a Pavement Condition Rating Procedure for Roads, Streets, and Parking Lots, Vol. II: Distress Identification Manual* (Shahin & Kohn, 1979), providing similar definitions for ACP, JPCP, and JRCP.

Lytton et al. developed the *Long Term Pavement Monitoring Data Collection Guide* (1985), in order to provide standards to evaluate and monitor the conditions of pavements within the Long Term Pavement Monitoring (LTPM) Program. Later, in order to optimize data collection efforts, the FHWA published the *Pavement Condition Rating Guide* (Zaniewski et al, 1985), where several pavement distresses were combined in distress types and, thus, data collection time and cost would be reduced.

However, arguably the most important effort to develop a national standard of pavement distress data collection practices is the *Distress Identification Manual for the Long –Term Pavement Performance Studies,* started by the Strategic Highway Research Program (Smith et al, 1989), and passed on to the FHWA. This manual was developed from the combination of the 1979 FHWA Distress Identification Manual, the 1985 LTPM Data Collection Guide, and the 1985 Pavement Condition Rating Guide, with the collaboration

and input from state DOTs. Updates of this manual were published by Miller et al (1993), and Miller and Bellinger (2003).

Although all these publications defined pavement distresses and their severities differently, the following distress types were found to be common in all of them (next page) (Gramling, 1994, descriptions and figures from Miller & Bellinger, 2003):

Asphalt Surfaced Pavements

Longitudinal Cracking: Cracks predominantly parallel to pavement centerline.

Figure 4. Longitudinal cracks on flexible asphalt pavement (from Miller & Bellinger, 2003).

Transverse Cracking: Cracks that are predominantly perpendicular to pavement centerline.

Figure 5. Sealed transverse crack on flexible pavement (From Miller & Bellinger, 2003).

Block Cracking: A pattern of cracks that divides the pavement into approximately rectangular pieces. Rectangular blocks range in size from approximately 0.1 m^2 to 10m^2.

Figure 6. Block cracks on flexible asphalt pavement (From Miller & Bellinger, 2003).

Reflection Cracking: Occurs in areas subjected to repeated traffic loadings (wheel paths). It can be a series of interconnected cracks in early stages of

development. Develops into many-sided, sharp-angled pieces, usually less than 0.3 meters (m) on the longest side, characteristically with a chicken wire/alligator pattern, in later stages.

(a) (b)

Figure 7. Reflection crack overview on flexible asphalt pavement (a), and reflection crack closeup (b) (From Miller & Bellinger, 2003).

Potholes: Bowl-shaped holes of various sizes in the pavement surface.

Figure 8. Pothole on flexible asphalt pavement (From Miller & Bellinger, 2003).

Rutting: A rut is a longitudinal surface depression in the wheel path. It may have associated transverse displacement.

Figure 9. Measure of a rut depth on flexible asphalt pavement (From Miller & Bellinger, 2003).

Bleeding: It is excess bituminous binder occurring on the pavement surface, usually found in the wheel paths. May range from a surface discolored relative to the remainder of the pavement, to a surface that is losing surface texture because of excess asphalt, to a condition where the aggregate may be obscured by excess asphalt possibly with a shiny, glass-like, reflective surface that may be tacky to the touch.

Figure 10. Bleeding on flexible asphalt pavement (From Miller & Bellinger, 2003).

Raveling and/or Weathering: Wearing away of the pavement surface caused by the dislodging of aggregate particles and loss of asphalt binder. Raveling ranges from loss of fines to loss of some coarse aggregate and ultimately to a very rough and pitted surface with obvious loss of aggregate.

Figure 11. Raveling on flexible asphalt pavement (From Miller & Bellinger, 2003).

Lane-Shoulder Separation/Drop-off: Difference in elevation between the traveled surface and the outside shoulder. Typically occurs when the outside shoulder settles as a result of pavement layer material differences.

Figure 12. Lane-shoulder drop-off (From Miller & Bellinger, 2003).

Jointed Concrete Pavements

Longitudinal Cracking: Cracks that are predominantly parallel to the pavement centerline.

Figure 13. Longitudinal cracks on rigid concrete pavement (From Miller & Bellinger, 2003).

Transverse Cracking: Cracks that are predominantly perpendicular to the pavement centerline.

Figure 14. Transverse crack on rigid concrete pavement (From Miller & Bellinger, 2003).

Durability "D" Cracking: Closely spaced crescent-shaped hairline cracking pattern. It occurs adjacent to joints, cracks, or free edges; initiating in slab corners. Dark coloring of the cracking pattern and surrounding area.

Figure 15. Durability cracking on rigid concrete pavement (From Miller & Bellinger, 2003).

Faulting of Transverse Joints: Difference in elevation across a joint or crack.

Figure 16. Faulted transverse joints on rigid concrete pavement (From Miller & Bellinger, 2003).

Blowups: Localized upward movement of the pavement surface at transverse joints or cracks, often accompanied by shattering of the concrete in that area.

Figure 17. Severe blowups on a rigid concrete road (From Miller & Bellinger, 2003).

Continuously Reinforced Pavements

Durability "D" Cracking: Closely spaced, crescent-shaped hairline cracking pattern. Occurs adjacent to joints, cracks, or free edges. Initiates at the intersection, e.g., cracks and a free edge. Dark coloring of the cracking pattern and surrounding area.

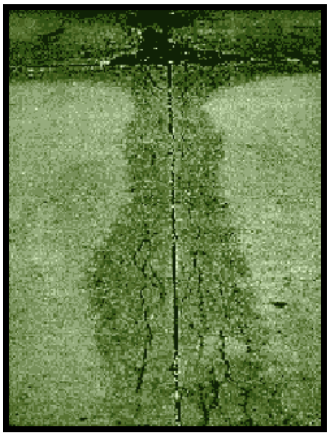


Figure 18. Durability cracks on continuously reinforced pavement (From Miller & Bellinger, 2003).

2.4.2.2. Sampling Procedures

Another aspect where discrepancy among state agencies is evident is on the sampling policies and procedures to determine what portion of the pavement network will be evaluated and the frequency in which pavement condition assessments are performed. Regarding the network sampling, some agencies evaluate 100% of the pavement network, while others only evaluate a portion of each road mile which varies from 100ft to more than one half mile. This also refers to what lanes are evaluated, particularly in the case of multi-lane roads, where some agencies evaluate a lane for each direction, while others only evaluate a single lane. Regarding the frequency of evaluations, most agencies conduct these efforts biennially or on a yearly basis (Gramling, 1994; Flintsch & McGhee, 2009; Papagiannakis et al, 2009). However, since the development of the *Highway Pavement Monitoring System* (HPMS) *Reassessment 2010+* (FHWA, 2008), state DOTs are mandated to report the Federal Government rutting and fatigue cracking data every year, and transverse cracking data every other year, in an effort to build a comprehensive database that would support budget allocations at the federal level.

2.4.2.3. Rating Procedures

Recent studies focused on the rating procedures, scoring, and indexes used by the different state DOTs (Gramling, 1994; Papagiannakis et al, 2009). The responses given by the state agencies vary to the point in which almost every state has its own signature rating system.

2.4.2.4. Evaluation Methods

Pavement distress data serves many different purposes, at different decision-levels. That, added to the development of technologies to record, store, and analyze data, has induced the development of many different methods to collect and analyze pavement distress data. Figure 19 (page 35) is a generic depiction of the most common data collection methods used by the different state transportation agencies (Haas et al, 1994; Gramling, 1994; McGhee, 2004; Flintsch & McGhee, 2009; Papagiannakis, 2009).

Pavement distress surveys may be performed by walking along the pavement section and recording the distresses observed and/or measured. These surveys provide the most precise data about the conditions of the evaluated section, but they require more time to be performed, thus being challenging to survey the entire surface of a highway network (Haas et al, 1994; Gramling, 1994).

Some agencies collect distress data while driving along the shoulder, at low speed (5 to 15 mph), and collecting data by viewing the pavement section. Since the evaluation is performed at higher speeds than at walking surveys, it is possible to cover the entire network surface at the risk, however, of collecting less detailed and/or accurate data (Haas et al, 1994; Gramling, 1994).

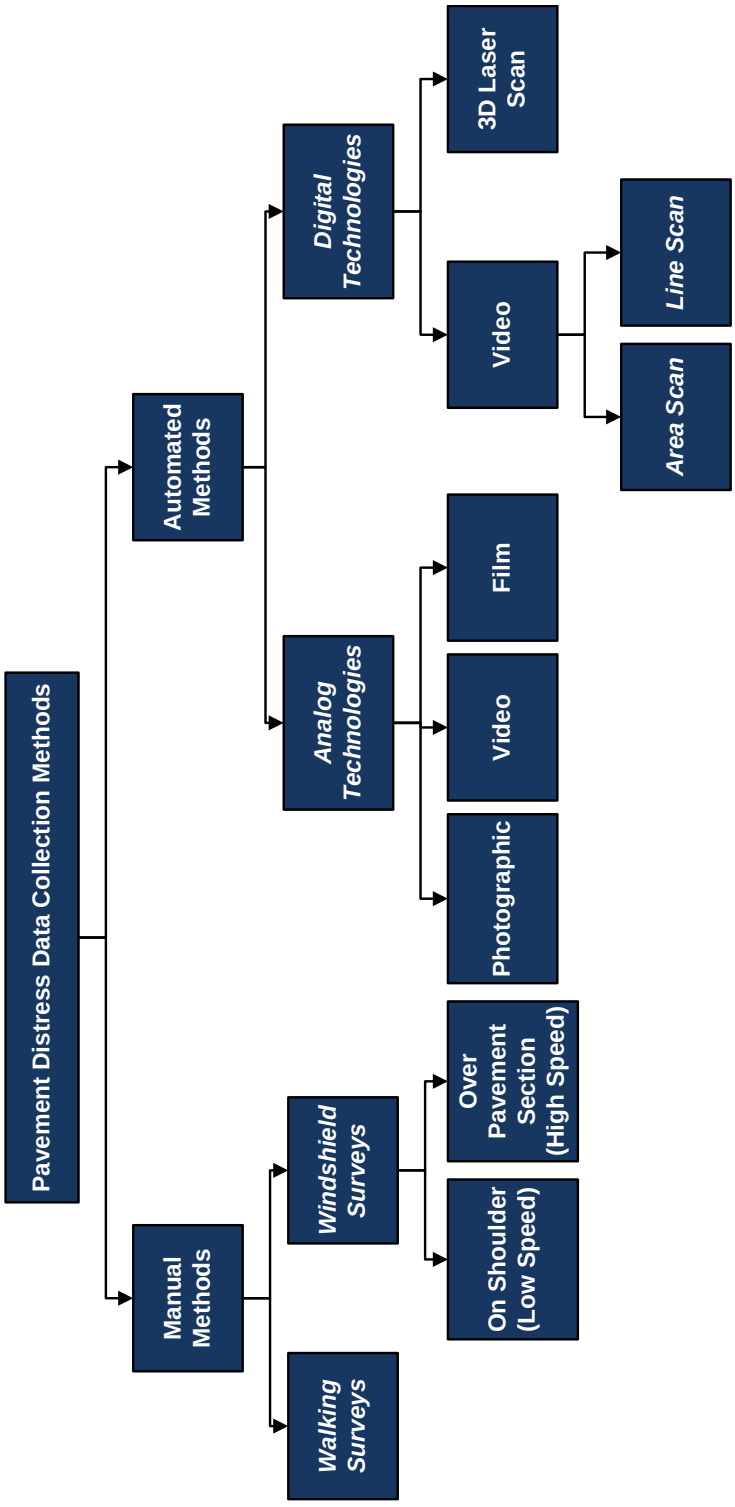


Figure 19. Pavement distress data collection methods (Adapted from Haas et al, 1994; Gramling, 1994; McGhee, 2004).

Some agencies have attempted to observe pavement distress through the windshield of a vehicle, while driving over the pavement section at high speed (5 to 55 mph). These surveys provide very general data since it is not likely to that the observer will see, recognize, and record distresses in a consistent manner. Most times, this method of evaluation is accompanied by roughness measurements, which represents the primary data element (Haas et al, 1994; Gramling, 1994).

More recently, automated surveys are conducted by using a survey system that automatically records pavement distresses on the section. These systems include taking video, filming photography, and using noncontact sensors. In general, these methods are divided into analog and digital, depending on the type of data collection device employed, as referred by McGhee (2004):

> Analog refers to the process wherein images are physically imposed on film or another medium through chemical, mechanical, or magnetic changes in the surface of the medium. Digital imaging refers to the process wherein images are captured as streams of electronic bits and stored on electronic medium. The digital bits can be read electronically for processing or reproduction purposes (pp.11-12)

As a National Cooperative Highway Research Program initiative, a survey was performed of all states and Canadian provinces transportation agencies regarding their practices for collecting data of their highways current conditions (Gramling, 1994). According to this study, the majority of the agencies still collect distress data through manual methods (i.e. walking and windshield surveys); however, the number of agencies

using automated data collection methods has increased from previous studies (Epps & Monismith, 1986)

This trend was confirmed by a later studies (McGhee, 2004; Papagiannikis et al, 2009), where most agencies responded that they are implementing, or in the process of implementing, automated methods for pavement distress data collection; although, most of these agencies are still performing manual assessments to complement the data obtained through automated methods.

It has also been reported that there is a trend to outsource the collection of pavement condition data due, in part, to the availability of technologies to collect large amounts of data (Flintsch & McGhee, 2009).

2.4.3. Quality Management of Pavement Distress Surveys

Quality, in a general sense, means "conformance to requirements" (Crosby, 1979). In any engineering project, it is common to perform quality inspections of the different deliverables –product quality, equipment functionality, construction/production processes, etc.- throughout all the stages of its life cycle to verify their compliance with the different standards that apply –owner's business scope, project specifications, equipment specifications, environmental normativity, etc. (Bogus, Migliaccio, and Cordova, 2010b). This brings the necessity for the implementation of a formal approach to organize, manage, and control quality. This approach should include methods, techniques, tools, and model problem solutions (Flintsch & McGhee, 2009). Figure 20 (page 39) shows a depiction of the components that form part of a quality management system. Through the interaction of a) processes, b) people, and c) technology, it is

possible to develop the elements necessary to perform the activities to run a system of this nature (Morian et al, 2003; McPherson and Bennett, 2005; ISO 9000:2000, 2000).

Within a quality management system, a tool of significant importance is the quality management plan, as asserted by Flintsch and McGhee (2009):

> A Quality Management Plan documents how the agency will plan, implement, and assess the effectiveness of its pavement data collection quality control and quality acceptance operations. It describes the quality policies and procedures; areas of application; and roles, responsibilities, and authorities. The Quality Management Plan is a program-specific document that describes the general practices of the program. It may be viewed as the "umbrella" document under which individual quality activities are conducted. (p. 21)

However, managing the quality of pavement distress data is challenging since 1) the end product is not clearly known (i.e. there is not a single characteristic that defines the quality of the data collected during condition assessments), and 2) the "ground truth" sometimes cannot be determined (Morian et al, 2002). These challenges are reflected in the state of the practice of this matter. The National Cooperative Highway Research Program (NCHRP) conducted a series of questionnaires to different state DOTs and Canadian provinces regarding their implementation of a formal data quality management plan (Flintsch and McGhee, 2009). The results showed that even when most of these agencies are implementing, or in the process of implementing a pavement data quality

management plan, still a large percentage (38%) do not have, or do not know if they have a plan under implementation.

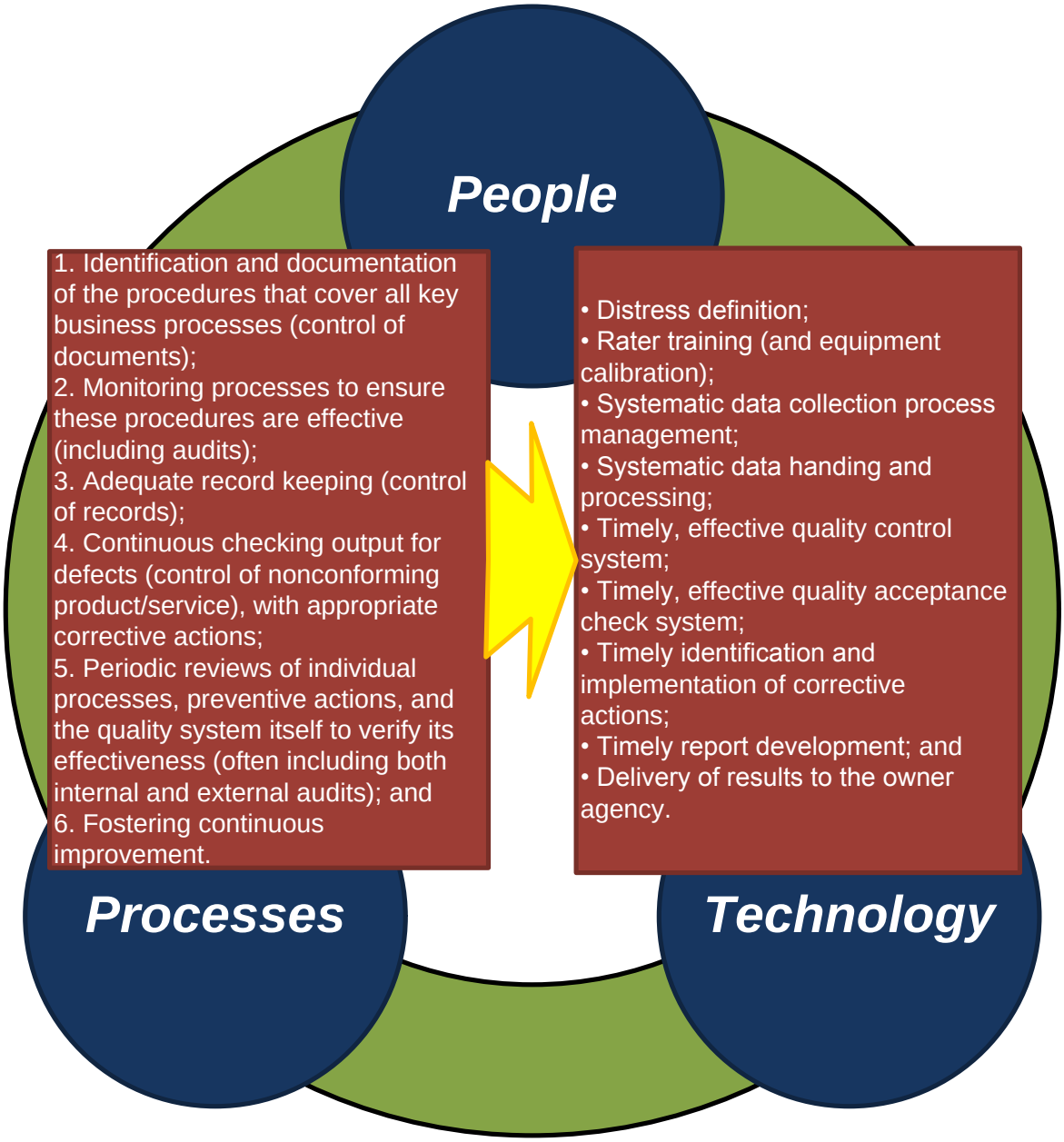


Figure 20. Quality management system components (Adapted from Morian et al, 2002; McPherson and Bennett, 2005; ISO 9000:2000, 2000).

Some of the tasks that are of significant importance within a quality management plan are quality control and quality acceptance. One of the main functions of the plan is defining how these will be carried out. These were defined by Flintsch and McGhee (2009) as follows:

> Quality control includes actions and considerations necessary to assess and adjust production processes to obtain the desired level of quality of pavement condition data […] Quality acceptance activities are those that govern the acceptance of the pavement condition data [… or the] actions taken by the buyer or user of the data to ensure that the final product is in compliance with the agreements, provisions, or specifications. (p. 23).

The most common procedures and tools used for these activities are, according to a survey performed by the NCHRP (Flintsch and McGhee, 2009):

> *Personnel Training and Certification:* Continuous training is very important to ensure that the personnel operating the equipment or conducting the visual surveys are properly trained. That the classification of the distresses is somewhat subjective makes training even more critical for the distress surveys. Some agencies require a formal "certification" of the pavement distress raters and equipment operators to verify that they have the required knowledge and skills.
>
> *Equipment and Method Calibration, Certification, and Verification* is to be conducted before the initiation of the data collection activities and

periodically thereafter to verify that equipment is functioning according to expectations and that the collection and analysis methods are being followed.

Data Verification Procedures by Testing of Control or Verification Sites are used for both quality control and acceptance before and during production. Typical verification techniques include periodic retesting of control or verification pavement segments, oversampling or cross-measurements, and reanalyzing or resurveying a sample of the sections measured by an independent evaluator. The locations of sections can be known or unknown (blind) to the data collection crews.

Software Data Checks are used during production for quality control, when the data are submitted for quality acceptance, and when the data have been entered into the pavement management database. Typical checks include network-level checks for ratings that are out of expected ranges, checks for detecting missing segments or data elements, and statistical analy[s]es to check for data inconsistencies.

Other Tools: In addition to the test described earlier, some agencies also conduct other tests, such as time-history comparisons, geographic information system (GIS)-based analysis, and verification of sample data by independent third parties.

2.5. Inter-rater Agreement –Adapted from Bogus, Migliaccio, and Cordova (2010a, 2010b)

2.5.1. Overview

The main concern regarding the use of manual pavement evaluations is its subjective nature that, accumulated along the whole team of evaluators, produces a degree of variability that may make the outputs of the evaluation not reliable enough to support a critical decision regarding rehabilitation and maintenance budgets. Reliability is the extent to which any measuring procedure yields the same or consistent results on repeated trials (Carmines & Zeller, 1979); in the case of manual pavement evaluations, reliability would refer to the extent to which pavement evaluators rate pavement the same way, regardless of the exterior factors involved as well as the differences in judgment among evaluators.

Two statistical concepts that address these concerns are inter-rater reliability (IRR) and inter-rater agreement (IRA). IRR refers to the relative consistency in ratings provided by multiple judges of multiple targets (Bliese, 2000; Kozlowski & Hattrup, 1992; LeBreton et al, 2003). Estimates of IRR are used to address whether judges rank order targets in a manner that is relatively consistent with other judges (LeBreton & Senter, 2008). In contrast, IRA refers to the absolute consensus in scores furnished by multiple judges for one or more targets (Bliese, 2000; James et al, 1993; Kozlowski & Hattrup, 1992; LeBreton et al., 2003). Estimates of IRA are used to address whether scores furnished by judges are interchangeable or equivalent in terms of their absolute value. The concepts of IRR and IRA both address questions concerning whether or not ratings furnished by one judge are ‘‘similar’’ to ratings furnished by one or more other judges (LeBreton et al.,

2003). These concepts only differ in how they define inter-rater similarity. Agreement emphasizes the absolute consensus between judges and is typically indexed via some estimate of within-group rating dispersion; reliability emphasizes the relative consistency or the rank order similarity between judges and is typically indexed via some form of a correlation coefficient (LeBreton & Senter, 2008).

2.5.2. IRA Indexes

2.5.2.1. James et al (r_{WG})

Arguably, the most popular estimates of IRA have been James et al's (1984, 1993) single-item $r_{WG(I)}$ and multi-item $r_{WG(J)}$ indices. When multiple evaluators rate a single target (e.g. a pavement or road sample) on a single variable (e.g. a distress' degree of severity) using an interval scale of measurement, IRA may be assessed using the r_{WG} index, which defines agreement in terms of the proportional reduction in error variance. The use of r_{WG} is based on the assumption that each target has a single true score on the construct being assessed (e.g., longitudinal cracking degree of severity). Consequently, any variance in evaluators' ratings is assumed to be error variance. Thus, it is possible to index agreement among evaluators by comparing the observed variance to the variance expected when judges respond randomly. Basically, when all evaluators are in perfect agreement, they assign the same rating to the target, the observed variance among judges is 0, and r_{WG} =1.0. In contrast, when evaluators are in total lack of agreement, the observed variance will asymptotically approach the error variance obtained from the theoretical null distribution as the number of evaluators increases. This leads r_{WG} to approach 0.0.

For the estimation of the inter-rater agreement over a single item, James et al. (1984) proposed the following (Formula 1):

$$r_{WG(I)} = 1 - \left(\frac{S_{x_j}{}^2}{\sigma_E{}^2}\right) \tag{1}$$

Where:

$r_{WG(I)}$ = Within group inter-rater agreement for a group of K evaluators on a single item Xj.

$S_{xj}{}^2$ = Observed variance on Xj.

$\sigma_E{}^2$ = Variance on Xj that would be expected if all evaluations were due exclusively to random measurement error.

Similarly, for the estimation of the inter-rater agreement over multiple items, James et al. (1984) proposed Formula 2:

$$r_{WG(J)} = \frac{J\left[1 - \left(\frac{\overline{S_{x_j}{}^2}}{\sigma_E{}^2}\right)\right]}{J\left[1 - \left(\frac{\overline{S_{x_j}{}^2}}{\sigma_E{}^2}\right)\right] + \left(\frac{\overline{S_{x_j}{}^2}}{\sigma_E{}^2}\right)} \tag{2}$$

Where:

$r_{WG(J)}$ = Within group inter-rater agreement for evaluators mean scores based on J parallel items.

$\overline{S_{x_J}^2}$ = Mean of the observed variances on the J items.

J = Number of items.

An important point has to be noted about the expected variance (σ_E^2) concept. This value depends on the type of statistical distribution that is followed by the expected variance due to random error. In this case, random error in pavement evaluation refers to those mistakes made in the pavement condition assessment for reasons that do not have anything to do with the protocols and procedures established. In other words, the σ_E^2 value depends on the expected numbers the evaluators will most likely assign to a pavement sample if they were not trained on how to evaluate, or if they did not have any previous knowledge on how to evaluate a pavement. This represents the major challenge in the employment of this method, because if the distribution assumed for σ_E^2 is not correct, the values of $r_{WG(I)}$ and $r_{WG(J)}$ will not be accurate. Therefore, the author recommends that the entity employing this method should make a conscious estimate of the type of distribution of σ_E^2. If data supporting the selection of a distribution is not available, the author suggests to assume a uniform distribution; that is, that an untrained evaluator is as likely to assign a number to a sample as to assign any other number.

The practical values of $r_{WG(I)}$ and $r_{WG(J)}$ range between 0 and 1; however, mathematically, the range of values expands below 0. A negative value of $r_{WG(I)}$ or $r_{WG(J)}$ clearly suggests

that something is wrong with either the evaluation system procedures or the data analysis. This could imply that the evaluation protocols are not being helpful in developing a better agreement among the evaluators. Another explanation for a negative value is that the distribution assumption of σ_E^2 is not correct, meaning that the evaluators' biases are not the same as assumed. Then, in the presence of a negative value of $r_{WG(I)}$ or $r_{WG(J)}$, it should be checked, first, if the distribution assumed for σ_E^2 is correct, or if it is necessary to use another distribution. If the distribution of σ_E^2 is not the issue (i.e. negative values are obtained with all the distributions tested), then the asset manager should verify and revise the protocols and procedures followed in the initial training of the evaluators, or provide additional training during the evaluation season.

2.5.2.2. Schmidt & Hunter (SD)

Schmidt and Hunter (1989) critiqued the r_{WG} and $r_{WG(J)}$ indices, largely based on semantic confusion arising from earlier writers' labels of the r_{WG} indices as reliability coefficients (James et al., 1984) versus agreement coefficients (James et al., 1993; Kozlowski & Hattrup, 1992). Their primary concern with r_{WG} was that it was not conceptually anchored in classical reliability theory –where reliability is defined as one minus the ratio of the variation of the error score and the variation of the observed score. Although this was an accurate statement, it is not necessarily a limitation of the r_{WG} indices because they are not reliability coefficients. In any event, Schmidt and Hunter recommended that when researchers seek to assess agreement among judges on a single target, researchers should estimate the standard deviation (SD_X, Formula 3, next page) of ratings and the standard error of the mean rating (SE_M, Formula 4, next page).

$$SD_X = \sqrt{\sum_{k=1}^{K} \frac{(X_k - \overline{X})^2}{K - 1}}$$

(3)

$$SE_M = \frac{SD_X}{\sqrt{K}}$$

(4)

Kozlowski and Hattrup (1992) rejected this approach to estimating agreement because the SEM is heavily dependent on the number of judges and because the Schmidt and Hunter approach failed to account for the level of agreement that could occur by chance. The sensitivity of the SE_M to sample size limits its usefulness as a measure of rating consensus (Lindell & Brandt, 2000; Schneider et al, 2002). These researchers have stated that the SD_X is most appropriately conceptualized as a measure of inter-rater dispersion or disagreement. Consequently, this index is not necessarily an optimal index of agreement.

2.5.2.3. Lindell et al (r^*_{WG})

Lindell and Brandt (1997) found that the concept of the James et al. (1984) multi-item $r_{WG(J)}$ could be erroneous theoretically and mathematically. It was found that $r_{WG(J)}$ is the equivalent to the Spearman-Brown correction to $r_{WG(I)}$. The recognition of r_{WG} as an agreement rather than a reliability coefficient (James et al, 1993; Kozlowski & Hattrup, 1992), calls into question the justification for the Spearman-Brown correction. Thus, Lindell et al. (1999) suggested the use of Formula 5 instead of $r_{WG(I)}$.

$$r^{*}_{WG(J)} = 1 - \frac{\bar{S}^{2}_{x}}{S^{2}_{EU}}$$

(5)

2.5.2.4. Burke et al (AD)

The average deviation (AD) index has been proposed as another measure of IRA (Burke et al, 1999). This measure, like r_{WG}, was developed for use with multiple evaluators rating a single target on a variable using an interval scale of measurement. The index is described as a "pragmatic" index of agreement because it estimates agreement in the metric of the original scale of the item (i.e., it has the same units as the item targeted). The AD index may be estimated around the mean (AD_M, Formula 6, next page) or median (AD_{Md}, Formula 7, next page) for a group of evaluators rating a single target (i.e. pavement) on a single item (i.e. pavement distress): where k=1 to K evaluators, X_{jk} is the kth evaluator's rating on the jth item, and $\bar{X}_j$ and Md_j are, respectively, the item mean and median taken over evaluators. It has been noted that the use of AD for medians may be a more robust test (Burke et al, 1999). Similar to $r_{WG(J)}$, AD can be calculated for J essentially parallel items rated by K evaluators as follows, where all terms are as defined above and j=1 to J essentially parallel items (Formulas 8 and 9, next page).

$$AD_{M(j)} = \frac{\sum_{k=1}^{K} \left| X_{jk} - \bar{X}_j \right|}{K} \tag{6}$$

$$AD_{Md(j)} = \frac{\sum_{k=1}^{K} \left| X_{jk} - Md_j \right|}{K} \tag{7}$$

$$AD_{M(J)} = \frac{\sum_{j=1}^{J} AD_{M(j)}}{J} \tag{8}$$

$$AD_{Md(J)} = \frac{\sum_{j=1}^{J} AD_{Md(j)}}{J} \tag{9}$$

As explained by Burke & Dunlap (2002), the AD index is actually a measure of disagreement, such that a value of zero (e.g., $AD_M = 0$ or $AD_{Md} = 0$) means that there is zero disagreement (i.e., total agreement). Since there is rarely total agreement among evaluators, a cut-off value of c/6 can be used to determine whether there is a consensus among evaluators, where c represents the number of response options (Burke & Dunlap, 2002). Values lower than the cut-off point mean acceptable levels of consensus, while a value that falls over the cut-off point would indicate a problem of consensus between evaluators. According to Burke and Dunlap (2002), this concept was developed from the fact that, historically, the lower limit for a meaningful reliability estimate, expressed as a

correlation between measures or ratings sources, has been in the range of 0.6 to 0.8 (Cronbach, 1990; Kaplan & Saccuzzo, 1993; Nunnally, 1978). Thus, starting with the assumption that 0.70 is a reasonable expected lower limit, Burke and Dunlap (2002) rearranged the correlation coefficient in terms of the variance and, by assuming a uniform response distribution for chance responding in the population of respondents and adjusting for average deviation, they determined that c/6 is an acceptable upper limit of consensus for the AD indexes.

2.6. Summary

All the concepts presented in Chapter 2 form the background of this study, and the basis for the methodology in Chapter 3. Sections 2.1 to 2.4 frame the environment within which the study presented in this thesis was developed. Section 2.5 introduced the concept of inter-rater agreement, and the indexes that were considered for this study. The estimation of inter-rater agreement measures is an important step within the methodology followed on this study by providing the degree of consensus between asset evaluators, in their asset condition ratings. The use of inter-rater agreement measures is presented in Chapter 3, and the process to estimate these are explained in Chapter 4 and in Appendix A.

CHAPTER 3. RESEARCH METHODOLOGY

3.1. Research Objectives

The main focus of the study was to create a process to evaluate and reduce the variability of manual asset condition assessments. More specifically, the objective is to assess and reduce the variance resulting from the subjective nature of the processes of observing and rating the conditions of an asset. This concern was explained by Bogus, Migliaccio and Cordova (2010a): Evaluators required to perform an evaluation may use a generally accepted body of knowledge (e.g., they are similarly trained), a detailed evaluation protocol (e.g., they use the same process to perform the evaluation), and/or their subjective judgment (e.g., they use their subjective experience and biases). A program that aims at being reliable would want to obtain the same results independently from who the evaluator is. Therefore, a process is needed to minimize the third effect (i.e., subjective judgment).

Thus, the main question the study answered was:

> ***How can variability of visual asset condition assessments, due to the evaluators' subjectivity, be reduced?***

In the case of rating the conditions of an asset, subjective judgment varies due to the concepts of *bias* and *experience*. Bias is a result of a person's background; it affects how this person perceives the world surrounding him or her, and how the person will react to the environment. Then, evaluators with similar exposure to the item to be rated can perceive the item's characteristics differently and, thus, submit a different rating than the other evaluator about the very same item. Experience is a result of the exposure of the

evaluator with the item to be rated, and the knowledge of the range of possibilities that item's characteristics may present. Therefore, it can be expected that variability will differ between experienced and inexperienced evaluators because experienced evaluators are expected to compensate for the variability of contextual conditions. The effects of these two concepts can be observed in the differences when different evaluators assess the conditions of the same sample, and when each evaluator assesses the same sample at different times. However, both concepts can be controlled. If the differences in background and exposure are reduced among the panel of evaluators, the differences in judgment will be reduced and, thus, the panel can produce repetitive evaluations.

Both the background and exposure of the panel of evaluators can be made uniform with additional training. However, additional training involves more time and money spent without producing the data needed from the evaluations. In addition, not all the concepts and procedures may need additional training. For this, there is a need for a framework to assess the variability of the data collected in condition assessments, and what aspects represent an issue to this matter.

Then, if the differences in visual evaluations due to the evaluators' subjectivity can be reduced and controlled within a panel of evaluators, the differences in their evaluations will also be reduced. However, contrary to what usually happens with the body of knowledge and the evaluation protocol, subjectivity is built and modified constantly; thus, controlling and/or reducing the degree of judgment differences requires continuous efforts throughout the length of the asset condition assessment. Thus, additional training, addressing the differences of bias and experience, is necessary to accomplish this.

Therefore, the hypothesis for the research question is that there is a positive relationship between the variability of visual asset condition assessments and the judgment differences among evaluators, and between the variability of visual asset condition assessments and the judgment differences of each evaluator with time. Figure 21 presents a depiction of these hypotheses, in which the construct *visual asset evaluations' variability* is measured in terms of the statistical variables obtained through the statistical procedures employed in this study; more specifically, this construct is measured in terms of inter-rater agreement indexes and linear regression analysis variables. Both constructs of *variability between evaluators* and *variability throughout time* are measured by the differences of the evaluations between evaluators and between each evaluator's assessment times, respectively.

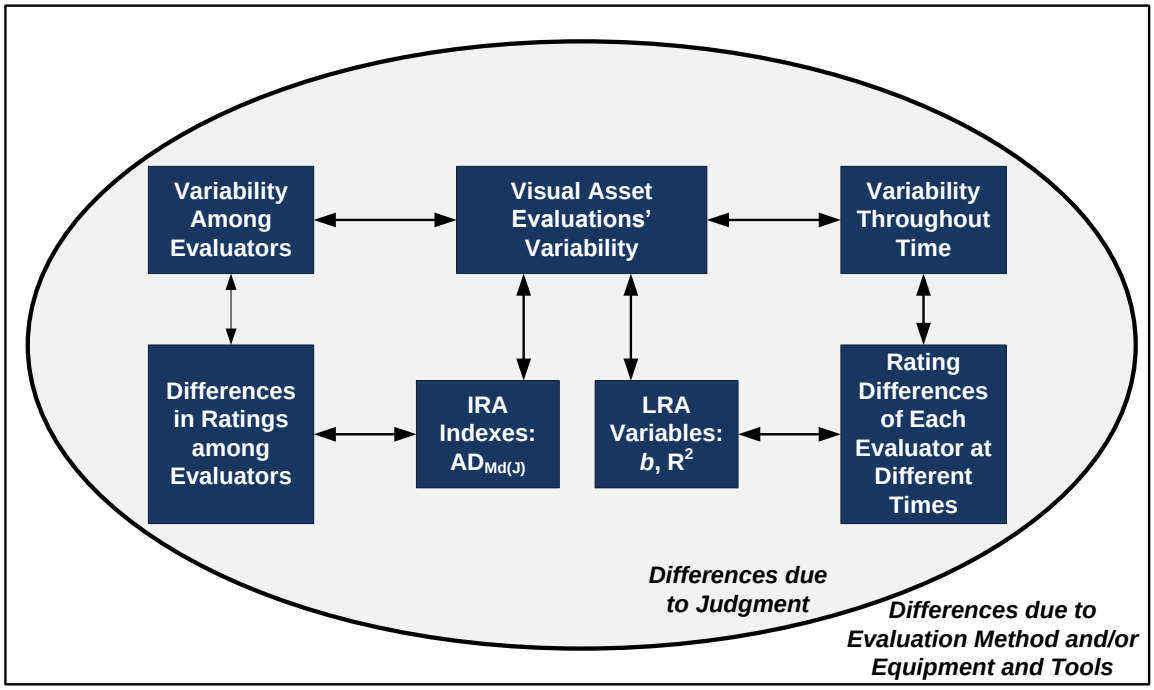


Figure 21. Constructs and variables of the research hypotheses.

With the development of the answers and hypotheses for the main question, secondary questions emerged regarding the methods that would be used to answer the main question:

Can statistical analysis be used to assess subjectivity variance?

Statistical analysis is used to monitor and control quality in the manufacturing industry, which is assessed by measuring the variance present in a specific property of the item evaluated with respect to a standard or a "ground truth". It is then hypothesized that the variance of the evaluations performed by different evaluators over the same items, at multiple times, can be assessed by performing statistical analysis of these evaluations.

What alternative can be used to identify variability among evaluators?

Of recent development, Inter-rater Agreement (IRA) indexes represent the proportion of systematic variance in relation to the total variance (Bogus, Migliaccio & Cordova, 2010a).., Then it is hypothesized that IRA measures can be used to assess variability among evaluators.

What alternative can be used to identify variability throughout time?

Linear Regression Analysis (LRA) has been proven to be an useful method to assess the differences of test outcomes measured at different times. This process is known as the Test-Retest Reliability process. Therefore, it is hypothesized that LRA can be used to assess variability of manual asset condition assessments throughout time.

3.2. Research Design

An overview of the process followed to conduct the study is depicted in the flowchart shown in Figure 22 (next page). The first step in the process was to formulate the research question. The main activities pertaining to this step were: a) setting research objectives, b) defining research scope limitations, and c) framing research sequence. These are covered in the introduction of this thesis and in this chapter. Then, a literature review of the concepts associated with the study was conducted to build a strong background to support the subsequent efforts that took place in the study. The main topics included in this step were:

a) Asset Management

b) Pavement Management Systems

c) Pavement Evaluation

d) Inter-rater Agreement

These are covered in Chapter 2 of this manuscript. With a strong background and the study objectives and process defined, the next step was to develop the process that was used to answer the research questions (i.e. the Data Quality Assessment & Improvement Framework). This part of the process is covered in Chapter 4.

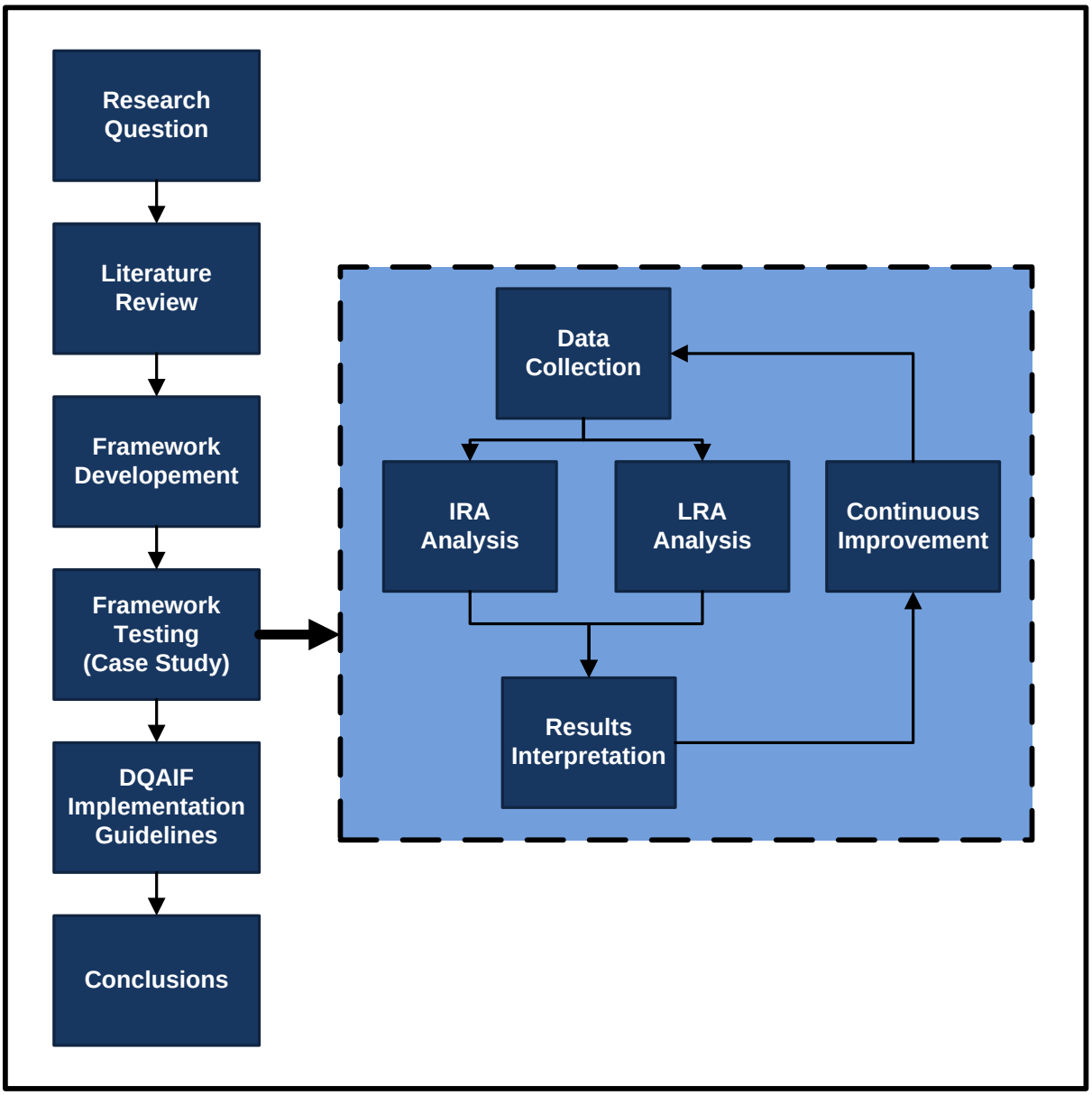


Figure 22. Research process

Chapter 5 covers the case study conducted to test the DQAIF. This step involves a sub process itself. Due to the fact that data were collected from more than one assessment time, a cycle involving 'data collection', 'IRA & LRA analysis', 'results interpretation', and 'continuous improvement' was performed. This way, after the collection of the first set of data, IRA analysis was conducted, but not LRA, because this analysis is done to compare between times of assessments. The results were interpreted from the analysis, and continuous improvement measures were taken (e.g. additional training to the evaluators) in order to control the variance among evaluators. Data were collected a second time and both IRA and LRA analyses were conducted.

The author developed guidelines for practitioners as a result of the experience obtained with the study. The guidelines address concerns about the implementation of the DQAIF within an asset management system, and questions that may arise at the moment of implementing this framework regarding data collection and analysis (e.g. the statistical methods used and their computations). These are covered in Chapter 6 of this thesis.

3.3. Data Collection

This section describes the data collection process, the case study –the Northern New Mexico Pavement Evaluation Program, and the nature of the data collected. This section has been adapted from Bogus, Migliaccio, and Cordova (2010a, 2010b).

In order to prove the usefulness of the DQAIF to monitor and control the quality of the data collected in manual asset condition assessments, data were collected from the 2009 Northern New Mexico Pavement Evaluation Project. Since 2006, the New Mexico Department of Transportation (NMDOT) has contracted with the University of New Mexico (UNM) to perform condition assessments of the public roads and highways in the northern half of the state of New Mexico's pavement network. The university hires 10-12 students as evaluators and provides them training prior to starting the assessment process. The evaluators perform evaluations on a road segment that is 0.1 mile-long, at each mile marker. The evaluations are performed by visually and subjectively assessing the severity and extent of eight different distresses for each test section [Figure 23, next page (UNM, 2009)], being:

a) The severity of a distress, the degree to which a particular distress affects the evaluated pavement, and

b) The extent of the distress, the proportion of the sample that is being affected by that particular distress.

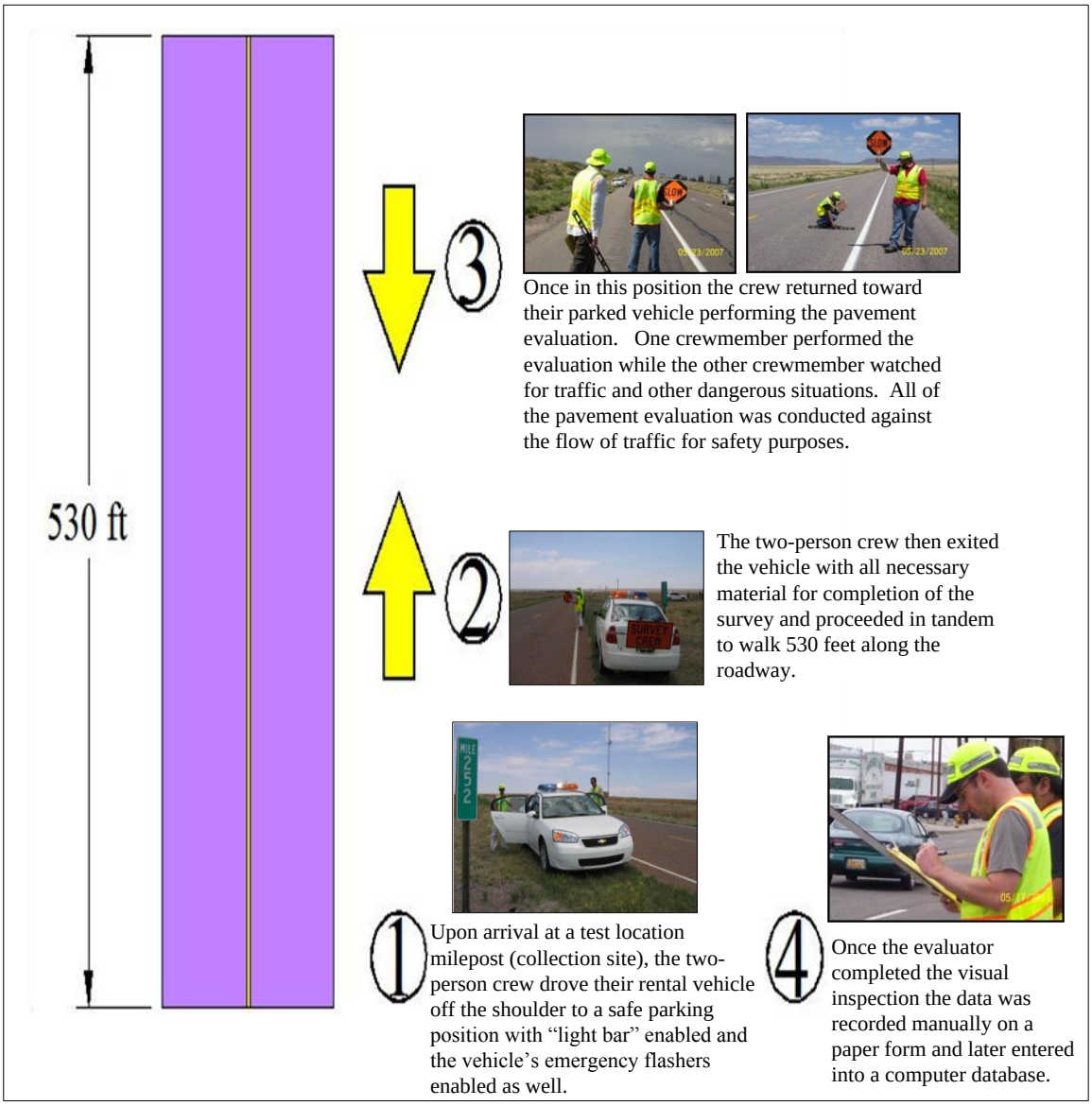


Figure 23. Field operations in the NMDOT pavement evaluation program (UNM, 2009).

Table 1 summarizes the eight different distresses measured on flexible pavement by the evaluators, and the description of each distress used by the NMDOT, and adapted from the Federal Highway Administration (FHWA) of Infrastructure Research and Development's Distress Identification Manual for the Long-Term Performance Program (LTPP) (Miller et al, 1993).

Table 1. NMDOT Distress descriptions.

Distress	Description
Raveling & Weathering	The wearing away of the pavement surface, due to dislodged aggregate particles and loss of asphalt binder.
Bleeding	A film of bituminous material on the pavement surface.
Rutting & Shoving	Longitudinal surface depressions in wheel path.
Longitudinal Cracks	Cracks predominantly parallel to pavement centerline. Location within the lane (wheel-track, mid-lane, center line) is not significant.
Transverse Cracks	Cracks that are predominantly perpendicular to pavement centerline and that extend over the entire width of the lane.
Alligator Cracks	Pattern of interconnected cracks resembling chicken wire or alligator skin.
Edge Cracks	Cracks which occur on the edge of the pavement.
Patching	An area where the original pavement has been removed and replaced with similar or different material.

For each distress, the severity and extent are rated on a 4-point scale ranging from 0 to 3. A value of 0 represents a "null" presence of the distress evaluated, or "no presence"; a value of 1 represents the "low" category; a value of 2, "medium" presence of that distress; and 3 means that the distress has a "high" presence in the road sample, according to NMDOT distress rating criteria (NMDOT, 2004). Figure 24 shows an excerpt of the

severity and extent criteria used for the Pavement Evaluation Project for one distress – rutting and shoving.

DISTRESS	SEVERITY		EXTENT	
Rutting and Shoving:	**Null: (0)**	This distress is not present.	**Null: (0)**	This distress is not present.
Longitudinal surface depressions in wheel path (Check with 4-foot rut bar).	**Low: (1)**	¼-inch to ½-inch in depth.	**Low: (1)**	1% to 30% of test section.
	Mid: (2)	½-inch to 1-inch in depth.	**Mid: (2)**	31% to 60% of test section.
	High: (3)	More than 1-inch in depth.	**High: (3)**	61% of test section, or more.

Figure 24. NMDOT Severity and extent descriptions for rutting and shoving (NMDOT, 2004).

The data collected through this program is used by NMDOT, at a network level, to compute the *distress rate* (DR), defined as shown in Equation 10:

$$DR = \sum_{i=1}^{n} (SR_i * ER_i * WF_i) \tag{10}$$

Where:

DR = Distress Rate of a particular pavement sample.

SRi = Severity Rating for the ith distress.

ERi = Extent Rating for ith distress.

WFi = Weighting Factor for the ith distress.

Here, *i* represents each of the eight distresses that are evaluated in the program; thus, *n* = 1. Then, the total DR value is the sum of the DR values of each distress (DR*i*). The

values of the weighting factors for flexible pavements are given in Table 2 (NMDOT, 2004). These are used to give each distress the effect it has in determining the performance of a pavement.

Table 2. Weighting factors for flexible pavement distresses (NMDOT, 2004).

Distress	Weighting Factor
Raveling & Weathering	3
Bleeding	2
Rutting & Shoving	14
Longitudinal Cracks	12
Transverse Cracks	12
Alligator Cracks	25
Edge Cracks	3
Patching	2

The DR value is used then by the NMDOT to compute the Pavement Serviceability Index (PSI). This is a pavement condition measure, and it is used by the NMDOT to make decisions regarding the programming and budgeting of their system efforts, at a network level. This index ranges from 0, meaning 'very poor condition', to 5, very good condition. It is calculated using one of the following empirical formulas (11, 12):

$$\text{PSI} = 0.041666 * \text{X} \quad \text{if X} \leq 60$$

(11)

or

$$\text{PSI} = [0.0625(\text{X} - 60)] + 2.4999, \quad \text{if X} > 60$$

(12)

Where X is given by Formula 13:

$$\text{X} = 100 - \left[\frac{0.6 * (\text{IRI} - 25) + (0.4 * \text{DR})}{2.9}\right]$$

(13)

Where IRI is the Interantional Roughness Index and DR is the Distress Rate. As per contract requirements, UNM has developed a quality management plan for the collection of pavement distress data. This plan applies TQM and CQI principles, at different levels of the program. Figure 25 (next page) depicts the TQM circle applied to the yearly level of the program. Figure 26 (next page) shows the different levels at which quality is controlled by UNM.

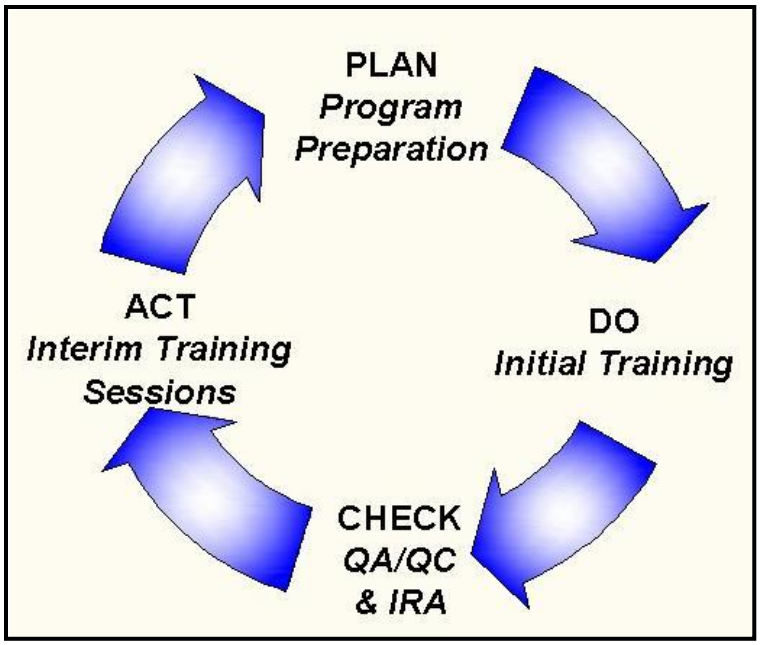


Figure 25. TQM circle of the northern New Mexico pavement evaluation program.

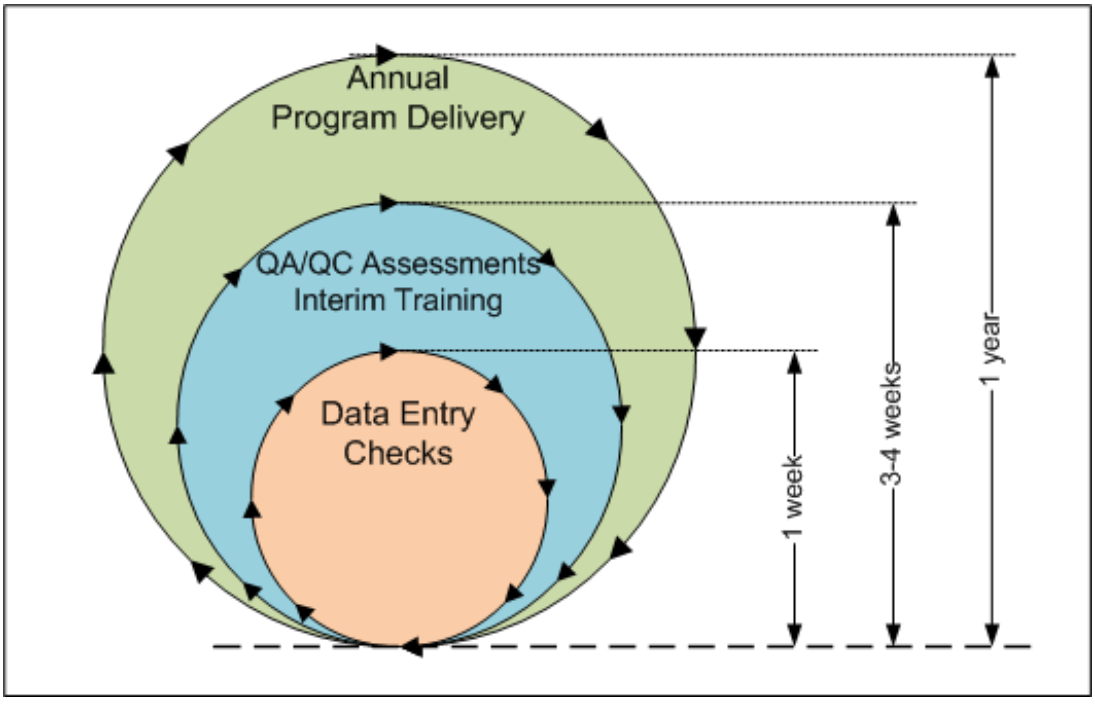


Figure 26. Quality control levels of the northern New Mexico pavement evaluation program.

As a quality control/quality assurance procedure, all the evaluators perform assessments of previously selected pavement sections at several times during the data collection project, with a span of 3-4 weeks between QA/QC round. With the data collected, DR*i* of each distress in every section are computed, and linear regression analysis is conducted to measure how the ratings vary between QA/QC rounds (UNM, 2009).

To collect data for the case study, 10 evaluators assessed the pavement condition on four stretches of pavement in each of 6 different routes in the state of New Mexico, at two different times. The roads used for the case study have AADT values that range between 500 and 16,000, representing a wide variety of roads, including one federal, and several state roads (Table 3).

Table 3. Roads used for case study data collection (From UNM, 2009).

Route	**Mileposts**
NM0041	0-3 Northbound
NM0041	29-32 Northbound
NM0014	0-3 Northbound
NM0006	0-3 Eastbound
US0550	0-3 Northbound
NM0556	12-15 Southbound

3.4. Data Analysis

The following subsections describe the processes followed to analyze the data collected as described in the previous section. An explanation of the variables involved in the computations, and what they represent is also provided.

3.4.1. Inter-Rater Agreement

There are a number of statistical methods that can be used to assess inter-rater agreement. However, not all of them can be used in situations like the ones present in manual asset condition assessments or, in the case of this particular study, manual pavement distress surveys. It has to be considered that, in most cases, a numerical discrete value is assigned in each rating; it is also of importance when selecting an inter-rater agreement measure to consider a method that is not sensitive to the size of the evaluators panel, so the results show a real picture of the variability of the data collected, regardless of how many evaluators performed the assessment.

For this study, there are three IRA sets of methods available:

a) James et al's (1984, 1993) r_{WG}

- Single-item $r_{WG(I)}$
- Multiple-item $r_{WG(J)}$

b) Lindell et al's (1999) r^*_{WG}

- Single-item $r^*_{WG(I)}$
- Multiple-item $r^*_{WG(J)}$

c) Burke et al's (1999) AD indexes

- Around the mean AD_M
- Around the median AD_{Md}

Even though all the items in the precedent list are IRA measures and their purpose is the same, each set of indexes shown in the above list are of different nature, and even in their meaning. For instance, Table 4 presents a summary of each set of indexes' characteristics.

Table 4. Summary of IRA indexes.

Index	Range of Values	Cut-off		Higher Values Mean…	Lower Values Mean…
		Value	Upper/Lower Limit		
$\mathbf{r_{WG(I)}}$, $\mathbf{r_{WG(J)}}$	$(-\infty, +\infty)$	0.7	Lower Limit	Higher Agreement	Lower Agreement
$\mathbf{r^*_{WG(J)}}$	$(-\infty, +\infty)$	0.7	Lower Limit	Higher Agreement	Lower Agreement
$\mathbf{AD_{M(j)}}$, $\mathbf{AD_{M(J)}}$	$[0, +\infty)$	c/6	Upper Limit	Lower Agreement	Higher Agreement
$\mathbf{AD_{Md(j)}}$, $\mathbf{AD_{Md}}$	$[0, +\infty)$	c/6	Upper Limit	Lower Agreement	Higher Agreement

James et al's (1984, 1993) indexes are arguably some of the most used within this group of methods. The use of these indexes is widespread in different fields, from health sciences to strategic management (LeBreton and Senter, 2008). Lindell et al's (1999) indexes were selected to backup other indexes' results, particularly in the case of extreme disagreement. In addition, these indexes have arisen as some of the most accepted alternatives of the r_{WG} indexes. Burke et al's (1999) indexes are less complex in their conception and computation, which make them a convenient and simple alternative for asset managers. For this reason the AD method is recommended to be carried forward in

the framework presented in Chapter 4. Of the two AD methods, the AD_{Md} is preferred because the value of the median of a dataset is not affected by outliers as the mean value is. Thus, the AD_{Md} measure was used in this study to estimate the consensus between evaluators.

3.4.2. Linear Regression Analysis

Linear regression analysis was performed to assess variability of pavement distress data over time, by using the DR values computed from the data collected in the study. The DR is a value that compounds both the severity and extent ratings assigned by the evaluators, which is more appropriate for the analysis, since the wide range of values that the DRs can take complies with the assumptions that rule the use of LRA. This will provide enough detail for an overall assessment (i.e. all the evaluators rating all the pavement sections), but not for a lower level analysis (e.g. assessment of variability of a particular distress, or sorting the results by pavement section, by evaluator, etc.). For analyses at a distress level, the author recommends the use of histograms for each distress, where the different rating combinations between assessment times are counted. These procedures will be further explained in Chapter 4.

CHAPTER 4. DATA QUALITY ASSESSMENT & IMPROVEMENT FRAMEWORK (DQAIF)

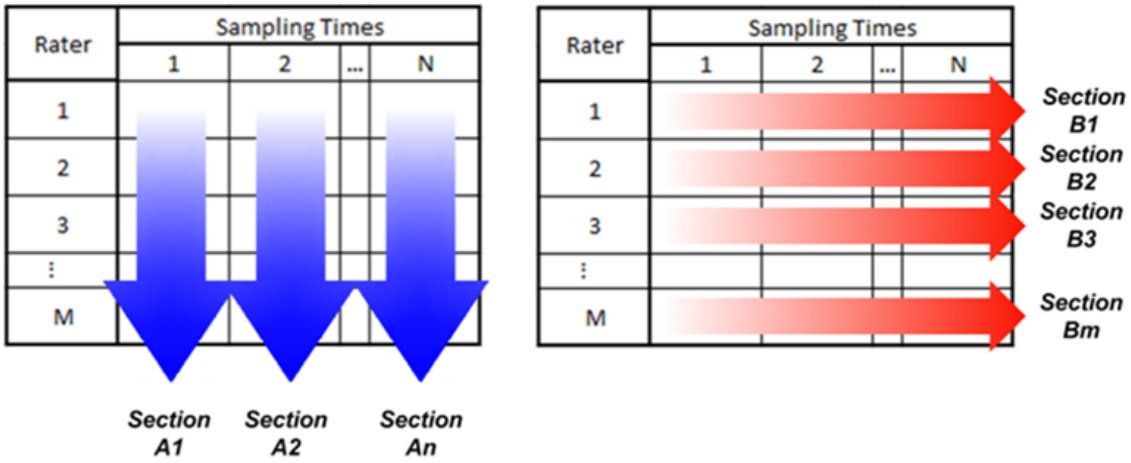
4.1. Overview

This chapter explains the DQAIF as the assessment part of the process proposed in this study. It will, first, provide an insight to the scope of the framework. The development of the DQAIF started with a conceptual structure which was then expanded as the concepts and processes where developed. The explanation of the DQAIF, thus, follows a top-down fashion, were main concepts are developed until reaching to the lowest and most detailed level of explanation: a step-by-step description of the process followed (e.g. each IRA method employed) during the assessments performed in this study. This chapter has been adapted from Bogus, Migliaccio, and Cordova (2010a, 2010b).

4.2. Conceptual Structure

The main objective of the DQAIF is to monitor and to control the degree at which data for asset condition assessments vary due to an evaluator's judgment. Since it is proposed that judgment variance has two main sources, namely a) bias and b) experience, this scope translates into monitoring and controlling these.

The DQAIF structure can be conceptualized as a two-dimensional array, like the one shown in Figure 27 (next page) (Bogus, Migliaccio, and Cordova, 2010a), where the rows represent each one of the evaluators, and the columns represent each time an assessment was performed. Variability among evaluators is assessed by computing IRA using, at each time the data stored in sections A1, A2… An, following the direction shown in Figure 27a. Variability between sampling times can be assessed by performing LRA using the data in sections B1, B2, B3… Bm, as shown in Figure 27b.



(a) (b)

Figure 27. DQAIF conceptual structure (from Bogus, Migliaccio, and Cordova, 2010a).

However, it is worth noting that each section of the array in Figure 27 represents a full set of computations where either IRA or LRA are performed. In other words, each section of the array represents a set of spreadsheets, like the example shown in Figure 28 (next page), which depicts the spreadsheets used in this study to compute AD_{Md} indexes. These spreadsheets will be explained with detail in the next sections.

Item	Judges					No. of Judges	$AD_{Md(I)}$
	1	2	3	...	K		
1	X_{11}	X_{12}	X_{13}	...	X_{1K}	k_1	$AD_{Md(1)}$
2	X_{21}	X_{22}	X_{23}	...	X_{2K}	k_2	$AD_{Md(2)}$
3	X_{31}	X_{32}	X_{33}	...	X_{3K}	k_3	$AD_{Md(3)}$
.	.	.	.	...	.	.	.
.	.	.	.	...	.	.	.
.	.	.	.	...	.	.	.
J	X_{J1}	X_{J2}	X_{J3}	...	X_{JK}	k_J	$AD_{Md(J)}$
						No. of Items	J
						$AD_{Md(J)}$	$\Sigma(AD_{Md(I)})/J$

Evaluators	Items				
	1	2	3	...	J
1	$\lvert X_{11}-Md_1\rvert$	$\lvert X_{12}-Md_2\rvert$	$\lvert X_{13}-Md_3\rvert$	...	$\lvert X_{1J}-Md_J\rvert$
2	$\lvert X_{21}-Md_1\rvert$	$\lvert X_{22}-Md_2\rvert$	$\lvert X_{23}-Md_3\rvert$	...	$\lvert X_{2J}-Md_J\rvert$
3	$\lvert X_{31}-Md_1\rvert$	$\lvert X_{32}-Md_2\rvert$	$\lvert X_{33}-Md_3\rvert$	...	$\lvert X_{3J}-Md_J\rvert$
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
K	$\lvert X_{K1}-Md_1\rvert$	$\lvert X_{K2}-Md_2\rvert$	$\lvert X_{K3}-Md_3\rvert$	...	$\lvert X_{KJ}-Md_J\rvert$
Σ	$\Sigma\lvert X_{k1}-Md_1\rvert$	$\Sigma\lvert X_{k2}-Md_2\rvert$	$\Sigma\lvert X_{k3}-Md_3\rvert$	...	$\Sigma\lvert X_{kJ}-Md_J\rvert$

Figure 28. Spreadsheet showing the overall process to compute Burke et al (1999) single- and multiple-item $AD_{Md.}$

4.3. DQAIF Process Flow

The implementation of the DQAIF follows the process flowchart illustrated in Figure 29 (next page).

4.3.1. Data Collection

The process starts with the collection of the data for which quality will be measured, in terms of the two dimensions of judgment variability. Since different asset management programs may follow different procedures and rating systems, and since the focus of the assessment and improvement efforts may be narrower than the entire body of data collected for current asset conditions, the DQAIF input data may be of a different nature from program to program. Whichever the case, the DQAIF can be used for different types of data, as long as the sets of data meet the following three conditions:

a) The protocols that define how data are collected and guide the rating rationale do not change throughout each data collection season, unless the ambiguity and/or format of these protocols are found to be the reason for unacceptable levels of data quality.

b) The input for the DQAIF has to be a set of data composed solely of discrete-number values. Some of the processes within the framework that are described in this chapter cannot be conducted with datasets different from this.

c) All the asset samples included in the input must have been evaluated by all the evaluators subjected to these assessment and improvement efforts.

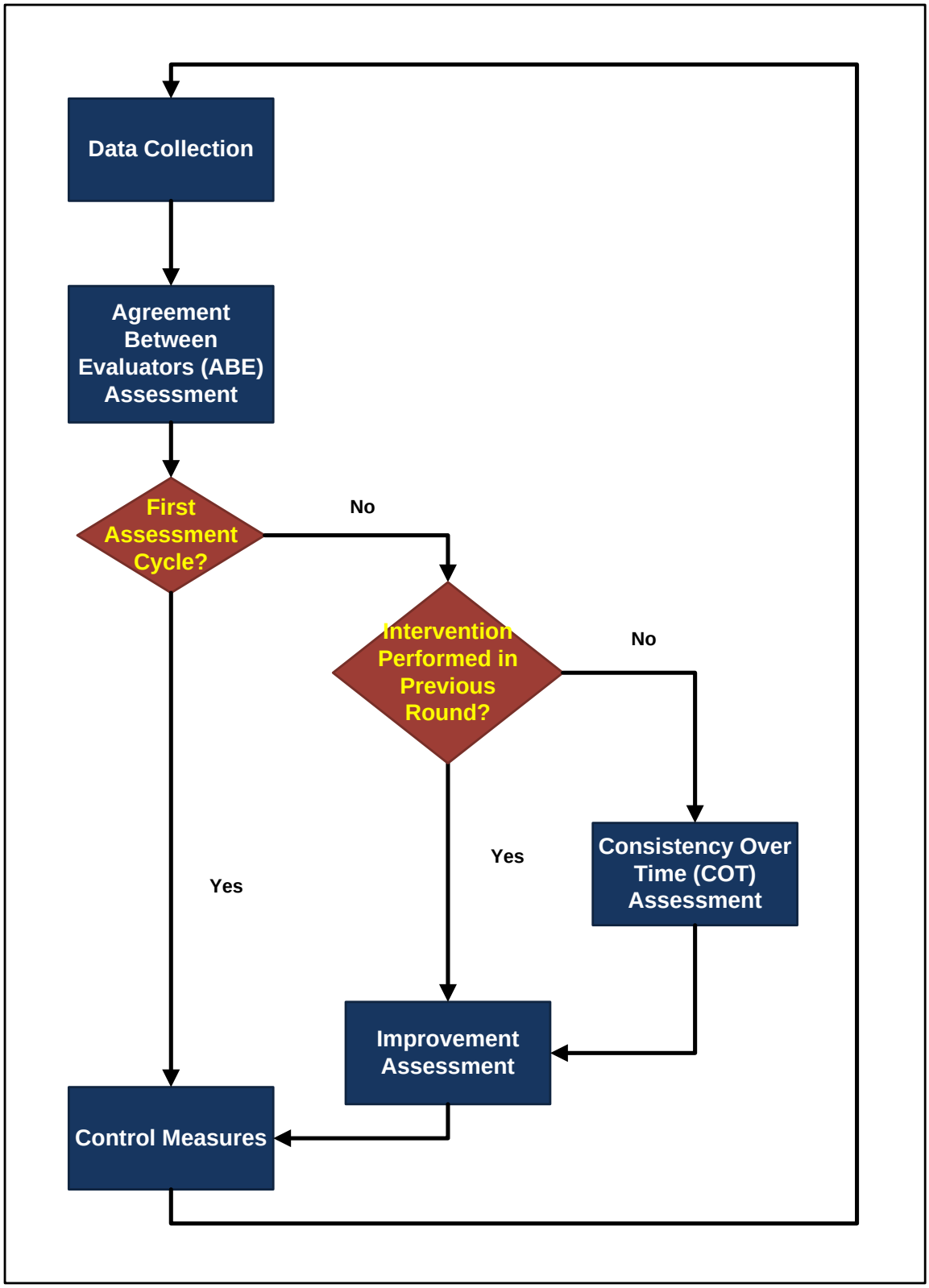


Figure 29. Overall sequence of the DQAIF process.

4.3.2. Agreement Between Evaluators (ABE) Assessment

The data will then be prepared and used to conduct an Agreement Between Evaluators (ABE) assessment in order to evaluate the panel of evaluators' rating variance. This assessment consists of the computation of IRA indexes for each item subjected to analysis (i.e. distress type, asset sample). Conclusions regarding the degree of variance can be made at this stage (i.e. identification of items whose variability meets standards and/or goals, and those which do not).Whether there are items that present data quality problems or not will direct the end of this stage, or the conduction of subsequent support analysis within the same stage.

In the case of the first cycle of assessment, the results and conclusions drawn upon the completion of this stage will constitute the basis to define the actions that will be taken to improve and/or control the quality of the data collected in the current asset conditions assessment program. In the case of subsequent assessment cycles, the process will further proceed to the conduction of a Consistency Over Time (COT) assessment and, right after, an Improvement assessment.

This section elaborates on ABE assessments, which have the goal to provide information regarding the degree of variance in the evaluation of an asset sample among the panel of evaluators through the performance of IRA analyses. The section covers the process that has to be followed during this DQAIF stage, including a step-by-step description of the procedures to perform IRA analyses and a rationale to read and interpret the results obtained during the ABE assessments.

Figure 30 depicts the framework of the process followed during the ABE assessment. In a general sense, the data quality analyst will assess the panel of evaluators' variance in this stage through the conduct of an IRA analysis, where an IRA single-item index is computed for each item (i.e. asset section or sample), and an IRA multi-item index is computed for each group of items (i.e. each distress). Once computed, the values of these indexes are compared against a cutoff value of c/6 (for AD_{Md} estimates) to pass the test. In the case of a pass, the analyst can proceed to the next stage.

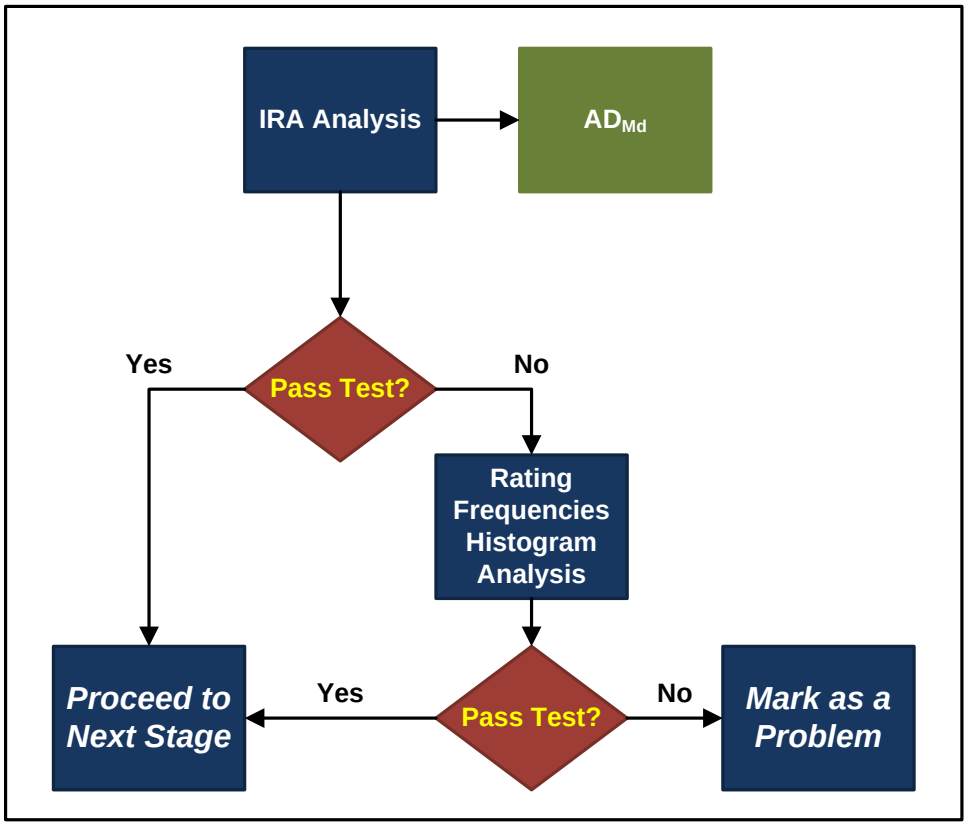


Figure 30. ABE assessment process framework.

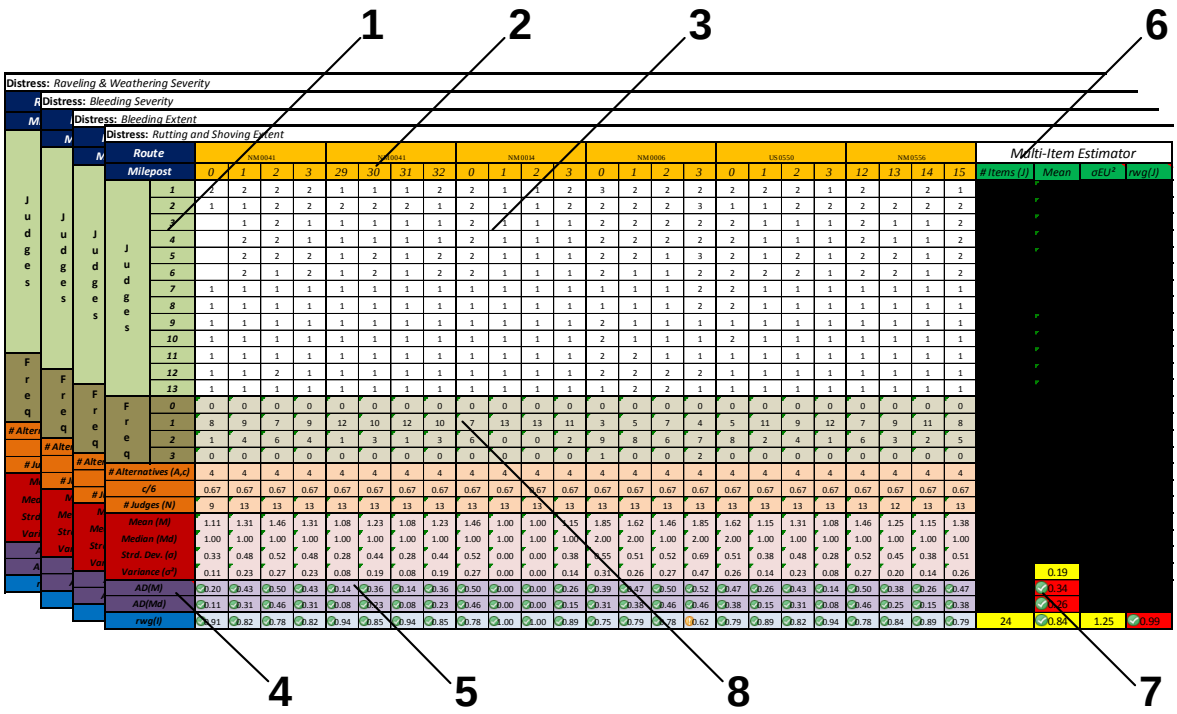
In the case of a no-pass, a second check is performed consisting of the analysis of the frequencies of each different rating value the panel assigned to an item (i.e. a pavement sample). The IRA indexes may be sensitive to outliers and the test may be failed when

almost all evaluators rated the same way except for a few who turned considerably different outputs. Thus, an analysis of the frequencies of each rating value can help the analyst understand what the situation behind poor IRA values is. Since the purpose of this test is to rule out the possibility that extreme cases are affecting the results, as a rule of thumb, the author suggests that in the case 75% or more samples present a rating that is repeated by more than half of the evaluators panel, to consider that the distress has localized issues (e.g. only a portion of the panel is rating differently from the rest of it); otherwise, it should be considered that the variability issues pertain to the entire panel of evaluators. In the latter case, the distress under assessment should be marked to be considered during the Improvement assessment.

The results will have to be analyzed as a whole for each distress (i.e. the multi-item index is the main criteria to be considered). An additional secondary criterion includes the proportion of single-item indexes at each distress level that passed the test against the total items evaluated for a distress. In this case, the author suggests that in order to pass the test, a distress should pass the single-item tests in a minimum of two thirds of the total items under assessment.

The computation of the IRA measures can be performed with the use of a spreadsheet, with a format and distribution as shown in Figure 31 (next page). Since more than one characteristic of an asset is usually assessed during asset evaluation (e.g. different asset distresses), and since the variability of the data collected for these characteristics may not be related to each other, the ABE assessment is conducted separately for each characteristic under assessment (i.e. distress severity or extent). Thus, the spreadsheet in Figure 31 should be reproduced for each one of these.

The following are step-by-step descriptions of the computation of the AD_{Md} indexes, within the ABE assessment. These descriptions were also developed for other IRA indexes (e.g. r_{WG}, r^*_{WG}, AD_M), and can be found in Appendix A.



Legend:

(1) List of Evaluators
(2) List of Items (Asset samples)
(3) Input Data (Ratings)
(4) Single-Item IRA Estimator
(5) Single-Item IRA Estimate
(6) Multi-Item IRA Estimator
(7) Multi-Item IRA Estimate
(8) Rating Frequency Counter

Figure 31. IRA analysis spreadsheet format.

As referred in Chapter 2, the estimation of the average deviation around the median (AD_{Md}) for a single item is obtained through Formula 14, and for multiple items, Formula 15.

$$AD_{Md(j)} = \frac{\sum_{k=1}^{K} \left| X_{jk} - Md_j \right|}{K} \tag{14}$$

$$AD_{Md(J)} = \frac{\sum_{j=1}^{J} AD_{Md(j)}}{J} \tag{15}$$

The procedure to calculate $AD_{Md(I)}$ and $AD_{Md(J)}$ for one distress extent or severity, based on the spreadsheet format in Figure 31, is the following:

1) *Collection of the data that will be subjected to analysis (x_{kj}):* The data (region 3 on Figure 31) should be organized by evaluator and by item. In the spreadsheet, the rows represent the data collected by the same evaluator (k) –See region 1 of Figure 31, and the columns represent the data collected in each sample (j) –See region 2 on Figure 31, or items subjected to analysis (Figure 32). However, it is worth mentioning that in the case of missed data, the cells related to that data should be left blank in order to not affect the results of the analysis.

Evaluators	Items				
	1	2	3	...	J
1	X_{11}	X_{12}	X_{13}	...	X_{1j}
2	X_{21}	X_{22}	X_{23}	...	X_{2j}
3	X_{31}	X_{32}	X_{33}	...	X_{3j}
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}

Figure 32. Spreadsheet format of x_{kj}

2) *Count the number of evaluators (k) and the number of alternatives (c):* It is important to make these counts at the beginning in order to generate the information required for the more complex calculations. It is worth noting that *k* may not be always the same for all items due to missing data or data that was not collected by any of the evaluators. However, *c* should never change during the assessment, because this will change the upper cut-off value. Figure 33 shows where these variables should be placed within the ABE Spreadsheet.

Evaluators	Items				
	1	2	3	...	J
1	X_{11}	X_{12}	X_{13}	...	X_{1j}
2	X_{21}	X_{22}	X_{23}	...	X_{2j}
3	X_{31}	X_{32}	X_{33}	...	X_{3j}
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}
# Alternatives	c				
# Evaluators	K_1	K_2	K_3	...	K_5

Figure 33. Location of the evaluator and alternatives counts within the IRA Spreadsheet.

3) *Estimation of each item median value (Md_j):* The next step is to estimate the median for each item. The median can be obtained by ordering the observations (i.e. ratings of a single sample) from the smallest to the largest value. If the number of ratings is odd, the median is the value of the rating in position (k+1)/2. If the number of ratings is even, the median will be the average of the ratings in positions k/2 and (k+2)/2. The median will be placed in the location within the IRA Spreadsheet indicated in Figure 34.

Evaluators	Items				
	1	2	3	...	J
1	X_{11}	X_{12}	X_{13}	...	X_{1j}
2	X_{21}	X_{22}	X_{23}	...	X_{2j}
3	X_{31}	X_{32}	X_{33}	...	X_{3j}
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}
# Alternatives	c				
# Evaluators	K_1	K_2	K_3	...	K_5
Median	Md_1	Md_2	Md_3	...	Md_5

Figure 34. Estimate of the median in the IRA spreadsheet.

4) *Development of the Deviation around the Median Matrix (DM_{Md}):* The upper element of Formula 14 is the sum of the absolute differences between each of the ratings of an item and their median. For this, an additional spreadsheet has to be created, called the $DM_{Md.}$. This matrix is built in a similar fashion as Figure 32 (see Figure 35), with the difference that the input data consists of the absolute value of x_{jk} and its respective Md_j, which has been computed in the previous step (Figure 34).

Below the data array, the sums of each column have to be calculated. These values represent the numerator in the $AD_{Md(j)}$ Formula (14).

Evaluators	Items				
	1	2	3	...	J
1	$\lvert X_{11}-Md_1 \rvert$	$\lvert X_{12}-Md_2 \rvert$	$\lvert X_{13}-Md_3 \rvert$	...	$\lvert X_{1J}-Md_J \rvert$
2	$\lvert X_{21}-Md_1 \rvert$	$\lvert X_{22}-Md_2 \rvert$	$\lvert X_{23}-Md_3 \rvert$	...	$\lvert X_{2J}-Md_J \rvert$
3	$\lvert X_{31}-Md_1 \rvert$	$\lvert X_{32}-Md_2 \rvert$	$\lvert X_{33}-Md_3 \rvert$	...	$\lvert X_{3J}-Md_J \rvert$
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
K	$\lvert X_{K1}-Md_1 \rvert$	$\lvert X_{K2}-Md_2 \rvert$	$\lvert X_{K3}-Md_3 \rvert$	...	$\lvert X_{KJ}-Md_J \rvert$
Σ	$\Sigma\lvert X_{k1}-Md_1 \rvert$	$\Sigma\lvert X_{k2}-Md_2 \rvert$	$\Sigma\lvert X_{k3}-Md_3 \rvert$	...	$\Sigma\lvert X_{kJ}-Md_J \rvert$

Figure 35. The deviation around the median matrix (DM_{Md}).

5) *Estimation of the Single-Item Average Deviation around the Median Indexes ($AD_{Md(j)}$):* With both the numerator and the denominator already estimated (see Figures 35 and 33, respectively), the estimation of the $AD_{Md(j)}$ values may proceed. These will be placed in the single-item IRA estimate region (no.5 in Figure 31), as shown in Figure 36.

Evaluators	Items				
	1	2	3	...	J
1	X_{11}	X_{12}	X_{13}	...	X_{1j}
2	X_{21}	X_{22}	X_{23}	...	X_{2j}
3	X_{31}	X_{32}	X_{33}	...	X_{3j}
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}
# Alternatives	c				
# Evaluators	K_1	K_2	K_3	...	K_5
Median	Md_1	Md_2	Md_3	...	Md_5
Single-Item AD_M	$AD_{M(1)}$	$AD_{M(2)}$	$AD_{M(3)}$	...	$AD_{M(j)}$

Figure 36. Estimation of the single-item AD_{Md} indexes in the IRA spreadsheet.

4) *Estimation of the Multi-Item Average Deviation around the Median Index ($AD_{Md(J)}$):* The next step is to estimate the value of the index that determines the overall status of a particular distress. According to Formula 15, the multi-item estimate represents the average of the all the single-item indexes for that distress. Thus, all these have to be summed and the result is divided by the number of items, which is also defined in the IRA Spreadsheet (Figure 37). The multi-item estimate will be placed in the single-item row, falling within the 'Average' column, in region 7 (see Figure 31) as shown in Figure 37.

Evaluators	Items					# Items	Average
	1	2	3	...	J		
1	X_{11}	X_{12}	X_{13}	...	X_{1j}		
2	X_{21}	X_{22}	X_{23}	...	X_{2j}		
3	X_{31}	X_{32}	X_{33}	...	X_{3j}		
.	.	.	.	...	.		
.	.	.	.	...	.		
.	.	.	.	...	.		
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}		
# Alternatives	c						
# Evaluators	K_1	K_2	K_3	...	K_5		
Median	Md_1	Md_2	Md_3	...	Md_5		
Single-Item AD_M	$AD_{M(1)}$	$AD_{M(2)}$	$AD_{M(3)}$	...	$AD_{M(j)}$	J	$AD_{M(J)}$

Figure 37. Estimate of the multi-item AD_{Md} in the IRA spreadsheet.

Once all the estimations have been performed, the results can be tested against the cut-off value, which represent the upper limit of disagreement. For the case of the AD_{Md} indexes, the maximum degree of disagreement is derived from the expression: Number of rating alternatives/ 6. As indicated before (Table 4 in Chapter 3), as the agreement between evaluators increases, the values of $AD_{Md(j)}$ and $AD_{Md(J)}$ decrease.

At this point, the analyst will decide whether the results for a particular distress are acceptable within the program, or not. For those elements that don't fully satisfy the requirements, a second test is performed consisting on the plot of a frequency histogram, where each possible rating will be counted from the evaluators' assessments. This tool will graphically help the analyst define whether the reason for the failure of the first test was due to disagreement within the panel of evaluators, or if it was the result of a small proportion of outliers, in which case it can be decided that a particular distress complies with the standards defined ahead.

In order to build the frequencies histogram, an eighth region has been added to the IRA analysis spreadsheet, where the frequencies of each rating value will count for each item, as shown in Figure 38 (next page). Then, a histogram can be built for each item so it can be determined the acceptance or rejection of the degree of disagreement in a particular distress. Figure 39 (page 86) depicts the frequencies histogram of an item evaluated with a 4-point scale rating system. In this case, it can be noted that most ratings were for the same, and that there is tendency towards giving a rating of 3 to the sample assessed. Therefore, the distress in question would be declared as compliant with ABE requirements.

Evaluators	Items					# Items	Average
	1	2	3	...	J		
1	X_{11}	X_{12}	X_{13}	...	X_{1j}		
2	X_{21}	X_{22}	X_{23}	...	X_{2j}		
3	X_{31}	X_{32}	X_{33}	...	X_{3j}		
.	.	.	.	...	.		
.	.	.	.	...	.		
.	.	.	.	...	.		
K	X_{K1}	X_{K2}	X_{K3}	...	X_{Kj}		
Freq. X_A	$F(X_{A1})$	$F(X_{A2})$	$F(X_{A3})$	...	$F(X_{Aj})$		
Freq. X_B	$F(X_{B1})$	$F(X_{B2})$	$F(X_{B3})$	...	$F(X_{Bj})$		
.	.	.	.	...	.		
.	.	.	.	...	.		
.	.	.	.	...	.		
Freq. X_N	$F(X_{N1})$	$F(X_{N2})$	$F(X_{N3})$	...	$F(X_{Nj})$		
# Alternatives	C						
# Evaluators	K_1	K_2	K_3	...	K_5		
Median	Md_1	Md_2	Md_3	...	Md_5		
Single-Item AD_M	$AD_{M(1)}$	$AD_{M(2)}$	$AD_{M(3)}$	...	$AD_{M(j)}$	J	$AD_{M(J)}$

Figure 38. Ratings frequencies counts within the IRA spreadsheet.

In the case of a considerably scattered distribution along the histogram, according to the analyst criteria, the element would be then marked as a data quality problem, in terms of agreement between evaluators, for improvement assessment.

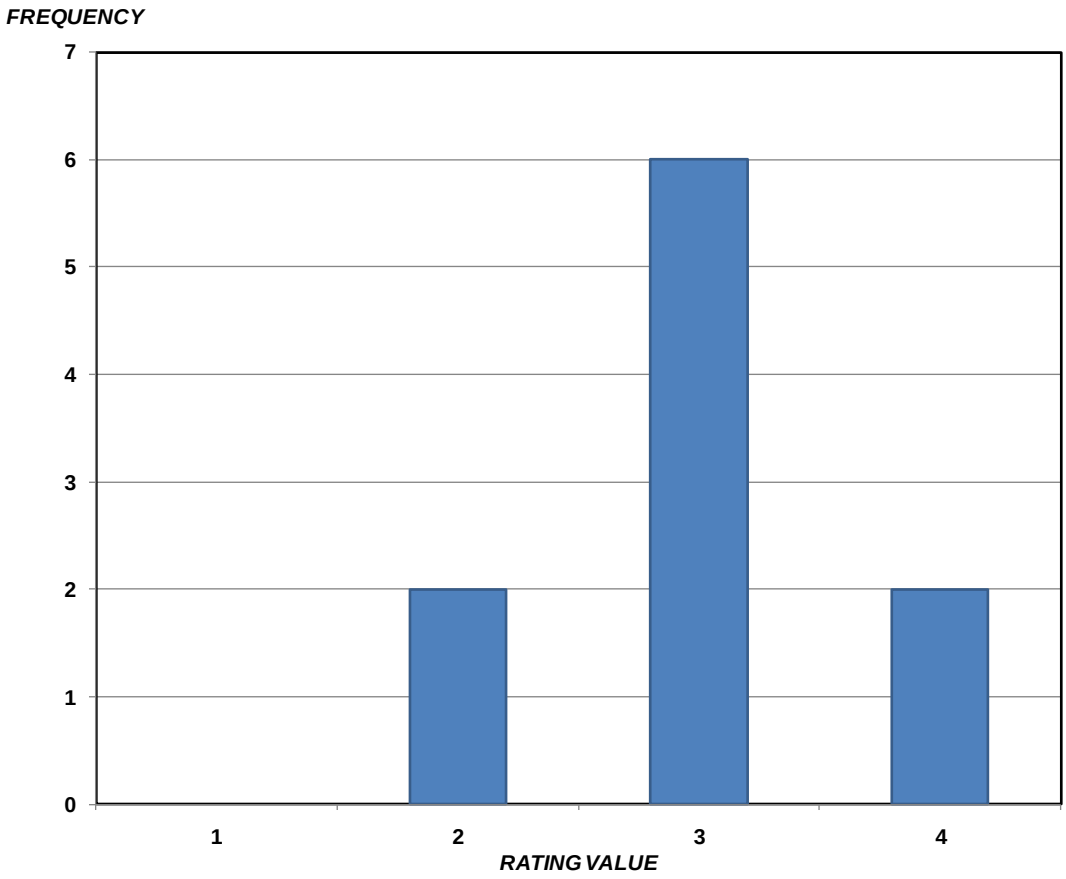


Figure 39. Ratings frequencies histogram based on a 4-point scale rating protocol.

4.3.3. Consistency Over Time (COT) Assessment

The COT assessment is performed in order to evaluate each evaluators' variance throughout time. The assessment consists of a linear regression analysis over the data collected on the current and the previous assessment cycles. With the computation of the data fitting line slope and the coefficient of determination, it can be determined whether the ratings of the evaluators have changed within acceptable ranges, or if these are changing to the degree of compromising the repeatability of manual condition assessments. However, the conclusions obtained from this analysis should be considered partial, until these are complemented with the analyses performed in the Improvement assessment.

This section elaborates on COT assessments, which goal is to provide a snapshot of the degree of variance between assessment times through the performance of linear regression analyses. This section covers the process that has to be followed during this stage of the DQAIF. A description of the procedures to perform LRA, and the rationale to read and interpret the results coming out from the COT assessments are provided.

Figure 40 (next page) shows the flow of the process that is performed in a COT assessment. It starts with the execution of a LRA to measure the variance by computing the equation of the line through the origin ($y = bx$) that best fits the pair of coordinates, which values represent the ratings of each of the two assessment times under analysis; and by estimating the coefficient of determination (R^2). These will be compared a cut-off value of 0.7 to determine whether the variance between both rounds is acceptable. If so, the COT assessment is completed by proceeding to the Improvement assessment. There is not a value that has been declared in the literature as the lower limit for these parameters, but the author suggests the value of 0.7, as being the most used in reliability measures (Cronbach, 1990; Kaplan & Saccuzzo, 1993; Nunnally, 1978).

It should be noted that the LRA is performed in such a way that the fitting line is set to pass through the origin (with pair of coordinates 0,0). This is done because in the case that is being assessed, an ideal relationship between two different times of assessments would have data points with pair of coordinates of the same number. Thus, if a zero is given on one assessment, a zero should be given in the other assessment.

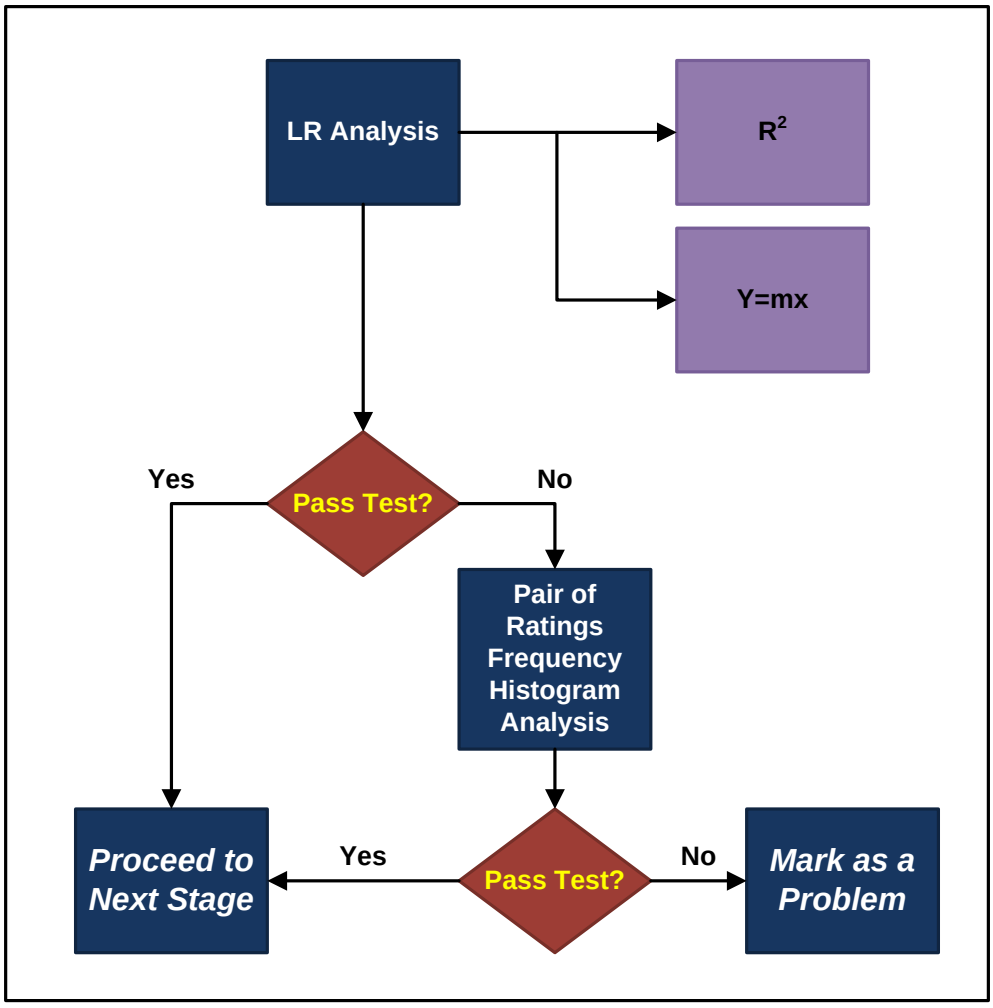


Figure 40. COT assessment process framework.

If the results do not pass the test, the analyst would proceed to perform an analysis with the use of frequency histograms of all the possible rating combinations the evaluators can make in both times of assessment, for each distress.

The procedures that will be described in this section can be performed in a spreadsheet, just like in the ABE assessments. Even though there are software packages that can perform the LRA by their own, the author recommends to make sure the tools contained in these packages can perform LRA by setting the fitting line at the origin correctly before their use.

In order to perform a LRA, a spreadsheet like the one shown in Figure 41 can be used. It contains all the elements necessary to perform this type of analysis. The steps to complete the analysis are as follow:

Evaluator	Rating Current Assess't Round	Rating Previous Assess't Round	Rating Squares		Rating Product	Fit Line Value	Residual	Residual Square
	x_i	y_i	x_i^2	y_i^2	$x_i y_i$	$\hat{y}_i$	e_i	e^2_i
1	x_1	y_1	x_1^2	y_1^2	$x_1 y_1$	bx_1	$y_1-\hat{y}_1$	$(y_1-\hat{y}_1)^2$
2	x_2	y_2	x_2^2	y_2^2	$x_2 y_2$	bx_2	$y_2-\hat{y}_2$	$(y_2-\hat{y}_2)^2$
3	x_3	y_3	x_3^2	y_3^2	$x_3 y_3$	bx_3	$y_3-\hat{y}_3$	$(y_3-\hat{y}_3)^2$
…	…	…	…	…	…	…	…	…
N	x_N	y_N	x_N^2	y_N^2	$x_N y_N$	bx_N	$y_N-\hat{y}_N$	$(y_N-\hat{y}_N)^2$
Σ=	Σx_i	Σy_i	Σx_i^2	Σy_i^2	$\Sigma x_i y_i$	$\Sigma(bx_i)$	$\Sigma(y_i-\hat{y}_i)$	$\Sigma(y_i-\hat{y}_i)^2$
			Fit Line Slope b		$\Sigma x_i y_i / \Sigma x_i^2$	Coefficient of Determination R^2		$1-(\Sigma e_i^2/\Sigma y_i^2)$

Figure 41. LRA spreadsheet.

1) *Collection of the data that will be subjected to analysis (x_i, y_i):* The data comprises the ratings collected during the current and its previous assessment cycles (x_i and y_i, respectively). The data can be analyzed separately for each distress extent and severity, or it can be compounded in a single value for each item (asset sample).It is always encouraged to perform the most detailed analysis, but since a detailed assessment will be performed during the next step of the DQAIF (Improvement assessment), the author recommends using a compounding value of all distresses extents and severity for each item (e.g. distress rates), in order to assess the overall trend of the experience variance. The data is entered in the second and third columns in Figure 41. At the bottom, the summations of both columns have to be estimated, as shown in the same figure.

2) *Estimation of ratings squares and products (x_i^2, y_i^2, and x_iy_i):* As indicated in Figure 41, the next step is to estimate the squares and products of the ratings in the fourth, fifth, and sixth columns. At the end, summations of the three columns have to be estimated.

3) *Estimation of the fitting line's slope (b):* One of the objectives of a regression model is to develop a function that fits, at the best possible, the set of data under analysis. In this particular case, since it is assumed that a sample that is given a rating of zero in the previous round should also receive a rating of zero during the current assessment round – ideally, the fitting line should then be forced to cross through the origin of the plot, for which the intercept would be zero. Thus, the equation that is modeled in this analysis is Formula 16. For this, the only parameter that has to be estimated is *b*, which represents the slope of the fitting line –this is done with Formula 17. This value will be placed in the lowest row within the LR spreadsheet, as shown in Figure 41.

$$y_i = bx_i \tag{16}$$

$$b = \frac{\sum_{i=1}^{N}(x_i y_i)}{\sum_{i=1}^{N}(x_i{}^2)} \tag{17}$$

4) *Estimation of the fitting line dependent values ($\hat{y}_i$):* Once the slope of the fitting line is known, values of the dependent variable (in this case, the previous assessment round) for each independent value (current assessment round ratings) will be estimated using Formula 18. These values will be placed in the seventh column within the LRA

spreadsheet. At the end, the sum of these values will be estimated and placed at the bottom of the column, as shown in Figure 41.

$$\hat{y}_i = bx_i \tag{18}$$

5) *Estimation of the model residuals and their squares (e_i, e_i^2):* The next steps involve the computation of the residuals of the regression model, which represent the difference between the true dependent variable value (y_i) and the one estimated with the model ($\hat{y}_i$), as shown in Formula 19. Then, the squares of the residuals will be also computed. These two concepts will be placed in the eighth and ninth columns, respectively, as shown in Figure 41. Summations of both concepts will also be estimated.

$$e_i = y_i - \hat{y}_i \tag{19}$$

6) *Estimation of the coefficient of determination (R^2):* The final step within this process is to compute the coefficient of determination, through the use of Formula 20. This value will be then placed at the bottom right corner of the LRA spreadsheet, as shown in Figure 41.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(e_i{}^2)}{\sum_{i=1}^{N}(y_1{}^2)} \tag{20}$$

Once all the estimations have been performed, the results can be tested against pre-established limits of consensus. For the case of the R^2, which varies from 0 to 1 –being 1 the value of total consensus, the literature suggests a value of 0.7 as the minimum degree of consensus. The author suggests the use of the same value, unless future studies determine otherwise. In the case of the slope of the fitting line (b), a similar approach is suggested by the author. This is, unless future research supports otherwise, a range of b between 0.7 and 1.4 can be considered as acceptable. The range comes from the same meaning of b, which is the slope of the fitting line, representing the proportion of the vertical units per horizontal unit; then, a value of 0.7 is equal to 0.7 vert.units/hor.unit; likewise, a value of 1.4 represents 0.7 hor.units/vert.units.

At this point, the analyst will decide whether the results are acceptable within the program, by considering as failed all those elements (i.e. distresses) which either value is lower than the cutoff value. For those elements that do not fully satisfy the requirements, the analyst would proceed to perform an analysis with the use of frequency histograms of all the possible rating combinations the evaluators can make in both times of assessment, for each distress, like the example in Figure 42. In the figure, the count of each combination of two different assessment times was performed for one distress. The rating scale ranges from 0 to 3. In order to pass this test, most ratings should fall within the pair of ratings with the same numbers (columns 0-0, 1-1, 2-2, and 3-3). In this case, when a 0 was given in the first assessment time, the same was given in the subsequent assessment time. However, the distribution is more spread in the other sets of columns. Since the columns with the pair of the same numbers are not predominant in all the sets of

columns, the distress under assessment in Figure 42 is declared as a no-pass. In the case of a no-pass, the distress is marked as of consideration in the improvement assessment.

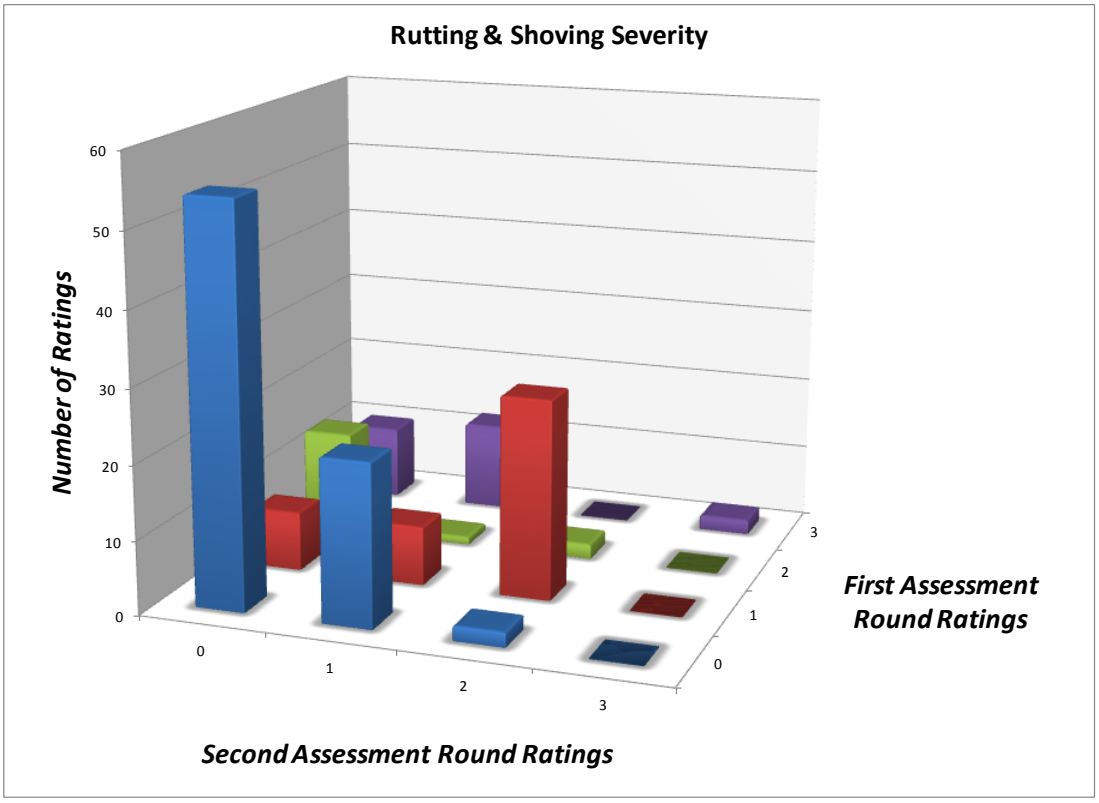


Figure 42. Rating count histogram of two assessment times for one distress.

4.3.4. Improvement Assessment

The next step is the conduction of an Improvement assessment. This stage consists of performing a graphical analysis of the results obtained in the ABE assessments of the current and the previous cycles. By plotting these results in a radar graph, like the one shown in Figure 43, conclusions can be drawn regarding the sources of data quality issues, and the measures that can be taken to counter these. It can also show whether the quality of the data has improved. For instance, if all the data of the current assessment period show higher variability than their predecessors, then it can be concluded that the quality of the data is poorer. This could be due to the fact the evaluators build their

judgments differently, because of a different exposure to the asset distresses in the samples each of them assess. Then, additional training on all the different distresses may be needed to address this issue. This stage should be conceptualized as the problem-and-solution seeker step, where the specific quality problems are identified, and their solution or improvement can also be envisioned, while the previous stages are the variability red spotters, where the presence of quality issues can be noted. At this stage, comparisons are made between assessments cycles and against established levels of acceptance and/or degree of variability goals.

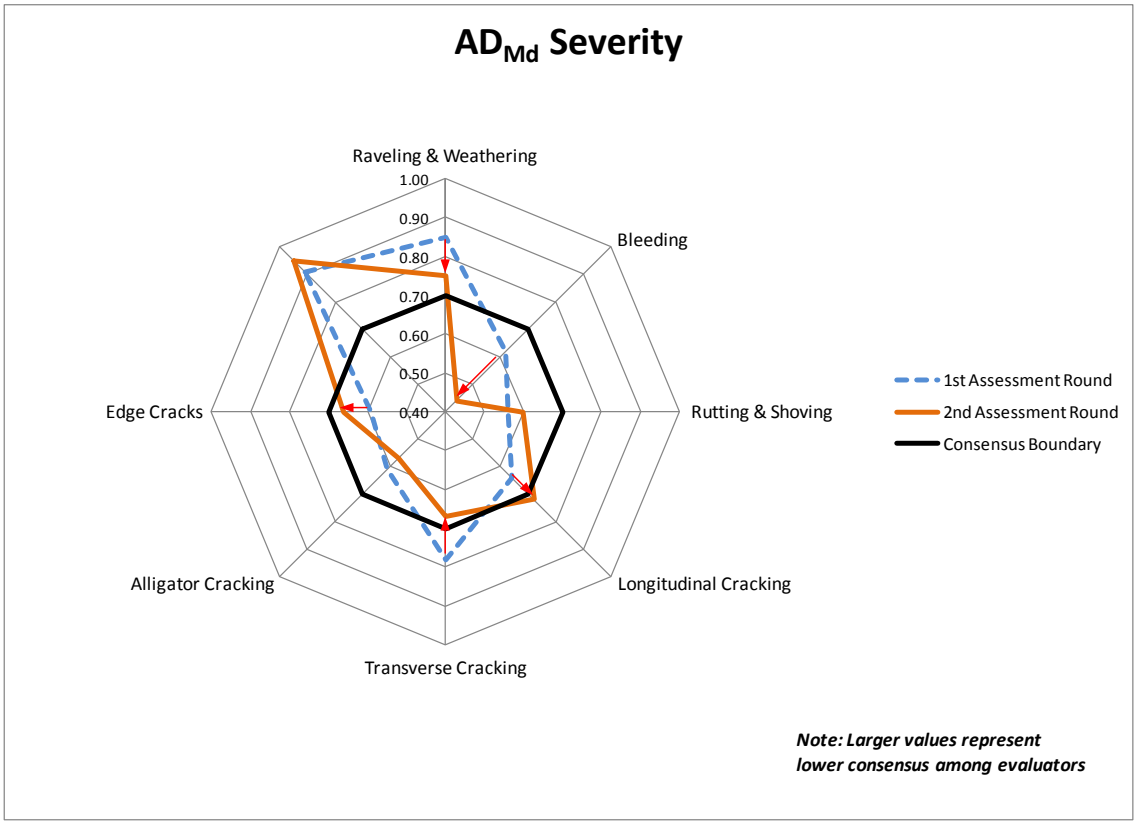


Figure 43. Radar graph used during the improvement assessment stage.

This section elaborates on improvement assessments, which have the goal to provide the analyst information regarding the overall and particular status of the data collected through the evaluation program. The latter is done by performing an analysis of a radar graph, which includes data from the current and the previous assessments rounds.

Figure 44 (next page) depicts the framework of the process followed during the Improvement assessment. In a general sense, the data quality analyst will assess the status of the data collected within the program during this stage, through the analysis of the data from both the current and the previous assessment rounds plotted in a radar graph, like the one shown in Figure 43. Analyses of both the overall trend and each item trends will be performed, by comparing the results against the pre-established limits discussed in the previous stages of the DQAIF, and against other assessment times. If after the analyses it is concluded that the results are satisfactory, then an Improvement Opportunities Analysis will be conducted to find potential actions that may improve the quality of the data collected in the asset evaluation program, or it can be found where there is a potential risk for quality issues to be present in the future. This is part of the CQI philosophy, which is an element of the proposed DQAIF.

In the case where the results are not satisfactory, a Problems Roots Analysis has to be performed in order to find the reasons that led to the current situation. Reasons for an overall unsatisfactory status, or particular element's unsatisfactory results (e.g. asset distresses where the issues are focalized), should be identified during this step of the process.

Once each of these analyses are completed, the analyst together with the management team can design improvement efforts. In the case of controlling the evaluators' judgment, most of these efforts would consist of additional training on the protocols followed by the program, with an emphasis on those elements which data quality need to be addressed. Once this is done, the team is ready to proceed to the next step: the implementation of the defined control measures.

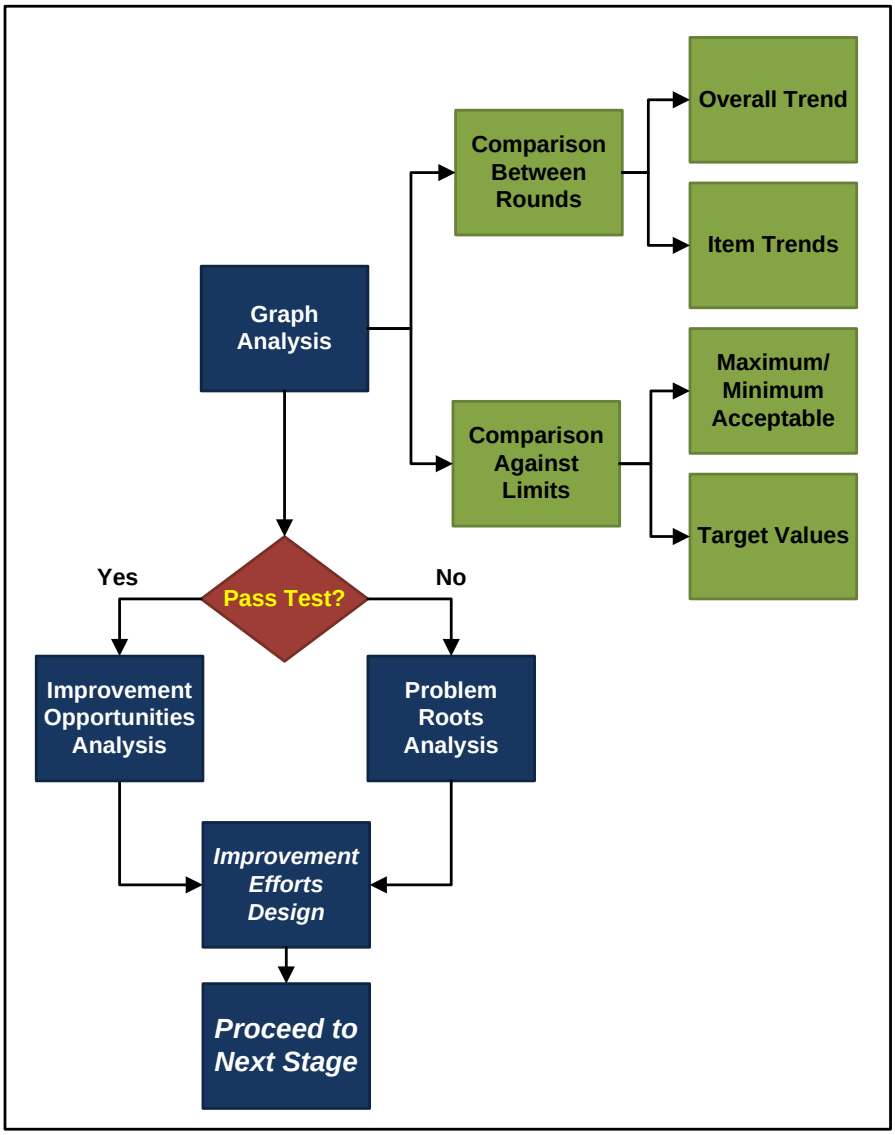


Figure 44. Improvement assessment process framework.

It is worth to noting that the analyses conducted during this DQAIF stage should consider the results and conclusions from previous stages. These might provide information that could help the analyst make better-informed decisions during the Improvement assessment.

The construction of the radar graph that will be the tool of analysis during this stage is left to the judgment of the analyst, according to the needs of the program, or to the results shown in previous stages. However, it is important to maintain consistency in the analyses in order to avoid comparing elements that are not comparable (e.g. rating data from different distresses, trying to compare extents' results against severities').

4.3.5. Control Measures

At the end, the results from the analyses and the comparisons between rounds can be interpreted. If it is interpreted that there are variability issues and their sources are identified, then the asset manager can address these by conducting additional training of the evaluators prior to the next time of assessment. In the case of extreme variance, it may be considered to revise the protocols and descriptions used to perform the asset evaluations.

4.4. Summary

This chapter introduced the DQAIF, and elaborated on each of its processes. In general, the DQAIF consists of a set of sequenced procedures to monitor and control the data collected on manual asset evaluation programs. These procedures have the function of assessing three main elements: 1) variance among evaluators, though ABE assessments;

2) variance with time, through COT assessments; and 3) overall and itemized status of the data quality, and improvement opportunities, through Improvement assessments.

ABE assessments are conducted by performing IRA analyses, with the help of the IRA spreadsheet and the deviation matrixes; and it is supported by performing a ratings frequencies analysis, with the help of a frequencies histogram. COT assessments are conducted by performing LRA, with the help of the LR spreadsheet; and it is supported by the use of scatter plots and pair of ratings frequency histograms analyses. Improvement assessments are carried out by performing radar graph analyses, and improvement measures are developed as result of these analyses.

CHAPTER 5. CASE STUDY: NORTHERN NEW MEXICO PAVEMENT EVALUATION PROGRAM

Adapted From Bogus, Migliaccio, And Cordova (2010a, 2010b)

5.1. Overview

In order to show how the DQAIF can be used to assess and monitor the performance of manual asset condition assessments, data were collected from the 2009 Northern New Mexico Pavement Evaluation Project. The New Mexico Department of Transportation (NMDOT) contracted with the University of New Mexico (UNM) for the condition assessment of the northern half of NMDOT's pavement network. For this, UNM hired 10 pavement evaluators who would stop at each mile marker along all Interstate, United States Federal, and New Mexico highways in the northern half of New Mexico, to assess the condition of a 0.1 mile-long, one lane-wide flexible pavement segment. The evaluators assigned a discrete number ranging from 0 to 3 to eight different distresses severities and extents, based on criteria developed by the NMDOT.

UNM developed and implemented a quality assurance and quality control (QA/QC) plan, as part of the agreement with NMDOT. Part of the plan consisted of having each evaluator perform a distress evaluation at the same 24 locations at several different times, so that the results could be compared across evaluators and across time. In order to avoid biases among the evaluators, these two rounds of evaluations were performed several weeks apart. The data gathered during these quality checks were analyzed with the DQAIF.

The remaining parts of this chapter present the results obtained in the different analyses the data were subjected to. It also explains the reasoning behind the DQAIF

implementation that took place during the same study. Since the DQAIF itself is a continuous cyclic process, the chapter has been organized in a chronological manner, in order to not confound the reader by mixing results from different times of assessments.

5.2. Data Analysis & Results

5.2.1. Assessments from Previous Years

In order to have a better perception of the effects of an intervention to control the quality of the data collected in visual asset evaluations, the data collected during the years of 2007 and 2008 in the Northern New Mexico Pavement Evaluation Program were used to perform IRA and LRA. These analyses provided results that could be compared against the results obtained during the 2009 season, were the DQAIF was implemented for both measuring and improving the data collected.

In order to obtain a perception of the trends of the quality of data collected through visual inspections without intervention or additional training, only the first test of each step in the DQAIF were performed. In other words, only the computation of $AD_{Md(J)}$, IRA, and plotting of the radar graphs were carried out for the data from years 2007 and 2008. The reason why only the first test in each DQAIF step was performed is that these were considered by the author to satisfactorily provide a global picture of the data quality in those years. The following subsections present the results obtained in these analyses.

5.2.1.1. 2007 Northern New Mexico Pavement Evaluation Program Analysis

Figure 45 (next page) shows the two radar graphs of the IRA indexes computed for the three assessment rounds performed during the 2007 season, one for all the distress severities (top graph) and another for the distress extents (bottom graph). From these figures, it can be noted that all distresses passed the ABE tests (lower values of AD_{Md} indicate better agreement between evaluators or less variability). However, there is an overall trend of the results to become poorer after each assessment time, as can be noted from the higher values of AD_{Md} of the second and third assessment rounds, compared to the ones of the first assessment round, in most distresses. This supports the idea that the data also varies with time, and that the tendency is on the negative side.

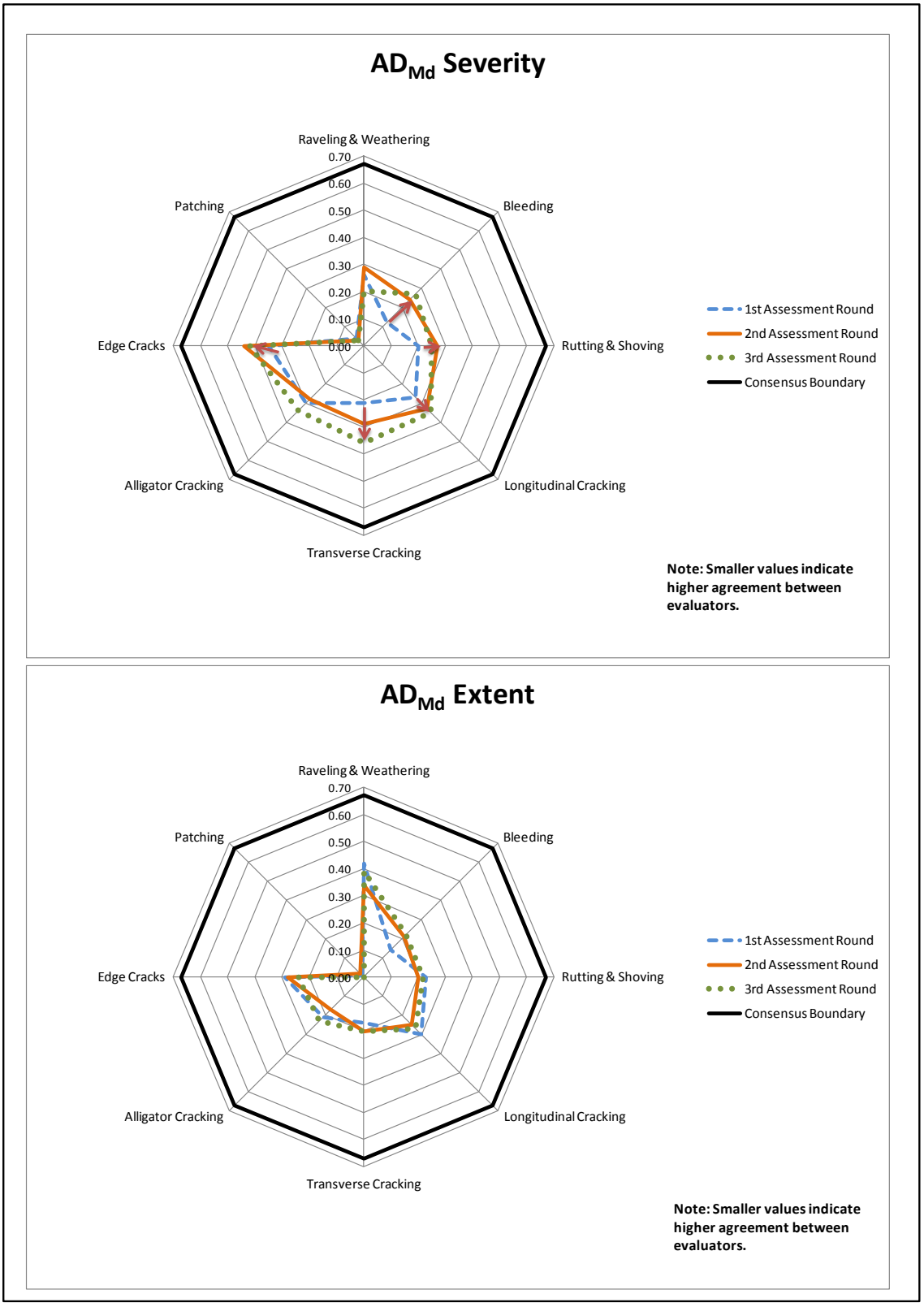


Figure 45. AD_{Md} results of the 2007 season.

The results of the LR analysis are shown in Table 5. The values of $b < 1$ indicate that there was an overall trend to give lower ratings (lower distress severities and extents) as the program progressed. Moreover, the differences in ratings between assessment times decreased from the first assessments to the last. The values of R^2 close to 1.0 suggest that the above-mentioned trends apply to all the evaluators. By isolating the results of the COT assessment, it could be concluded that there was a decrease in variability between assessment times as the program progressed. However, when looking at these results together with the ABE assessments', the impression is that as data collection progressed, the evaluators were building their judgment differently from the others. In conclusion, there were no variability issues during the season of 2007, but there was a tendency where the ratings were becoming more dissipated between evaluators as the project progressed.

Table 5. 2007 COT season results.

Assessment	*b*	R^2
First Round vs Second Round	0.688	0.852
Second Round vs Third Round	0.961	0.916
First Round vs Third Round	0.679	0.819

5.2.1.2. 2008 Northern New Mexico Pavement Evaluation Program Analysis

Figure 46 (next page) shows the radar graphs of the AD_{Md} indexes computed for the three assessment rounds performed during the 2008 season. Both graphs show that almost all the distresses passed the ABE assessments throughout the 2008 season, except for bleeding severity which represented an issue during the three assessment times. On the other hand, just as it happened in 2007, the data in collected during 2008 tended to vary more between evaluators as the program progressed.

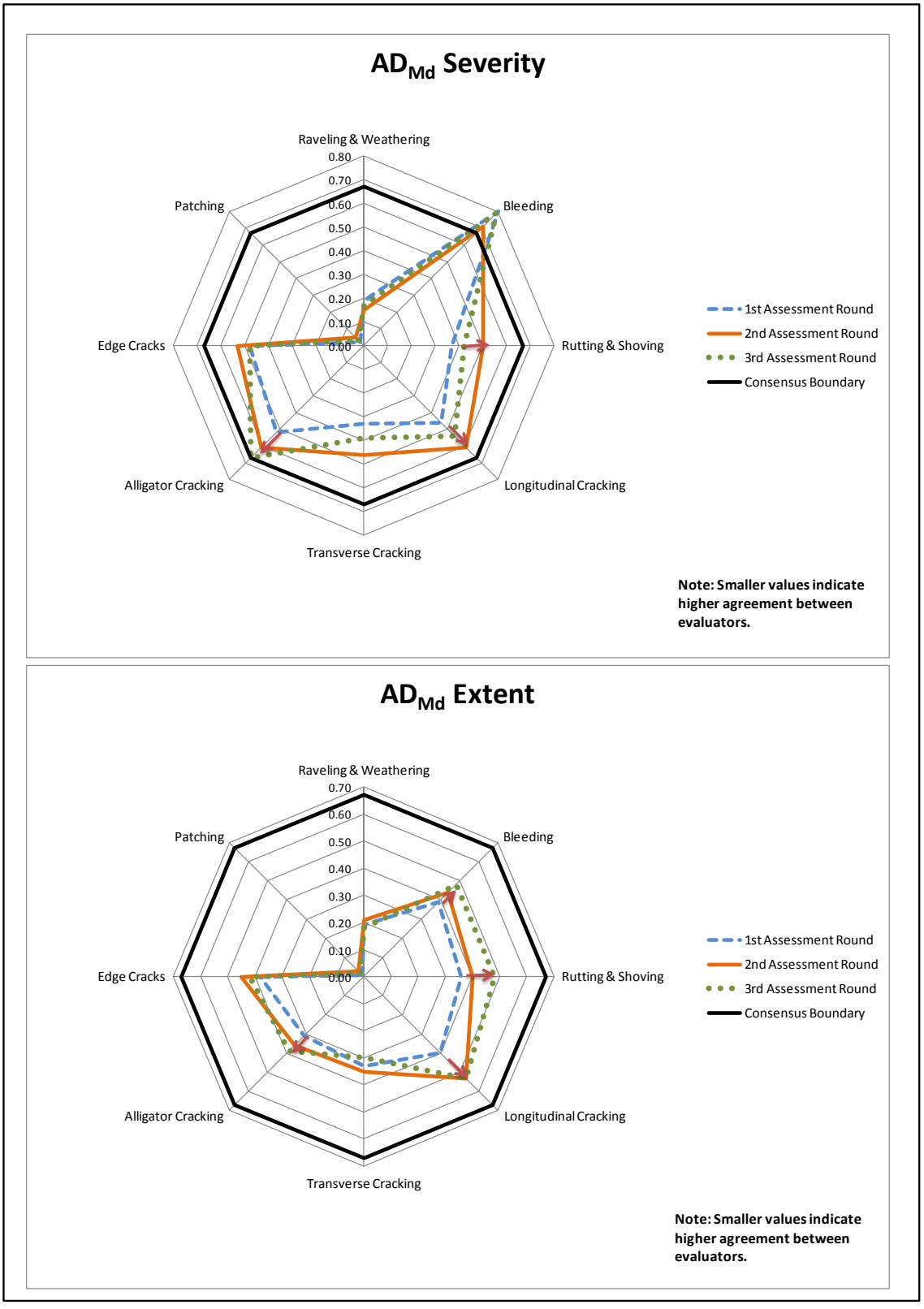


Figure 46. AD_{Md} results of the 2008 season.

A summary of the results obtained from the LRA are shown in Table 6. All the results passed the minimum cut-off value of 0.7 and are close to 1.0. It can be seen that the variability of each evaluator's data was decreasing with time (i.e. the results improved from the first comparison to the second, see first two rows in Table 6). By analyzing these results, it can be concluded that variability over time was not a problem in 2008. However, when the results from the IRA are included, the conclusion is that, in either case of a pass or a no-pass, there is a trend in the data analyzed in both 2007 and 2008 that shows that, even when the inspections of each evaluator were being less variable with time, the evaluators' judgment were becoming more different to the others' as data collection progressed. These conclusions support the idea of the need for continuous efforts to improve and maintain data quality of visual asset inspections.

Table 6. 2008 COT season results.

Assessment	*b*	R^2
First Round vs Second Round	0.921	0.829
Second Round vs Third Round	0.945	0.902
First Round vs Third Round	0.901	0.831

5.2.2. 2009 Northern New Mexico Pavement Evaluation Program Results and Analysis

5.2.2.1. First Assessment Round

ABE Assessment

In order to proceed with the implementation of the DQAIF, the data collected were subjected to the ABE assessment, as described in Chapters 3 and 4. The data consist of pavement ratings that range from 0 to 3, assigned to the extent and severity of 8 different pavement distresses, on 24 different pavement samples, by 10 evaluators

In order to maintain consistency of the concepts during the IRA analysis, it was decided by the author to separate each distress and separate their severities from their extents. This way, the researcher created 16 IRA analysis spreadsheets. Each pavement sample was defined as an item for the IRA analysis. There was no need to manipulate the data to fit the criteria needed to implement the DQAIF, since the ratings consist of discrete numbers. In addition, the number of alternatives (c) was set to 4, since the range used by the program extends from 0 to 3, for each rating.

In order to conduct the analysis with the AD_{Md} indexes, the cut-off value has been defined as $c/6 = 0.67$, as discussed in the previous chapter. This value represents the upper limit of consensus in order to determine whether the results are satisfactory or not.

Figure 47 shows the results of the multi-item AD_{Md} indexes for the first round of assessments. From the graph, it can be seen that all the distresses passed the AD_{Md} tests. Nevertheless, one particular distress caught the attention of the author. In the tests, bleeding severity was the distress that obtained the worst results, by a considerable difference respect to the other distresses (see column A in Figure 47). Therefore, with the

purpose of testing the ability of the DQAIF to improve the quality of the data collected, the author decided to pick bleeding severity for intervention. To support this decision, the author performed a rating frequency analysis over bleeding severity, in order to learn whether the high variability observed was due to issues with only part of the evaluators panel, or if there is an issue throughout this panel to assess bleeding severity.

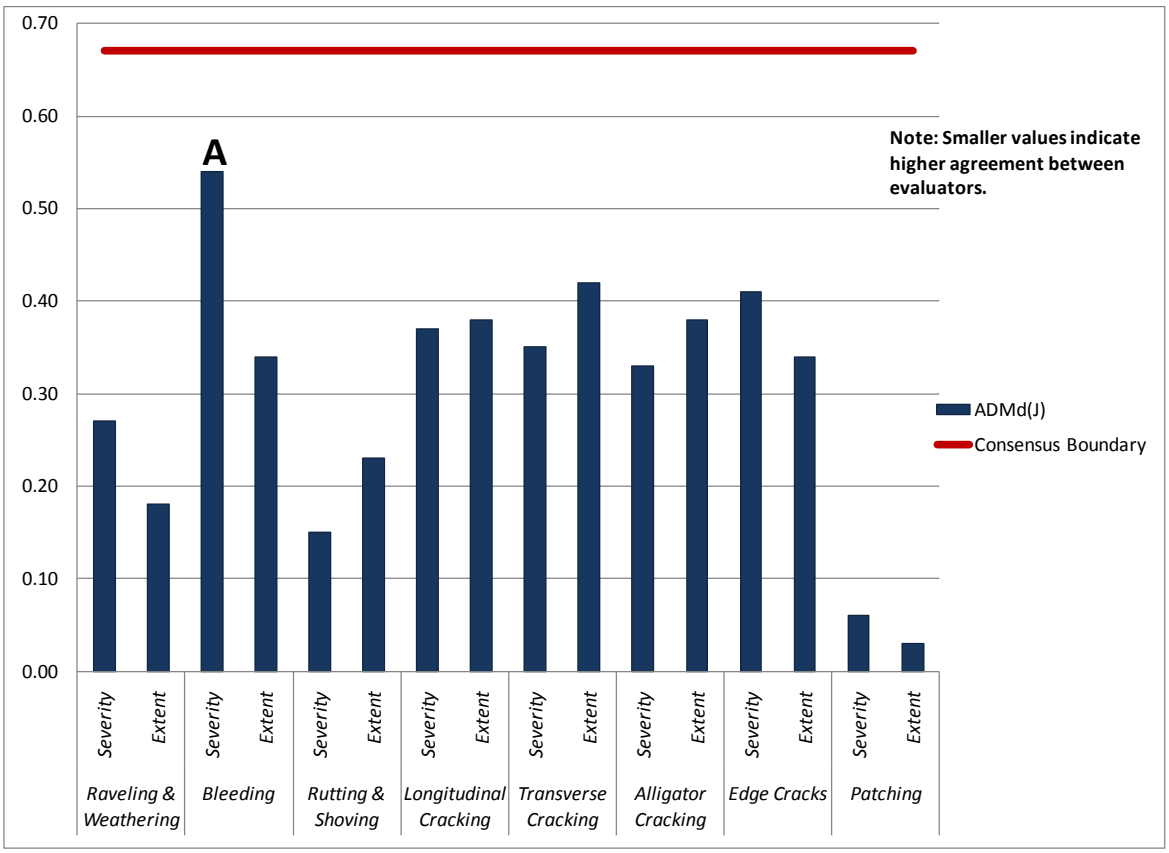


Figure 47. IRA analysis results for the first assessment round.

Figure 48 shows the frequency histograms for bleeding severity, for all the 24 pavement samples that were assessed. As it can be noted from this figure, there was a considerable discrepancy of the ratings among the evaluators. For instance, only 13 out the 24 evaluations presented a rating that was repeated by more than half the panel (i.e. five out of the ten evaluators), and in almost all this cases bleeding seemed to not be present in the

sample (implied from the strong consensus on giving a zero value to this distress). Moreover, most of the evaluations even presented ratings that differ by two or more (i.e. some evaluators rated the bleeding severity as a level 1 when others did as a level 3). The author concluded that bleeding severity results pertain to the entire panel of evaluators and, thus, it can be picked for further intervention, in order to test the DQAIF.

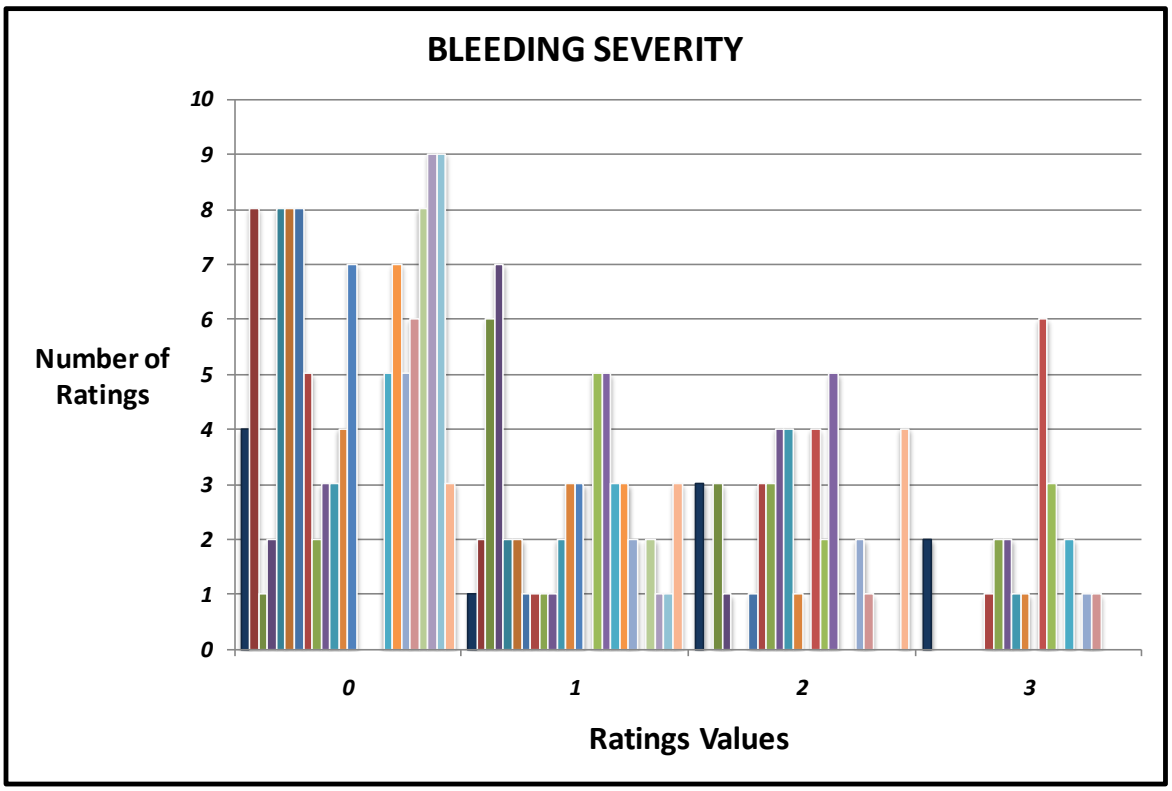


Figure 48. Ratings frequencies histograms of bleeding severity in four items.

Improvement Assessment

In this phase, since the only data available comes from a single assessment round, the conclusions did not differ from the ABE assessment. Moreover, a review of rating descriptions and protocols for the distress in question has shown that the evaluation of bleeding severity is merely based on qualitative characteristics of the distress, and the

language used is ambiguous by making use of adjectives that can be interpreted differently by each evaluator. This led the author to determine that an additional training session with the evaluators could be a good measure of intervention, and that this session should start by obtaining feedback from the evaluators regarding the judgment used to rate this distress and ending with establishing a common criteria for the entire panel, based on both the results shown here and the evaluators' feedback. A note was created for the following assessment round to observe the progress of this particular distress.

5.2.2.2. Second Assessment Round

ABE Assessment

The same considerations were followed in the second assessment round regarding the use and the definition of the different variables involved in the IRA analyses. Data were collected four weeks after the first assessment round from the same evaluators assessing the same pavement samples. Figure 49 depicts the results of $AD_{MD(J)}$ indexes for the second assessment round. The results show, as in the first assessment round, that none of the distresses fell above the cut-off value for this index (in the case of this study, this value was 0.67 for four alternatives). However, even when all the distresses passed the IRA test, the author found of particular interest the cases of rutting & shoving, longitudinal, and transverse cracking extents; as well as both the severity and extent of edge cracks (Columns A to E in Figure 49). These five distresses presented the highest $AD_{Md(J)}$ values, which translates into higher variance among evaluators. These results show a trend of disagreement among the evaluators in rating cracks and ruts, which not only represents a great proportion of all the distresses evaluated, but they are also considered as the most important to the NMDOT, according to the weight factors table in

Chapter 3 (Table 2). Thus, all of these distresses were considered by the author to require further analysis through the use of frequency analysis.

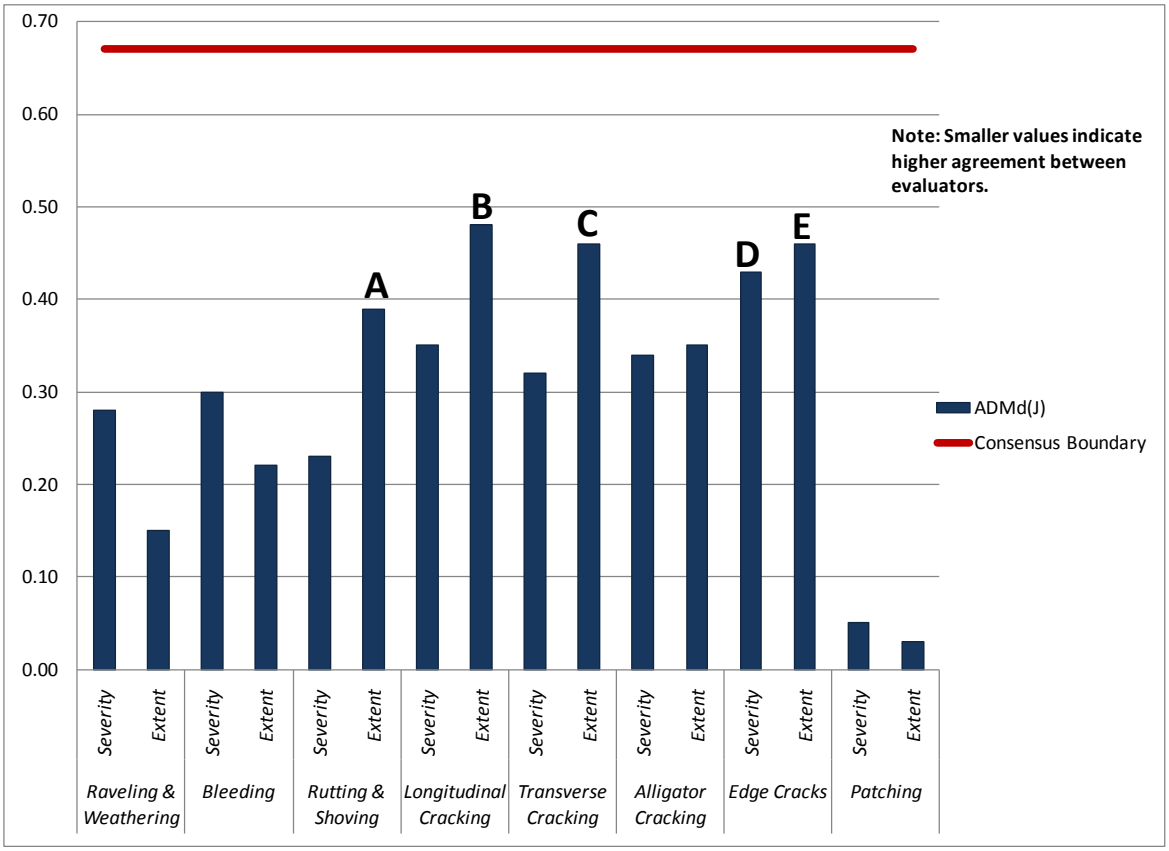


Figure 49. IRA analysis results for the second assessment round.

Figure 50 contains the rating frequency histograms for rutting & shoving, and longitudinal, and transverse cracks extents, and both edge cracks severity and extent. In all these cases, only around half or two thirds of the samples were rated with the same value by six or more evaluators. Thus, the author concluded that the IRA results pertain to the entire panel of evaluators, and that these five distresses should be marked to be considered in the improvement assessment.

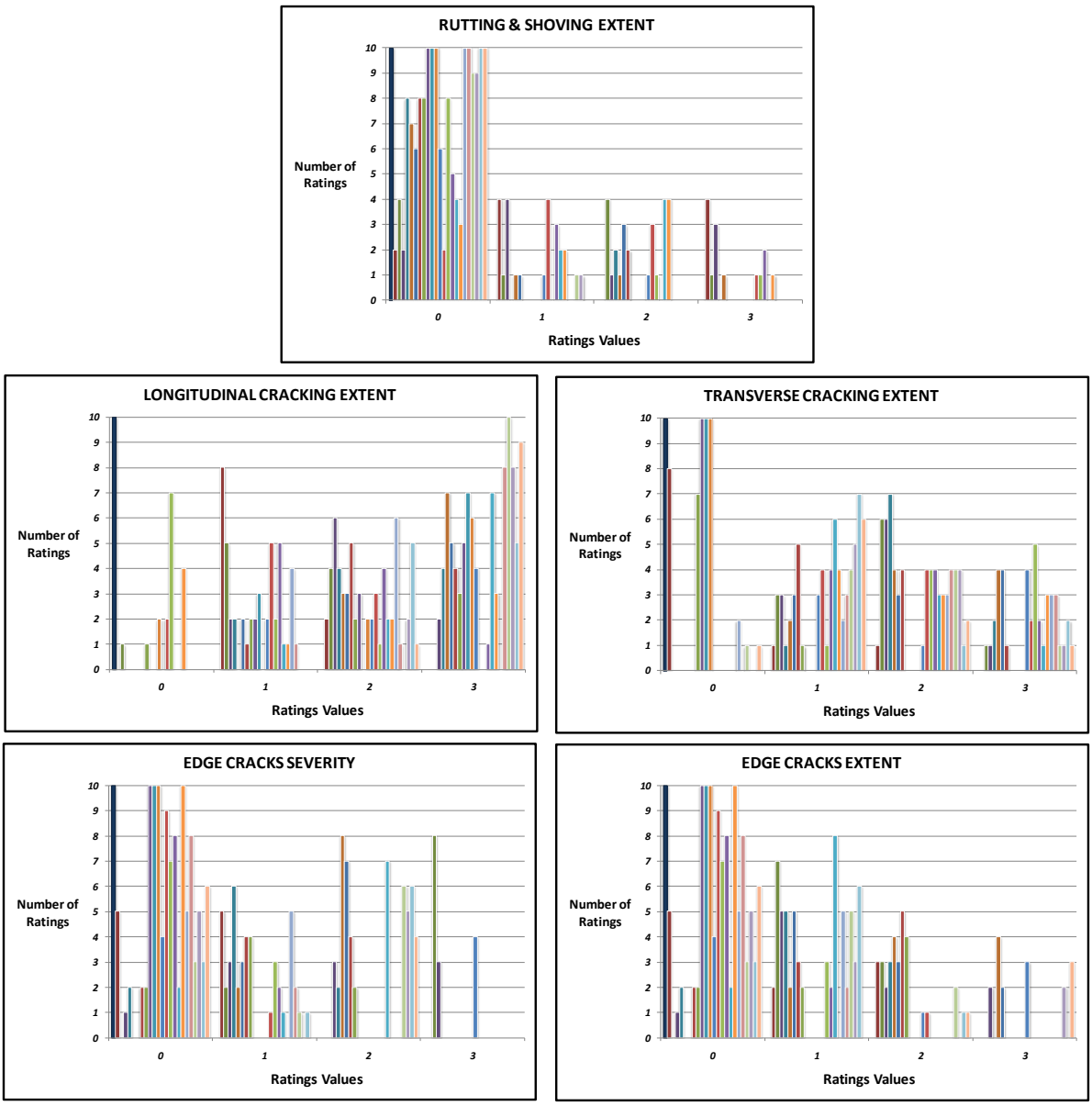


Figure 50. Ratings frequencies histograms of the five different distresses.

COT Assessment

According to Figure 29, a COT assessment should not be performed since the intervention conducted at the end of the last assessment round would affect the results obtained in this step. In other words, since bleeding severity was addressed in the first assessment round, the results of the LRA may not pass due to the changes on bleeding severity which high variance between rounds could represent a good change since this distress was intervened during the last round.

However, a COT was still performed in this study for demonstrational purposes. A LRA was performed on the DRs computed over ratings of each of the 10 evaluators, at two different times of assessment. The results of the analysis are shown in Figure 51 (next page), where *b* is the slope of the fitting line, and R^2 is the coefficient of determination. Roughly speaking, the value of *b* fell below 1.0, which means that the fitting line has an inclination of less than 45°, which is the inclination expected in a situation of perfect consistency throughout time. This suggests that, at an overall level, the ratings during the first assessment were higher than those of the second assessment. The R^2 value represents the proportion of the total rating variance that can be explained by the regression model .In this case, since the model was set to intercept the origin, it is represented by *b*. Since the values of the coefficient of determination fell below 1.0, the pairs of ratings were relatively scattered.

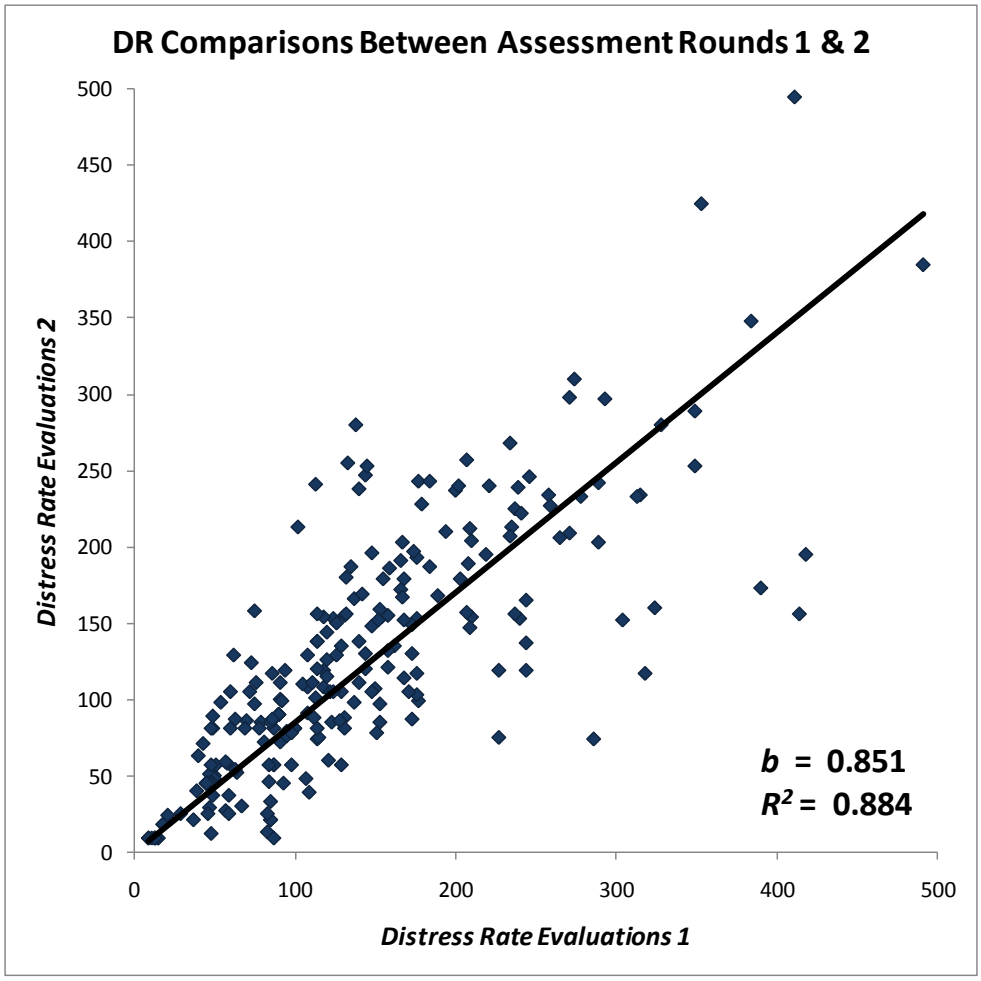


Figure 51. Scatter plots comparing the ratings at different times of assessment.

Since both values of *b* and R^2 fell within acceptable ranges, according to the criteria explained in Chapter 4, the author concludes that in general, the variability throughout time did not represent a problem between the first and the second assessment rounds. However, it has to be remembered that the concept of the DR is more sensitive to the ratings of some distresses (e.g. rutting & shoving, longitudinal, transverse, and alligator cracking) than others (e.g. raveling & weathering, bleeding, edge cracks, and patching). Thus, further analysis of the distresses that not affect the DR values as much is a safe practice to assure there are no problems with these distresses. Moreover, ratings

frequency histogram analyses were conducted on all distresses, in order to show their use within the DQAIF process.

The ratings frequency histogram analysis consists on the development of a histogram graph showing the counts of all the rating combinations possible between the two evaluation rounds under assessment (e.g. a level 1 rating on the first round and a level 3 rating on the second round, etc.). Figure 52 (next page) shows an ideal situation for a distress under this analysis. The depth axis represents the number assigned in the rating on the first assessment, while the numbers assigned during the second assessment are represented in the horizontal axis. The vertical axis represents the count of the combinations present in the ratings. The summation of all the counts (columns) represent the total number of evaluations (i.e. samples times evaluators), so maximum count a column may have is the total number of evaluations which, in turn, would leave all the other columns empty. Thus, reading column A in Figure would be: "In 51 times, when an evaluator assigned a value of zero to the distress under analysis during the first assessment round, this evaluator assigned a value of zero to this distress during the second assessment round". Then, the situation depicted in Figure 52 is ideal because the large counts in columns A to D denote a strong tendency of the evaluators to assign the same value to the distress under assessment during both assessment rounds, which is a perfect COT.

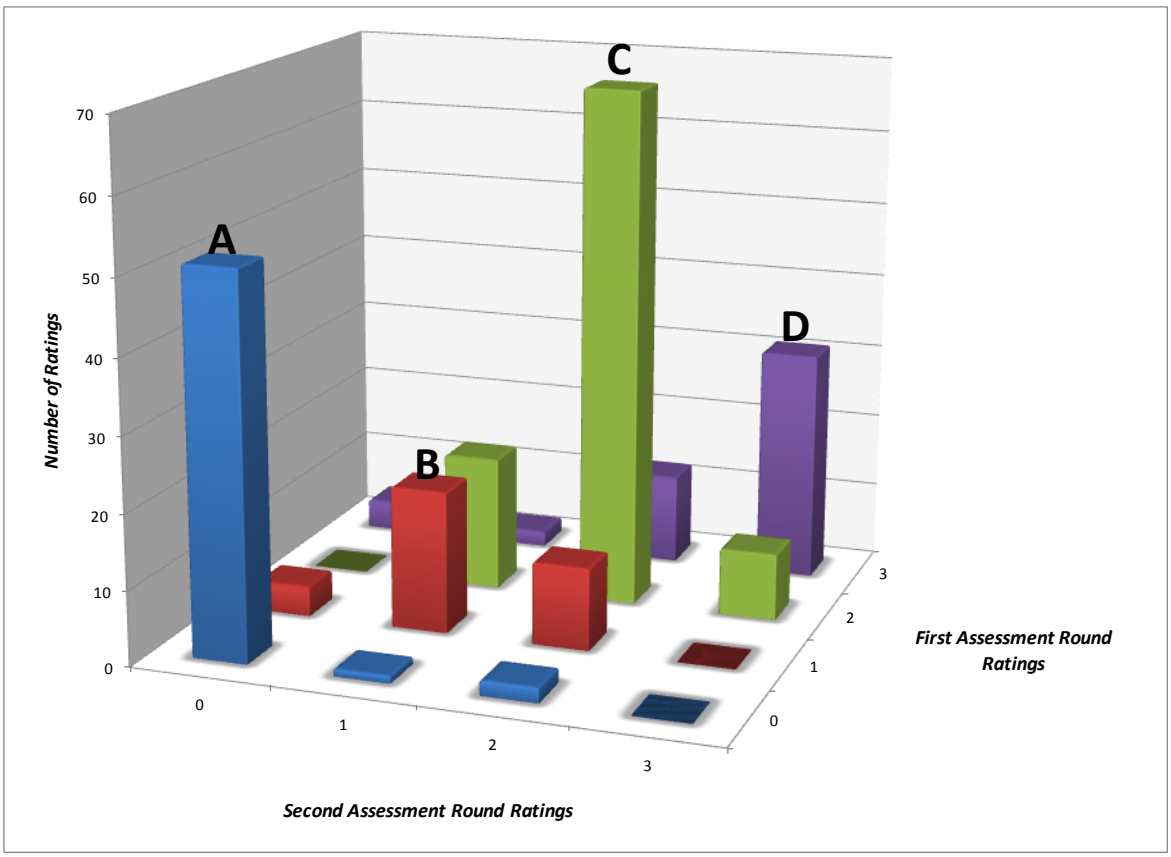
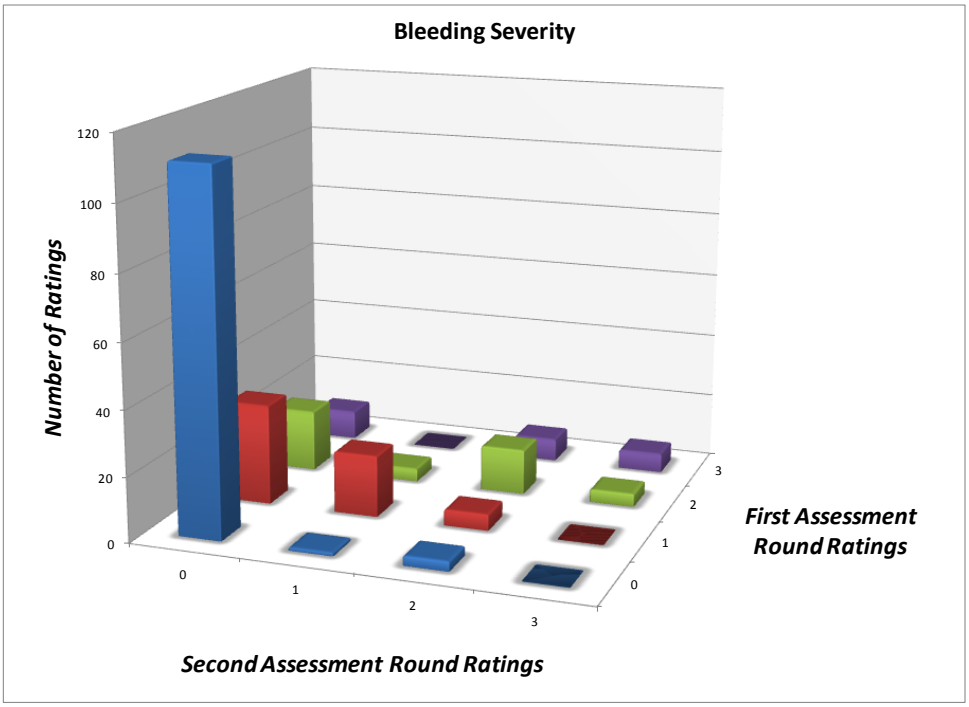


Figure 52. Example of a rating count histogram.

Figures 53 and 54 (next two pages) show the results obtained for bleeding severity and extent, and edge cracks extent, respectively. The graphs in both figures show that the evaluators were not consistent with their evaluations at different times, since the distributions of the columns in the graphs are more spread than those in Figure 52. Both bleeding and edge cracks are two of the distresses that do not affect the values of the DRs as much as most other distresses, and this is why the overall results in the LRA are positive, in spite of these distresses. Thus, the author concluded that bleeding severity and extent, and edge cracks extent vary considerably between rounds, and that these should be considered during the improvement analysis.



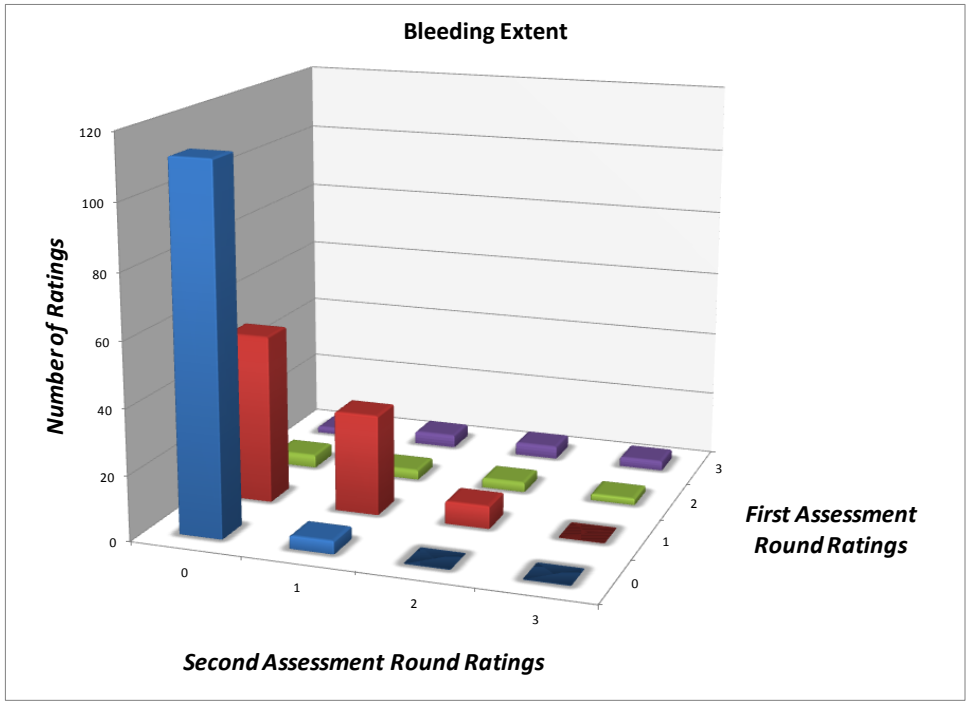


Figure 53. Ratings count histograms for bleeding severity and extent.

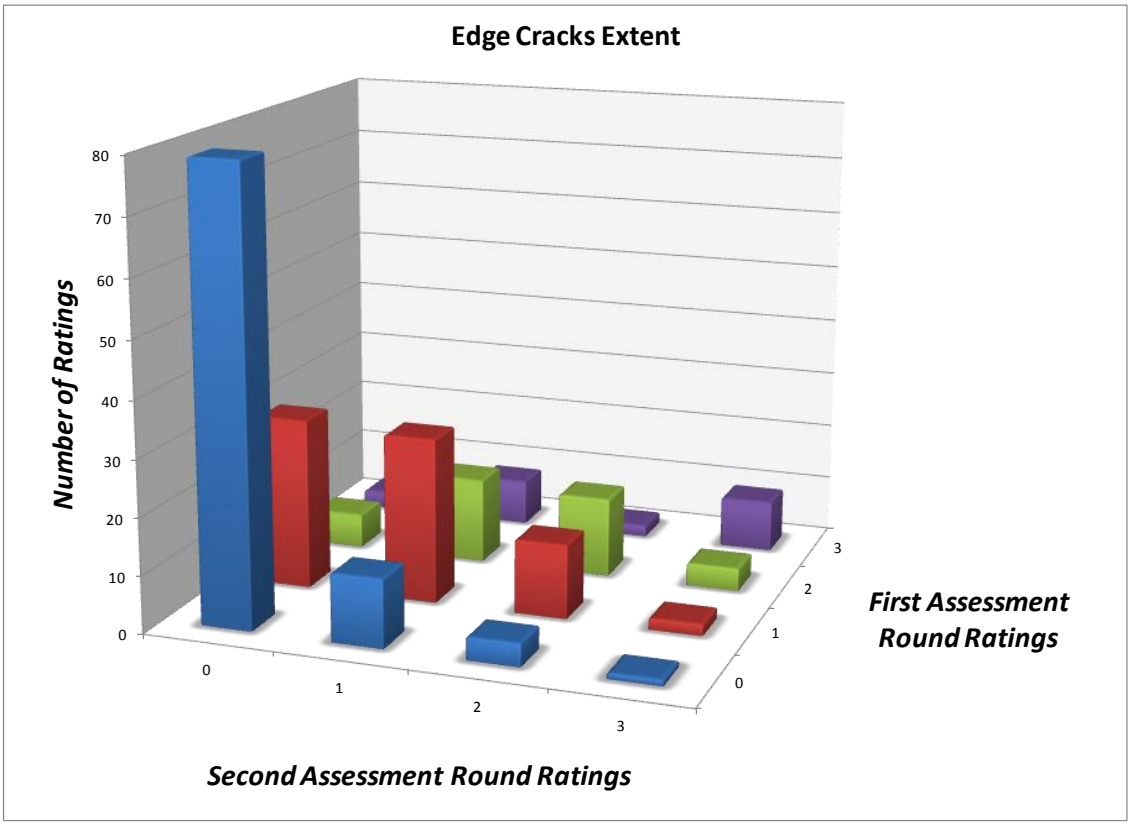


Figure 54. Ratings count histogram for edge cracks extent.

Improvement Assessment

After the completion of both the ABE and the COT assessments, the improvement assessment was performed in order to: 1) assess the overall status of the data quality; and 2) determine how this can be improved. The same data used in the ABE analyses were the focus of this assessment, as well as the results from the COT assessments.

Figure 55 (next page) shows the results for the analyses with the $AD_{Md(J)}$ index. The radar graphs show, separately, the results of the analyses of the distresses severities and extents. With these graphs, comparisons were made of the two assessment rounds against the upper cut-off value of consensus (consensus boundary in the radar graphs of Figure 55), and between each round.

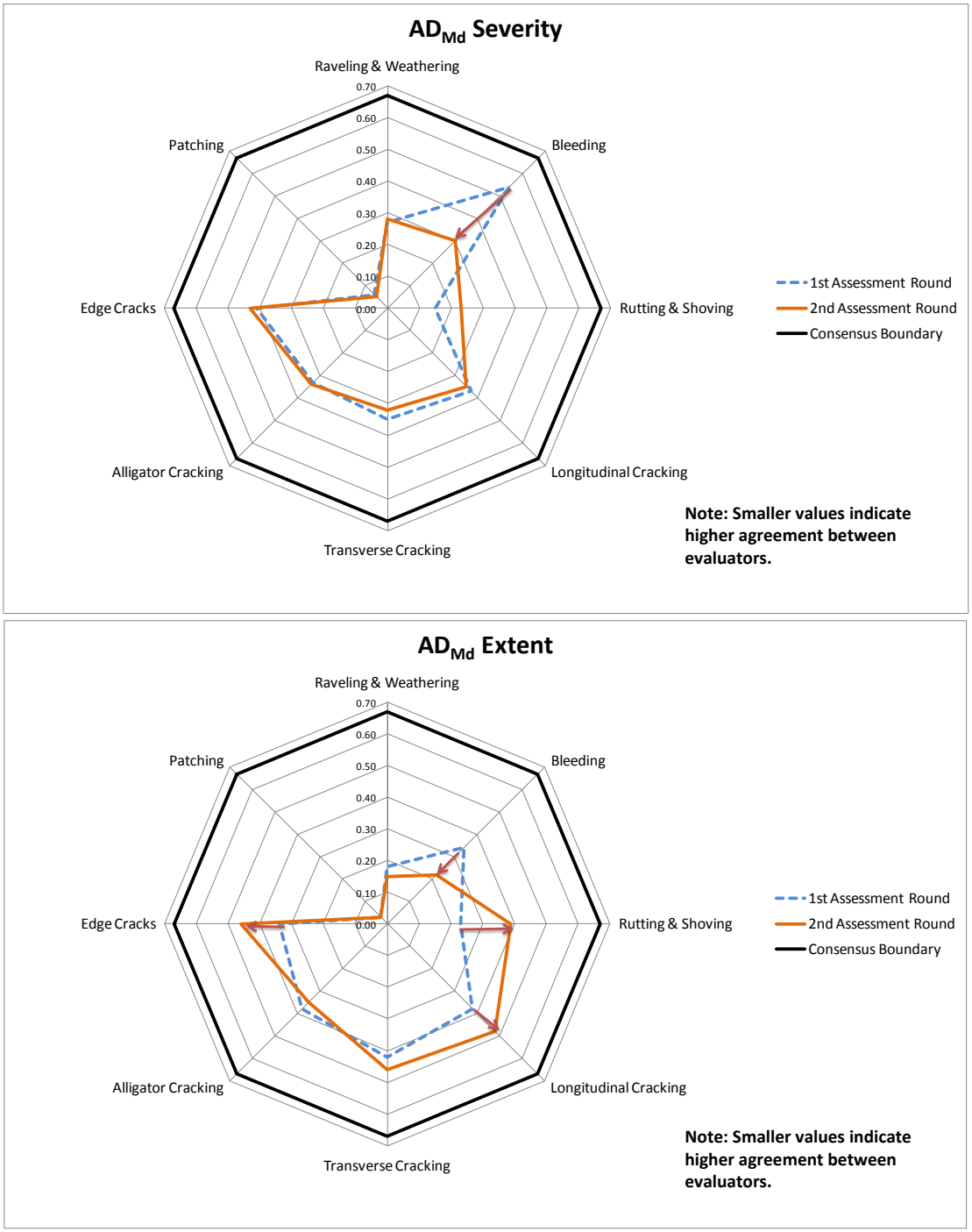


Figure 55. Radar graph of the $AD_{Md(J)}$ results from the first and second assessment rounds.

First, a comparison of the results against limits of consensus was performed. The analysis shows that, in general, the degree of variability of the data is within acceptable ranges, since all the distresses passed the $AD_{Md(J)}$ cut-off value (0.67 for $AD_{Md(J)}$ with four rating alternatives -0 to 3). However, looking back at each single distress, some concerns were placed upon bleeding severity during the first assessment round; and over rutting & shoving, and longitudinal and transverse cracks extents, and both edge cracks severity and extent at the second assessment round. It can be interpreted from this that even when the overall protocol is clear enough to have consistency among evaluators, there are still a few situations that are not clear enough for evaluators to assign the same values. An analysis of these special situations could be helpful in identifying the cause of the disagreement, in order to improve the training protocols, or even that particular distress description. For this reason, it is concluded that no Problem Roots Analysis will be performed, in addition to Improvement Opportunities Analysis, after the completion of this stage of analysis.

After comparing against the cut-off values, an analysis was made to compare the results between both assessment rounds. At a global level, the distress severities did not suffer significant changes between both rounds, except for bleeding severity, which showed a considerable improvement from the first to the second round. This overall trend suggests that the concepts have been stabilized among the panel of evaluators.

The ability of the DQAIF to influence improvements on the data quality of manual asset evaluations can be proven with the results shown in bleeding severity and extent. An itemized comparison among rounds of these two items reveal that, after the control measure implemented at the end of the first assessment, the results of both IRA tests were

significantly favorable on the second assessment round, especially when these are compared with the first round's. The dramatic change in these two distresses led bleeding to move from being a data quality concern to be at the same level of quality as most distresses.

However, the distress extents results show that the variability, in general, has gotten worse on the second round. Since the extents of the distresses assessed in this program are not a qualitative characteristic (i.e. they are objectively measurable), it can be interpreted that the evaluators are spending less time and/or effort in assessing the distresses extents (particularly the extents of different types of cracks). In other words, it could be explained as a source of the increase in the evaluators ratings' variance the development of a sense of comfortability as the evaluators gained experience. Particularly of concern were the progress shown in the evaluation of longitudinal and edge cracks extents, were both IRA indexes showed increased disagreement.

Of particular concern were the results obtained for both severity and extent of rutting & shoving, which showed more variability with their $AD_{Md(J)}$ values. The concern relates to the fact that neither the severity nor the extent of this distress were measured using qualitative properties. For instance, the severity is a function of the rut depth, which is measure with the help of a manual device; while the extent is measured as a proportion of the entire length of the sample. However, the reason for this result could be found in another fact: in order to measure ruts, the evaluator has to get inside the pavement lane under assessment, without interrupting the traffic passing along the sample. This is a significant implication, because not only is this procedure the most complicated to perform, but it is also the most hazardous procedure the evaluator has to perform during

an assessment. It can be interpreted from the above information that the evaluators were performing fewer rut checks due to one or both of the following: 1) the evaluators felt reluctant to perform as many checks as needed due to the effort it entails; or 2) the evaluators deterred from performing as many checks as needed by the protocol due to safety concerns –i.e. they did not feel safe performing these checks, or the traffic did not allow them to do it.

Then, an improvement opportunities analysis was performed to suggest potential improvement needs. As demonstrated in Figure 55, and in overall with the previous assessments, there are no distresses that need immediate control from the program manager. However, there are two elements that should be addressed before they become significant issues:

a) Cracking extents; and

b) Rutting & shoving assessments

It is of importance to address these two issues, not only because the results suggest that the progress of these are significantly negative; but also because all these distresses affect the calculations the NMDOT performs in order to assess the overall conditions of their pavement network. With this, it has been proven that the DQAIF can be used not only to monitor and control data quality of manual asset evaluations, but it has been proven to be useful to predict future issues before they become an actual problem.

5.3. Discussion & Recommendations –Adapted from Bogus, Migliaccio, and Cordova (2010a)

There are different aspects that may arise regarding the implementation of the proposed system because of its use of concepts and procedures that might not be familiar to the asset management environment. The protocol followed by NMDOT is appropriate for the implementation of the proposed framework. Nevertheless, the elements that any other asset management system would need to include for the adequate implementation of DQAIF are few. The most important is the use of manual or visual procedures for data collection of the asset conditions. The other condition that has to be met is the conversion of the data collected into a scale of positive discrete values. NMDOT uses a 4-point scale ranging from 0 to 3, but any other range of positive integer values that include a different number of ratings within the range can be used (i.e. five- or seven-point scale ratings). The same applies to the size of the panel of evaluators. In the case study, ten evaluators were used because this number allowed the team to complete the assessment of the whole Northern New Mexico road network within the desired time frame. While the author does not suggest a particular size for the implementation of the DQAIF, a consideration to take into account is that a greater number of evaluators will introduce more sources of variability. Nevertheless, the author recommends that, instead, this decision be based on the evaluation program constraints of time and budget. Regarding the amount and the roads to be selected as part of the quality assessment plan, the sample selected should be representative of the network while meeting the program's constraints. For instance, the asset manager can identify a set of roads that encompass the overall characteristics that

are expected to be found in the network, and that also can be travelled as part of a single path, so the quality assessment inspections can be optimized.

The precision that should be expected from this system has to be decided by the asset manager, considering the efforts and actions that have to be performed in order to achieve and maintain a desired level of precision, as well as time and budget availability. Another aspect that should be considered is the purpose that the data obtained in manual surveys will serve. The degree of precision required for project-level management is not the same as for network –level management, which has higher tolerances. The precision of the DQAIF can be refined between quality assessment rounds. Data quality issues found by the implementation of DQAIF can be addressed by additional training of the evaluators on the processes and protocols followed by the inspection program. The effect of the measures taken can be assessed during subsequent evaluation rounds. This way, the DQAIF not only can serve the assessment of the variability of manual condition surveys, but it also represents a tool for controlling this variability.

CHAPTER 6. CONCLUSIONS

6.1. Summary of Study

This study aimed to develop a process to monitor and control the variability of data collected manually or visually on asset evaluations, where the judgment of the evaluators has been regarded as a source of variability that has not been controlled until now. The Data Quality Assessment & Improvement Framework is a system that follows a continuous quality improvement approach, by constantly assessing the quality of the data collected in manual asset evaluations, through statistical process control, in order to find opportunities for data quality improvement.

The DQAIF was then implemented in the Northern New Mexico Pavement Evaluation Program in order to test its capacity to monitor and control the variability of the data collected by ten pavement evaluators at two different assessment times. The results of the study prove that the systematic conduction of IRA and LRA is not only capable of identifying data quality concerns in the short term, but it also provides the means to identify trends of potential risk for the asset data collection programs before these become an actual problem. In addition, the DQAIF was proven to be useful to justify decisions made regarding the implementation of procedures, tools, and protocols within an asset data collection program, with respect to their influence in data variability. The DQAIF was proven to be useful as a quality control tool within an asset evaluation program.

6.2. Research Questions Rationale & Findings

A review of the concepts discussed in Chapter 3 led to the definition of the research questions. The questions formulated in Chapter 3 asked concerns about different aspects of the matter of variability of visual asset evaluations. These questions were:

a) *How can variability of visual asset condition assessments, due to the evaluators' subjectivity, be reduced?*

b) *Can statistical analysis be used to assess subjectivity variance?*

c) *What alternative can be used to identify variability among evaluators?*

d) *What alternative can be used to identify variability throughout time?*

The development of the answers for these questions led to the development of a framework based on the concepts of continuous improvement and statistical process control. The framework developed (DQAIF) was then assessed based on the premises:

a) The capacity of the continuous improvement approach to reduce the variability of manual asset condition assessments; and

b) The capacity of the statistical process control procedures to support a continuous improvement approach to reduce the variability of manual asset condition assessments.

The challenge of answering the stated questions resides in the fact that, in order to achieve positive answers, both concepts should pass the assessment of the study. In other words, the failure of the DQAIF could be due not to the failure of the entire system, but

to the failure of either one of these two concepts. Thus, a close analysis of the results of the study was of significant importance in order to reach the right conclusions.

Reasoning the answer to the research question, it was observed that:

a) In a global sense, the variability of the data collected during the study, in terms of the two dimensions of judgment established at the beginning of the study, was within pre-established standards. However, at a local level, it could be observed that focalized concerns were present since the beginning of the assessment (e.g. bleeding severity in the first assessment round). Additionally, at the end of the assessment, it was observed the presence of trends that could compromise the successfulness of the case studied in the long term. These observations were possible through the appropriate implementation of a set of statistical process control procedures. Thus, the author concludes that this element complied with its function within the system under assessment.

b) Two different trends were present among the distresses under assessment in the study: a) for that under the action of control measures, it was observed a strong tendency in the reduction of its data variability –significantly better than the rest of the distresses; and b) for those that were not under the action of control measures, it could be observed that, not only they did not reduced their data variability, but –in most cases- these got worse. Thus, it is concluded by the author that this element complied with its function within the system under assessment.

Then, since both elements of the system under assessment complied with their functions, it was concluded by the author that the DQAIF complies with its objective in reducing the data variability of manual assets condition assessments.

6.3. Opportunities for Future Research

The development and testing of the DQAIF experience provided the author with ideas for continuing research related to the use and implementation of the procedures that are part of this system. For instance, the development of a software package or a tool that encompasses all the processes and procedures that form part of the DQAIF could represent an important contribution in order to facilitate the implementation of the DQAIF in the industry.

Another potential continuation of the present study could be the assessment of the bias evaluators present when performing visual asset evaluations. Particularly, comparing the ratings performed by trained and untrained evaluators could be of use.

Additionally, it was observed by the author the need to define the limits of consensus in order to these to be meaningful to the particular task of assessing current asset conditions. For instance, a study could be conducted in order to determine what cut-off value should be used when estimating each of the IRA measures used in the DQAIF. Finally, a potential topic for a study could be to test the usefulness of the DQAIF to maintain acceptable levels of data variability when the panel of evaluators suffers changes.

REFERENCES

American Association of State Highway and Transportation Officials. (1990). *Guidelines for Pavement Management Systems*. Washington, DC.

American Association of State Highway and Transportation Officials. (2001). *Pavement Management Guide* (p. 254). American Association of State Highway and Transportation Officials.

Bliese, P. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. *Multilevel theory, research, and methods in organizations*, 349381, 349–381.

Bogus, S., Migliaccio, G. C., & Cordova, A. (2010a). Data Quality Assessment for Manual Pavement Distress Evaluations. *Transportation Research Record: Journal of the Transportation Research Board, forthcoming*.

Bogus, S., Migliaccio, G. C., & Cordova, A. (2010b). Performance of Manual Condition Surveys Using Inter-Rater Agreement Measurements. *In Proceedings of the 2010 CIB World Congress (CD)* (pp. 10-13). Salford Quays, United Kingdom: CIB.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods*, 2(1), 49.

Burke, M., & Dunlap, W. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods*, 5(2), 159. Res Methods Div.

Carey Jr, W. N., & Irick, P. E. (1960). The Pavement Serviceability-Performance Concept. *HRB Bulletin*, 250, 40-58.

Carmines, E., & Zeller, R. (1979). *Reliability and validity assessment* (10 ed., p. 71). Sage Publications, Inc.

Clough, P., Duncan, I., Steel, D., Smith, J., & Yeabsley, J. (2004). Sustainable infrastructure : A policy framework Report to the Ministry of Economic Development.

Cronbach, L. (1990). *Essentials of Psychological Testing*. New York (5th.). New York: Harper-Row.

Crosby, P. B. (1979). *Quality is free: The art of making quality certain* (p. 270). New York: New American Library.

Epps, J., & Monismith, C. (1986). *Equipment for Obtaining Pavement Condition and Traffic Loading Data, NCHRP Synthesis of Highway Practice 126* (p. 118). Washington, DC: Transportation Research Board.

Federal Highway Administration. (2008). HPMS Reassessment 2010+ Final Report. Washington, DC.

Federal Highway Administration. (1997). Asset Management: Advancing the State of the Art Into the 21st Century Through Public-Private Dialogue.

Flintsch, G. W., & McGhee, K. K. (2009). *Quality Management of Pavement Condition Data Collection, NCHRP Synthesis of Highway Practice 401* (Vol. 8, p. 152). Washington, DC: Transportation Research Board.

Gramling, W. (1994). *Current practices in determining pavement condition, NCHRP Synthesis of Pavement Practice 203* (p. 63). Washington D.C.: Transportation Research Board.

Haas, R. C., & Hutchinson, B. G. (1970). A Management System for Highway Pavements. *Proc., Australian Road Res. Board.*

Haas, R., Hudson, W. R., & Zaniewski, J. (1994). *Modern Pavement Management*. Malabar, Florida: Krieger Publishing.

Hadfield, C. (1986). *World canals: inland navigation past and present*. London: David & Charles.

Hudson, W., & Haas, R. (1991). Research and Innovation Toward Standardized Pavement Management. In F. B. Holt & W. L. Gramling, *Pavement Management Implementation*. Philadelphia: American Society for Testing and Materials.

Hudson, W., Finn, F., McCullough, B., & Nalr, K. (1968). Systems Approach to Pavement Design. Systems Formulation. Performance Definition and Materials Characterization, Final Report, NCHRP Project 1-10. *Materials Research and Development, Inc*.

Hutchinson, B., & Haas, R. (1968). A systems analysis of the highway pavement design process. *Highway Research Record*, (239), 1-24.

ISO 9000:2000. (2000). *Quality Management Systems— Requirements*. Geneva, Switzerland.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of applied psychology*, 69(1), 85-98.

James, L. R., Demaree, R. G., & Wolf, G. (1993). r-sub (wg): An assessment of within-group interrater agreement. *Journal of Applied Psychology*, 78(2), 306–309.

Kaplan, R., & Saccuzzo, D. (1993). *Psychological Testing: Principles, Applications, and Issues. Applications and Issues, 3rd ed*. Brooks/Cole. Pacific Grove (3rd., p. 715). Cengage Learning.

Kozlowski, S. W., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology*, 77(2), 161–167.

Lay, M. G., & Vance, J. E. (1992). *Ways of the World: A History of the World's Roads and of the Vehicles that Used Them* (p. 424). New Brunswick, New Jersey: Rutgers University Press.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11(4), 815.

Lebreton, J. M., Burgess, J. R., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, 6(1), 80.

Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement*, 23(2), 127.

Lindell, M., & Brandt, C. (1997). Measuring interrater agreement for ratings of a single target. *Applied Psychological Measurement*, 21(3), 271-278.

Lindell, M., & Brandt, C. J. (2000). Climate quality and climate consensus as mediators of the relationship between organizational antecedents and outcomes. *Journal of Applied Psychology*, 85(3), 331–348.

Lytton, R., Rauhut, J., & Darter, M. (1985). *Long Term Pavement Monitoring Data Collection Guide*. Washington, D.C.: U.S. Department of Transportation, Federal Highway Administration.

McGhee, K. H. (2004). *Automated Pavement Distress Collection Techniques, NCHRP Synthesis of Highway Practice 334* (p. 84). Washington, DC: Transportation Research Board.

McPherson, K., & Bennett, C. (2005). Success Factors for Road Management Systems. Washington, DC.

Mifflin, H. (2000). *The American heritage dictionary of the English language*. (A. H. Company)Boston: Author (Fourth., p. 951). Houghton Mifflin Company.

Miller, J. S., & Bellinger, W. Y. (2003). *Distress Identification Manual for the Long-Term Pavement Performance Program (FHWA-RD-03-031).* Federal Highway Administration (4th revised.). Federal Highway Administration.

Miller, J., Rada, G., & Rogers, R. (1993). *Distress Identification Manual for the Long-Term Pavement Performance Project*. (SHRP) (3 ed., p. 147). Washington, DC: Strategic Highway Research Program, National Research Council.

Morian, D., Stoffels, S., & Frith, D. (2003). Quality Management of Pavement Performance Data. In I. Al-Qadi & T. Clarck, *Pavement Evaluation 2002*. Roanoke, Va.

Moteff, J., & Parfomak, P. (2004). Critical infrastructure and key assets: definition and identification. *Time*.

Needham, J., Ling, W., Gwei-Djen, L., Wang, L., & Lu, G. D. (1971). *Science and Civilisation in China: Physics and Physical Technology: Civil Engineering and Nautics*. Cambridge University Press.

New Mexico Department of Transportation. (2009). The New Mexico Department of Transportation's Pavement Evaluation Inspection Program: A Successful Partnership. Santa Fe, NM.

New Mexico Department of Transportation. (2004). The New Mexico Department of Transportation's Network Level Pavement Management.

Nunnally, J. (1978). *Psychometric theory*. NY: McGraw-Hill Inc. New York: McGraw-Hill.

O'Sullivan, A., & Sheffrin, S. M. (2003). *Economics: Principles in action*. Upper Saddle River, NJ: Prentice Hall.

Papagiannakis, A., Gharaibeh, N., Weissmann, J., & Wimsatt, A. (2009). Pavement Scores Synthesis. College Station, Tx.

Peterson, D. (1987). *Pavement Management Practices, NCHRP Synthesis of Highway Practice 135*. Transportation Research Board, National Research Council, Washington, DC. Washington, DC: Transportation Research Board.

Rada, G. R., Bhandari, R. K., Elkins, G. E., & Bellinger, W. Y. (1997). Assessment of Long-Term Pavement Performance Program Manual Distress Data Variability: Bias and Precision. *Transportation Research Record: Journal of the Transportation Research Board*, 1592(-1), 151–168.

Rodda, J. C., & Ubertini, L. (2004). *The Basis of Civilization–water Science?: Water Science?* (p. 334). IAHS Press.

Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology*, 74(2), 368–370.

Schneider, B., Salvaggio, A. N., & Subirats, M. (2002). Climate strength: A new direction for climate research. *Journal of Applied Psychology*, 87(2), 220–229.

Scrivner, F., Moore, W., Mcfarland, W., & Carey, G. (1968). A Systems Approach to the Flexible Pavement Design Problem. *Research Report*. College Station, Tx.

Shahin, M. Y. (2005). *Pavement Management for Airports, Roads, and Parking Lots*. (2nd., p. 572). Springer.

Shahin, M., & Kohn, S. (1979). *Development of a Pavement Condition Rating Procedure for Roads, Streets, and Parking Lots. Volume II. Distress Identification Manual*. (p. 120). Defense Technical Information Center.

Shahin, M., Kohn, S., & Darter, M. (1977a). *Development of a Pavement Maintenance Management System. Volume I. Airfield Pavement Condition Rating*. (p. 232). Defense Technical Information Center.

Shahin, M., Kohn, S., & Darter, M. (1977b). *Development of a Pavement Maintenance Management System. Volume II. Airfield Pavement Distress Identification Manual*. (p. 115). Defense Technical Information Center.

Smith, K., Darter, M., & Hall, K. (1989). *Distress identification manual for the Long-Term Pavement Performance (LTPP) studies*. Strategic Highway Research Program. Washington, DC: Strategic Highway Research Program.

Smith, R., Darter, M., & Herrin, S. (1979). Highway pavement distress identification manual for highway condition and quality of highway construction survey. Federal Highway Administration/U.S. Department of Transportation.

United States Department of Transportation. (1999). *Asset Management Primer*. Federal Highway Administration (FHWA), Office of Asset Management (p. 31). FHWA, Office of Asset Management.

University of New Mexico. (2009). 2009 Pavement Evaluation Report: Northern New Mexico. Albuquerque, NM.

Wilkins, E. (1968). Outline of a Proposed Management System for the CGRA Pavement Design and Evaluation Committee. In *Proc. Can. Good Roads Association*. Ottawa.

Zaniewski, J., Hudson, S., & Hudson, W. (1985). Pavement Condition Rating Guide. *ARE Inc., Austin, Texas*.

APPENDIX A: ESTIMATION PROCESSES FOR INTER-RATER AGREEMENT ALTERNATIVE MEASURES

James, Demaree, and Wolf (1984) single- and multiple-item r_{WG} -Rearranged from Bogus, Migliaccio, and Cordova (2010a)

for the estimation of the inter-rater agreement over a single item, James et al. (1984) proposed the Formula A:

$$r_{WG(I)} = 1 - \left(\frac{{S_{x_j}}^2}{{\sigma_E}^2}\right) \qquad \text{(A)}$$

Where:

$r_{WG(I)}$ = Within group inter-rater agreement for a group of K evaluators on a single item Xj.

Sx_j^2 = Observed variance on X_j.

σ_E^2 = Variance on X_j that would be expected if all evaluations were due exclusively to random measurement error.

Similarly, for the estimation of the inter-rater agreement over multiple items, James et al. (1984) proposed (Formula B):

$$r_{WG(J)} = \frac{J\left[1 - \left(\frac{\overline{S_{x_J}^2}}{\sigma_E^2}\right)\right]}{J\left[1 - \left(\frac{\overline{S_{x_J}^2}}{\sigma_E^2}\right)\right] + \left(\frac{\overline{S_{x_J}^2}}{\sigma_E^2}\right)}$$

(B)

Where:

$r_{WG(J)}$ = Within group inter-rater agreement for evaluators mean scores based on J parallel items.

$\overline{S_{X_J}^2}$ = Mean of the observed variances on the J items.

J = Number of items.

The procedure to calculate $r_{WG(I)}$ and $r_{WG(J)}$ for one distress extent or severity, with the use of the spreadsheet format in Figure 31, is the following:

1) *Collection of the data that will be subjected to analysis (x_{kj}):* The data should be organized by evaluator and by item. In the spreadsheet, the rows represent the data collected by the same evaluator (k), and the columns represent the data collected in each sample (j), or items subjected to analysis (Figure A). It is worth to mention, though, that in the case of missed data, the cells related to that data should be left blank in order to not affect the results of the analysis.

Evaluators	Items				
	1	2	3	...	J
1	X_{11}	X_{12}	X_{13}	...	X_{1j}
2	X_{21}	X_{22}	X_{23}	...	X_{2j}
3	X_{31}	X_{32}	X_{33}	...	X_{3j}
. . .	. . .	. . .	. . .	... ...	. . .
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}

Figure A. Spreadsheet format of x_{kj}

2) *Calculation of the Mean Rate value($\bar{x}_j$)* : These are the average of the rate values for each item (i.e. the mean value of the ratings given by all the evaluators to a sample). It is calculated by the following Formula C. Figure B is an expansion of Figure A, and it shows where in the IRA spreadsheet these values should be computed.

$$\bar{x}_j = \frac{\sum_{k=1}^{K} x_{kj}}{K} \qquad (C)$$

Where:

K = Number of evaluators.

Evaluators	Items				
	1	2	3	...	J
1	X_{11}	X_{12}	X_{13}	...	X_{1j}
2	X_{21}	X_{22}	X_{23}	...	X_{2j}
3	X_{31}	X_{32}	X_{33}	...	X_{3j}
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
K	X_{k1}	X_{k2}	X_{k3}	...	X_{ki}
# Evaluators	K_1	K_2	K_3	K_4	K_5
Mean	X_1	X_2	X_3	X_4	X_5

Figure B. Computation of the mean rate value in the IRA spreadsheet.

As can be noted in Figure B, the spreadsheet leaves opened the option of assigning a different K for each item. This does not necessarily means that the size of the panel of evaluators can be changed constantly, but only that in the case not all the data can be collected from the entire panel. It is desirable, however, that the size remains as constant as possible.

3) *Estimation of the Expected Variance (σ_E^2, Figure C):* A probability distribution of the expected variance has to be assumed in order to compute this value. James et al. (1984) suggest formulas and values for σ_E^2 with uniform and triangular distributions, as well as skewed distributions with different degrees of skew. All of these are a function of the number of alternatives (A) an evaluator has to give a rate (i.e. rating scale). This value represents the expected variance that would be present if evaluations were performed by untrained people.

Evaluators	Items					...	σ_E^2
	1	2	3	...	J		
1	X_{11}	X_{12}	X_{13}	...	X_{1j}	...	.
2	X_{21}	X_{22}	X_{23}	...	X_{2j}		.
3	X_{31}	X_{32}	X_{33}	...	X_{3j}	...	.
.	.	.	.	...	.		.
.	.	.	.	...	.	...	.
.	.	.	.	...	.		.
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}	...	.
# Evaluators	K_1	K_2	K_3	K_4	K_5		.
Mean	X_1	X_2	X_3	X_4	X_5	...	.
# Alternatives	A					...	σ_E^2

Figure C. Estimation of σ_E^2

It can be seen in Figure 34 that only one value of A can be entered into the spreadsheet. This is because; unlike with the size of the panel of evaluators, the number of alternatives should remain constant, or else meaning that the rating system itself has been modified. In this case, a separate analysis should be conducted with the data collected with a different rating system.

4) *Calculation of the Variance of each Item (S_{Xj}^2, Figure D):* The variance of each item (i.e. asset sample) is calculated using Formula D. As can be seen in Figure D, the A function has been rearranged to be right below the last row of input data.

$$S_{x_j}{}^2 = \frac{1}{K-1}\sum_{k=1}^{K}\left(x_{kj} - \bar{x_j}\right)^2 \qquad \text{(D)}$$

Evaluators	Items					...	σ_E^2
	1	2	3	...	J		
1	X_{11}	X_{12}	X_{13}	...	X_{1j}	...	.
2	X_{21}	X_{22}	X_{23}	...	X_{2j}	...	.
3	X_{31}	X_{32}	X_{33}	...	X_{3j}	...	.
. . .	. . .	. . .	. . .	... ...	. . .	...	. . .
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}	...	.
# Alternatives	A					...	σ_E^2
# Evaluators	K_1	K_2	K_3	K_4	K_5	...	
Mean	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$	$\bar{X}_5$	...	
Variance	S_{X1}^2	S_{X2}^2	S_{X3}^2	S_{X4}^2	S_{X5}^2	...	

Figure D. Estimation of S_{Xj}^2 in the IRA spreadsheet.

5) *Computation of the Single-Item Inter-rater Agreement Index ($r_{WG(I}$,* Figure E)*):* In this step, the computation of the inter-rate agreement index of a particular distress for a single item (i.e. each asset sample) is performed by using Formula A.

Evaluators	Items					...	σ_E^2
	1	2	3	...	J		
1	X_{11}	X_{12}	X_{13}	...	X_{1j}	...	.
2	X_{21}	X_{22}	X_{23}	...	X_{2j}	...	.
3	X_{31}	X_{32}	X_{33}	...	X_{3j}	...	.
.	.	.	.	...	.		.
.	.	.	.	...	.	...	.
.	.	.	.	...	.		.
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}	...	.
# Alternatives	A					...	σ_E^2
# Evaluators	K_1	K_2	K_3	...	K_j	...	
Mean	X_1	X_2	X_3	...	X_j	...	
Variance	S_{X1}^2	S_{X2}^2	S_{X3}^2	...	S_{Xi}^2	...	
Single-Item r_{WG}	$r_{WG(1)}$	$r_{WG(2)}$	$r_{WG(3)}$	...	$r_{WG(j)}$	...	

Figure E. Computation of $r_{WG(I)}$ in the IRA spreadsheet.

6) *Calculation of the Mean Variance of all Items* $\left(\overline{S_{X_J}^2}\right)$ *as in Formula E and in Figure F*:

The items, in this case, represent the asset samples evaluated.

$$\overline{S_{X_J}^2} = \frac{\sum_{j=1}^{J} S_{x_j}^2}{J} \tag{E}$$

Evaluators	Items					...	# Items	Average	σ_E^2
	1	2	3	...	J				
1	X_{11}	X_{12}	X_{13}	...	X_{1j}	...	.	.	.
2	X_{21}	X_{22}	X_{23}	...	X_{2j}	...	.	.	.
3	X_{31}	X_{32}	X_{33}	...	X_{3j}	...	.	.	.
.	.	.	.	...	.		.	.	.
.	.	.	.	...	.	...	.	.	.
.	.	.	.	...	.		.	.	.
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}	...	.	.	.
# Alternatives	A					...	.	.	σ_E^2
# Evaluators	K_1	K_2	K_3	...	K_j	...	.	.	
Mean	X_1	X_2	X_3	...	X_j	...	.	.	
Variance	S_{X1}^2	S_{X2}^2	S_{X3}^2	...	S_{Xj}^2	...	J	Mean S_{XJ}^2	
Single-Item r_{WG}	$r_{WG(1)}$	$r_{WG(2)}$	$r_{WG(3)}$	...	$r_{WG(j)}$	...			

Figure F. Estimation of $\overline{S_{X_j}^2}$ in the IRA spreadsheet.

7) *Computation of the Multi-Item Inter-rater Agreement Index ($r_{WG(J)}$, Figure G):* This is the index of the distress extent or the distress severity that is being assessed. The computation is performed by using Formula B.

Evaluators	Items					...	# Items	Average	σ_E^2	Multi-Item r_{WG}
	1	2	3	...	J					
1	X_{11}	X_{12}	X_{13}	...	X_{1j}	...	.	.	.	.
2	X_{21}	X_{22}	X_{23}	...	X_{2j}	...	.	.	.	.
3	X_{31}	X_{32}	X_{33}	...	X_{3j}	...	.	.	.	.
.	.	.	.	...	.		.	.	.	.
.	.	.	.	...	.	...	.	.	.	.
.	.	.	.	...	.		.	.	.	.
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}	...	.	.	.	.
# Alternatives	A					...	.	.	σ_E^2	.
# Evaluators	K_1	K_2	K_3	...	K_j	...	.	.		.
Mean	X_1	X_2	X_3	...	X_j	...	.	.		.
Variance	S_{X1}^2	S_{X2}^2	S_{X3}^2	...	S_{Xj}^2	...	J	Mean S_{XJ}^2		.
Single-Item r_{WG}	$r_{WG(1)}$	$r_{WG(2)}$	$r_{WG(3)}$	...	$r_{WG(j)}$	...				$r_{WG(J)}$

Figure G. Estimation of $r_{WG(J)}$ in the IRA spreadsheet.

Lindell, Brandt, and Whitney (1999) multiple-item r*$_{WG}$

The estimation of this multi-item IRA index is done through the use of Formula F:

$$r^*_{WG} = 1 - \left(\frac{\overline{S_{x_J}^2}}{\sigma_E^2} \right) \tag{F}$$

Is it worth to mention that this expression is mathematically equal to the average of James et al's $r_{WG(I)}$. Thus, in the IRA spreadsheet, in order to simplify the analysis process, r^*_{WG} will be treated like that, locating it in the cell of the average value of the James' single-item indexes (Figure H).

Evaluators	Items					...	# Items	Average	σ_E^2	Multi-Item r_{WG}
	1	2	3	...	J					
1	X_{11}	X_{12}	X_{13}	...	X_{1j}	...	.	.	.	.
2	X_{21}	X_{22}	X_{23}	...	X_{2j}	...	.	.	.	.
3	X_{31}	X_{32}	X_{33}	...	X_{3j}	...	.	.	.	.
.	.	.	.	...	.		.	.	.	.
.	.	.	.	...	.	...	.	.	.	.
.	.	.	.	...	.		.	.	.	.
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}	...	.	.	.	.
# Alternatives	A					...	.	.	σ_E^2	.
# Evaluators	K_1	K_2	K_3	...	K_j	...	.	.		.
Mean	X_1	X_2	X_3	...	X_j	...	.	.		.
Variance	S_{X1}^2	S_{X2}^2	S_{X3}^2	...	S_{Xj}^2	...	J	Mean S_{Xj}^2		.
Single-Item r_{WG}	$r_{WG(1)}$	$r_{WG(2)}$	$r_{WG(3)}$	...	$r_{WG(j)}$	...		$r^*_{WG(J)}$		$r_{WG(J)}$

Figure H. Estimation of r^*_{WG} in the IRA spreadsheet.

Burke, Finkelstein, and Dusig (1999) Single- and Multi-item AD_M

As referred in chapter 2, the estimation of the average deviation around the mean (AD_M) for a single item is obtained through Formula G, and for multiple items, Formula H.

$$AD_{M(j)} = \frac{\sum_{k=1}^{K} \left| X_{jk} - \bar{X}_J \right|}{K} \tag{G}$$

$$AD_{M(J)} = \frac{\sum_{j=1}^{J} AD_{M(j)}}{J} \tag{H}$$

Thus, in order to include these in the IRA analysis spreadsheet, the following procedure is suggested:

1) *Development of the Deviation around the Mean Matrix (DM_M):* The upper element of Formula 20 is the sum of the absolute differences between each of the ratings of an item and their average. For this, an additional spreadsheet has to be created, called the DM_M. This matrix is built in a similar fashion as Figure A (see Figure J), with the difference that the input data consists of the absolute value of x_{jk} and its respective $\bar{X}_J$, which has been computed in step 2 of the estimation of r_{WG} (Figure B).

Below the data array, the sums of each column have to be calculated. These values represent the numerator in the AD_M formula (G).

Evaluators	Items				
	1	2	3	...	J
1	$\|X_{11}-X_1\|$	$\|X_{12}-X_2\|$	$\|X_{13}-X_3\|$	...	$\|X_{1J}-X_J\|$
2	$\|X_{21}-X_1\|$	$\|X_{22}-X_2\|$	$\|X_{23}-X_3\|$	...	$\|X_{2J}-X_J\|$
3	$\|X_{31}-X_1\|$	$\|X_{32}-X_2\|$	$\|X_{33}-X_3\|$	...	$\|X_{3J}-X_J\|$
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
K	$\|X_{K1}-X_1\|$	$\|X_{K2}-X_2\|$	$\|X_{K3}-X_3\|$	...	$\|X_{KJ}-X_J\|$
Σ	$\Sigma\|X_{k1}-X_1\|$	$\Sigma\|X_{k2}-X_2\|$	$\Sigma\|X_{k3}-X_3\|$	...	$\Sigma\|X_{kJ}-X_J\|$

Figure J. Deviation around the Mean Matrix (DM_M).

2) *Computation of the Single-Item Average Deviation around the Mean Indexes ($AD_{M(j)}$):* With both the numerator and the denominator already estimated (see Figures J and B, respectively), the estimation of the $AD_{M(j)}$ values may proceed. These will be placed in the Single-Item IRA Estimate region, as shown in Figure K.

Evaluators	Items					...	# Items	Average	σ_E^2	Multi-Item r_{WG}
	1	2	3	...	J					
1	X_{11}	X_{12}	X_{13}	...	X_{1j}	...	.	.	.	.
2	X_{21}	X_{22}	X_{23}	...	X_{2j}	...	.	.	.	.
3	X_{31}	X_{32}	X_{33}	...	X_{3j}	...	.	.	.	.
.	.	.	.	...	.		.	.	.	.
.	.	.	.	...	.	...	.	.	.	.
.	.	.	.	...	.		.	.	.	.
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}	...	.	.	.	.
# Alternatives	A					...	.	.	σ_E^2	.
# Evaluators	K_1	K_2	K_3	...	K_j	...	.	.		.
Mean	X_1	X_2	X_3	...	X_j	...	.	.		.
Variance	S_{x1}^2	S_{x2}^2	S_{x3}^2	...	S_{xj}^2	...	J	Mean S_{XJ}^2		.
Single-Item AD_M	$AD_{M(1)}$	$AD_{M(2)}$	$AD_{M(3)}$	...	$AD_{M(j)}$	...				.
Single-Item r_{WG}	$r_{WG(1)}$	$r_{WG(2)}$	$r_{WG(3)}$	...	$r_{WG(j)}$	...		$r^*_{WG(J)}$		$r_{WG(J)}$

Figure K. Estimation of $AD_{M(j)}$ within the IRA spreadsheet.

3) *Computation of the Multi-Item Average Deviation around the Mean Index ($AD_{M(J)}$):* The next step is to obtained the index that determines the overall status of a particular distress. According to Formula 21, the multi-item estimate represents the average of the all the single-item indexes for that distress. Thus, all these have to be summed and the result divided by the number of items, which was already defined in the IRA spreadsheet (Figure F). Just like in the case of r^*_{WG}, the multi-item estimate will be placed in the single-item row, falling within the 'Average' column, as shown in Figure L.

Evaluators	Items					...	# Items	Average	σ_E^2	Multi-Item r_{WG}
	1	2	3	...	J					
1	X_{11}	X_{12}	X_{13}	...	X_{1j}	...	.	.	.	.
2	X_{21}	X_{22}	X_{23}	...	X_{2j}	...	.	.	.	.
3	X_{31}	X_{32}	X_{33}	...	X_{3j}	...	.	.	.	.
.	.	.	.	...	.		.	.	.	.
.	.	.	.	...	.	...	.	.	.	.
.	.	.	.	...	.		.	.	.	.
K	X_{k1}	X_{k2}	X_{k3}	...	X_{kj}	...	.	.	.	.
# Alternatives	A					...	.	.	σ_E^2	.
# Evaluators	K_1	K_2	K_3	...	K_j	...	.	.		.
Mean	X_1	X_2	X_3	...	X_j	...	.	.		.
Variance	S_{X1}^2	S_{X2}^2	S_{X3}^2	...	S_{Xj}^2	...	J	Mean S_{Xj}^2		.
Single-Item AD_M	$AD_{M(1)}$	$AD_{M(2)}$	$AD_{M(3)}$	...	$AD_{M(j)}$	...		$AD_{M(J)}$		.
Single-Item r_{WG}	$r_{WG(1)}$	$r_{WG(2)}$	$r_{WG(3)}$	...	$r_{WG(j)}$	...		$r^*_{WG(J)}$		$r_{WG(J)}$

Figure L. Multi-Item AD_M estimate within the IRA spreadsheet.