An empirical evaluation of the Baldrige Performance Excellence Program framework using applicant data

Jonathan B. Hill

Follow this and additional works at: https://louis.uah.edu/uah-dissertations

Recommended Citation

This Dissertation is brought to you for free and open access by the UAH Electronic Theses and Dissertations at LOUIS. It has been accepted for inclusion in Dissertations by an authorized administrator of LOUIS.
AN EMPIRICAL EVALUATION OF THE BALDRIGE PERFORMANCE EXCELLENCE PROGRAM FRAMEWORK USING APPLICANT DATA

Jonathan B. Hill

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in The Department of Industrial and Systems Engineering and Engineering Management to The Graduate School of The University of Alabama in Huntsville December 2023

Approved by:

Dr. Sampson Gholston, Committee Chair, Department Chair
Dr. Nicholas Loyd, Research Advisor
Dr. Sherri Messimer, Committee Member
Dr. Eric Sholes, Committee Member
Dr. L. Dale Thomas, Committee Member
Dr. Shankar Mahalingam, College Dean
Dr. Jon Hakkila, Graduate Dean
Abstract

AN EMPIRICAL EVALUATION OF THE BALDRIGE PERFORMANCE EXCELLENCE PROGRAM FRAMEWORK USING APPLICANT DATA

Jonathan B. Hill

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

The Department of Industrial and Systems Engineering and Engineering Management

The University of Alabama in Huntsville

December 2023

Over the 30+ years of its existence, the Baldrige Performance Excellence Program (originally known as the Malcolm Baldrige National Quality Award Program) has sought to define the basis for high performing organizations. The Baldrige framework has evolved since its inception in 1988. The purpose of this study was to empirically assess the existing Baldrige framework to determine its validity in context of its continuing evolution (Research Objective 1). Specifically, this research attempted to empirically validate the causal relationships suggested by the framework (Research Objective 2) and determine which criteria items are more important (Research Objective 3). The use of recent, actual scoring data addressed a hole in the existing research literature. The model chosen was firmly established in the research literature and provided a basis of comparison with previous research. This research represents a contribution to the research literature with several theoretical implications. An empirical assessment of the Baldrige framework was conducted utilizing Partial Least Squares Structural Equation Modeling. This research provided strong support that the Baldrige framework is still valid and relevant today in context of its continuing evolution. Although the analysis
conducted found strong support that the Baldrige framework is still relevant through structural equation modeling, the path coefficients suggest that how the components relate to each other may have evolved. Seven of the path coefficients were found to not be statistically significant suggesting that the relationships between categories that these path coefficients represent may not be adequately or accurately defined by the model. The results of this research confirm the continued usefulness of the Baldrige framework to drive organizational performance through confirmation of the causal relationships between the Baldrige categories and the Results category. This research also confirms the continued importance of both Leadership and Measurement, Analysis, and Knowledge Management as the cornerstones of the Baldrige framework.
Acknowledgements

This work would not have been possible without the support and encouragement of numerous individuals.

I would first like to thank my dissertation committee for their influence and direction on this work. Specifically, I would like to thank Dr. Sampson Gholston for his unwavering support, guidance, and encouragement through this entire process. His expertise in formulating the research question and methodology and insightful feedback pushed my work to a higher level.

I would like to acknowledge my leadership and colleagues at the United States Army Integrated Fires Mission Command for their support throughout this lengthy process.

In addition, I would like to thank my parents for their wise counsel throughout my life. Their emphasis on the importance of education and the value of hard work have impacted my education, career, and life above all others. Finally, I could not have completed this dissertation without the support of my family—Kerry, Ava, and Evan.
# Table of Contents

Abstract .............................................................................................................................. ii

Acknowledgements ........................................................................................................... v

Table of Contents ............................................................................................................. vi

List of Figures ................................................................................................................... ix

List of Tables ..................................................................................................................... x

1. **Chapter 1. Introduction** ..................................................................................... 1
   1.1 Background ..................................................................................................... 1
   1.2 Research Objective ......................................................................................... 6
   1.3 Summary ......................................................................................................... 7

2. **Chapter 2. Literature Review** .......................................................................... 10
   2.1 Introduction .................................................................................................. 10
   2.2 Background on Model Evaluations and Criteria Revisions ......................... 10
   2.3 Research in Business/Non-Profit ................................................................. 16
   2.4 Research in Health Care ............................................................................... 24
   2.5 Research in Other Countries ......................................................................... 26
   2.6 Summary ....................................................................................................... 27
   2.7 Conclusions and Observations ..................................................................... 29

3. **Chapter 3. Research Questions and Hypotheses** ........................................... 31
   3.1 Research Question ........................................................................................ 31
5.7 Summary

6. Chapter 6. Conclusions and Recommendations

   6.1 Introduction

   6.2 Conclusions

   6.3 Assumptions and Limitations

   6.4 Future Research

   6.5 Summary

References
List of Figures

Figure 1.1 Baldrige Criteria Changes (Hertz, 2019) .......................................................... 4

Figure 1.2 Example of Baldrige Criteria in the Customers Category (NIST.gov/Baldrige) ........................................................................................................... 5

Figure 2.1 Malcolm Baldrige National Quality Award Original Causal Model (Wilson & Collier, 2000) ........................................................................................................... 17

Figure 2.2 Final Causal Model with Financial Results as the Performance Variable (Wilson & Collier, 2000) .......................................................... 19

Figure 2.3 Final Causal Model with Customer Satisfaction as the Performance Variable (Wilson & Collier, 2000) .......................................................... 19

Figure 2.4 Final Causal Model of the 1993 Baldrige Structure utilizing the Arizona Quality Award Day (Pannirselvam & Ferguson, 2001) .......................................................... 21

Figure 2.5 Flynn and Saladin’s (2001) Path Model Based on 1997 Baldrige Framework ........................................................................................................................ 24

Figure 2.6 Results of Testing Causal Relationships in Baldrige Health Care Criteria (Meyer and Collier, 2001) ........................................................................................................... 25

Figure 3.1 Structural Path Model ..................................................................................... 36

Figure 4.1 Example Path Model ...................................................................................... 47

Figure 4.2 Basic PLS-SEM Algorithm (Hair et al., 2021) .................................................. 49

Figure 5.1 Structural Path Model ..................................................................................... 59

Figure 5.2 PLS-SEM Analysis with Path Coefficients and R² Values .......................... 61

Figure 5.3 Statistical Significance of Path Coefficients via Bootstrapping ..................... 64
List of Tables

Table 2.1 Baldrige Categories, Subcategories, and Points by Major Revision Year (Lee et al., 2006) .......................................................................................................................... 13

Table 4.1 Example Applicant Scoring Data from 2008 to 2013 .............................................. 40

Table 4.2 Example Applicant Scoring Data from 2014 to 2018 .............................................. 41

Table 4.3 Number of Applicants Per Year ............................................................................... 42

Table 4.4 Descriptive Statistics of Examiner Data ..................................................................... 43

Table 4.5 Anderson-Darling Test for Normality Results ............................................................. 43

Table 4.6 Evaluation Parameters and Desired Results ............................................................. 55

Table 5.1 Total Effects by Category ........................................................................................ 62

Table 5.2 VIF Scores by Category ........................................................................................... 62

Table 5.3 Blindfolding Results ................................................................................................. 65

Table 5.4 Statistical Significance of Path Coefficients via Bootstrapping ............................... 67

Table 5.5 Holm-Bonferroni Correction Comparisons ............................................................... 68

Table 6.1 Summary of Previous Research ............................................................................... 74

Table 6.2 Evaluation Parameters, Desired Results, and Achieved Results ......................... 78
Chapter 1. Introduction

1.1 Background

In the 1980s, United States political leaders faced with the reality of an ever-expanding global market realized the need for American companies to focus on quality. Secretary of Commerce Malcolm Baldrige, who passed away in an accident in 1987, was an advocate of quality management. Congress would later that same year pass the Malcolm Baldrige National Quality Improvement Act as an effort to enhance the competitiveness of American companies. Congress created the program to (1) identify and recognize role-model businesses, (2) establish criteria for evaluating improvement efforts, and (3) disseminate and share best practices (“History: Who We Are”).

Over the 35+ years of its existence, the Baldrige Performance Excellence Program (originally known as the Malcolm Baldrige National Quality Award Program) has sought to define the basis for high performing organizations. Originally developed purely as an award program targeted at for profit businesses, it has evolved into a program that seeks to help organizations in all market sectors continuously improve and innovate in order to improve performance. Initially based on a single set of business award criteria, the program today boasts a suite of performance improvement tools tailorable to all market sectors. The program has been copied and emulated throughout the United States in regional, state, and local versions of the program as well as throughout the world in other countries. The Presidential award, though originally created for business, has expanded
to include applicant categories in education, health care, non-profit, small business, and Government organizations.

According to Hertz (2019), the original definition of performance excellence in the context of Baldrige was “to deliver ever-improving value to customers, while improving the overall effectiveness and efficiency of the organization.” Today, performance excellence has been redefined in the Baldrige context to reflect a more complex approach—“an integrated approach to organizational performance excellence that results in (1) delivery of ever-improving value to customers and stakeholders, contributing to ongoing organizational success; (2) improvement of your organization’s overall effectiveness and capabilities, and (3) learning for the organization and people in the workforce.”

There have been four major iterations of the Baldrige framework, which consists of the model with category interactions and the underlying criteria for each of those categories. The original framework developed in 1988 consisted of seven major categories—Leadership, Planning, Information and Analysis, Human Resources, Quality Assurance Products and Services, Results of Quality Assurance, and Customer Satisfaction. In 1992, the framework was updated to depict seven slightly different categories—Senior Executive Leadership, Management of Process Quality, Human Resource Development and Management, Strategic Quality Planning, Information and Analysis, Customer Focus and Satisfaction, and Quality and Operational Results.

In 1997, the framework was updated yet again, in what is widely viewed as the most significant change to the framework. The framework evolved from a focus on quality to focus on overall performance excellence and a systems approach to that overall performance excellence.
performance. Also in 1997, Results became its own category with a substantial emphasis from a points perspective (450 out of 1000 total possible points) to emphasize that a process could not be truly effective without achieving desired results. The 1997 update also once again saw a change to the categories in what would become a semi-stable version lasting until present day—Leadership, Strategic Planning, Customer and Market Focus, Information and Analysis, Human Resource Development and Management, Process Management, and Business Results. The framework today includes a visual change in framework representation made in 2015, but still includes the same 7 basic categories as when first deployed in 1997—Leadership; Strategy; Customers; Measurement, Analysis, and Knowledge Management; Workforce, Operations, and Results.

The Criteria for Performance Excellence were originally in 1988 a set of statements that evolved into a set of questions in 1999 and subsequent years. The criteria were developed anecdotally by industry leaders and have evolved over the years in an effort of continuous improvement and evolution to match the ever-changing organizational environment and landscape.

The criteria ask how you accomplish your mission in each category, how you plan for the future, and what your results are. Award applicants submit an organizational profile and an application package addressing how their organization addresses the content of each category. The application is scored by a board of examiners trained in assessing the implementation of the criteria. Individual scores produced by examiners are elevated for discussion and debated at the team level to produce a consensus score for the applicant. High performing applicant organizations at the national level are selected
for a site visit by a team of examiners and given a further refined consensus score based on the findings of that visit by the team of examiners. The highest consensus scored organizations in each market sector are selected as recipients for the Presidential award at the National level. State and regional award organizations have varying winning criteria that may allow for multiple awards within an individual market sector to allow and encourage individual organizational achievement over pure competition.

**BALDRIGE CRITERIA OVERVIEW**

![Baldrige Criteria Changes](image)

*Figure 1.1* Baldrige Criteria Changes (Hertz, 2019).
Bemowski and Stratton (1995) found that the criteria are used not just by award applicants alone. Businesses use the criteria to obtain information on how to achieve business excellence and the criteria typically met or exceeded expectations of the business. Additionally, the criteria are used by a wide range of organizations within a broad set of industries to include even the American Association of Retired Persons (Schaefer, 2016).

In 2022, Baldrige announced (Fangmeyer, 2022) suspension of the award process for that year to “initiate a comprehensive, independent review of the Baldrige program. This review will assess how the program can best advance U.S. competitiveness and
address the challenges most relevant in today's business environment, as well as examine how its impact and accessibility could be increased.” In an announcement published by Quality Progress (Baldrige Process Paused, 2022)—“After much deliberation and consideration of multiple factors, including participation data over the past several years, we have collectively decided to initiate a comprehensive, independent review of the Baldrige Program. This review will assess how the program can best advance U.S. competitiveness and address the challenges most relevant in today's business environment, as well as examine how its impact and accessibility could be increased.”

1.2 Research Objective

The Baldrige framework has evolved since its inception in 1988. While this evolution has been well documented and researched, a final validation of the framework is ever elusive due to the continuous evolution of the criteria to account for societal and organizational performance improvement. NIST provided historical scoring data in 2011 of applicants’ submittals from 1988 to 2009. The scoring data provided was appropriately analyzed and published in 2018 by Mai et al. representing the most recent evaluation of the framework. While the results from Mai et al. (2018) research successfully validated the existing framework, major changes in the criteria but not the model since 2009 represent a research gap in the framework validation.

The purpose of this study was to empirically assess the existing Baldrige framework to determine its validity in context of its continuing evolution (Research Objective 1). Specifically, this research attempted to empirically validate the causal relationships suggested by the framework (Research Objective 2) and determine which criteria items are more important (Research Objective 3). This research did not attempt
To draw conclusions about the overall effects on the organization of being an award winner (e.g., increased performance, profitability, etc.). The Baldrige framework was developed based on knowledge and collaborative input from various Subject Matter Experts. Research objective two was to empirically validate the causal relationships suggested by the framework.

This research made several contributions to the field of study. First, this research utilizes analytical techniques that have not been practical for use in some previous studies, such as structural equation modeling. Second, this research used a data set of applicant consensus scoring data from 2008 to 2017 from the southern region of the United States (modeled after the Baldrige Performance Excellence Award). This examiner data is significantly more recent than anything that has been published in the existing research literature. Third, this research made several observations about the relationships expressed in the model that are practical and relevant to managers seeking to utilize the Baldrige framework to improve their organization as well as researchers seeking to better understand the Baldrige framework and its application.

1.3 Summary

Over the 35+ years of its existence, the Baldrige Performance Excellence Program has sought to define the basis for high performing organizations. There have been four major iterations of the Baldrige framework—the original framework developed in 1988, a slight change to the categories in 1992, a significant change to the model and categories in 1997, and a visual change in model representation made in 2015.

The Criteria for Performance Excellence were originally in 1988 a set of statements that evolved into a set of questions in 1999 and subsequent years. The criteria
were developed anecdotally by industry leaders and have evolved over the years in an effort of continuous improvement and evolution to match the ever-changing organizational environment and landscape.

The Baldrige framework evolution since its inception in 1988 has been well documented and researched; a final validation of the framework is ever elusive due to the continuous evolution of the criteria to account for societal and organizational performance advancement. While the 2011 release by NIST of blinded applicant scoring data from 1988 to 2009 has fueled recent research, continuous changes in the criteria since that time represent a significant gap in the research literature.

The purpose of this study was to empirically assess the existing Baldrige framework to determine its validity in context of its continuing evolution (Research Objective 1). Specifically, this research attempted to empirically validate the causal relationships suggested by the framework (Research Objective 2) and determine which criteria items are more important (Research Objective 3). The Baldrige framework was developed based on knowledge and collaborative input from various Subject Matter Experts. Research objective two was to empirically validate the causal relationships suggested by the framework.

This research made several contributions. First, this research utilized analytical techniques that have not been practical for use in some previous studies. Second, this research used a data set of applicant consensus scoring data from 2008 to 2017 from a state Baldrige Award (modeled after the Baldrige Performance Excellence Award). This data is significantly more recent than anything that has been published in the existing research literature. Third, this research highlighted several observations about the
relationships expressed in the framework that are practical and relevant to managers seeking to utilize the Baldrige framework to improve their organization as well as researchers seeking to better understand the Baldrige framework and its application.
Chapter 2. Literature Review

2.1 Introduction

An extensive literature search was conducted to examine the body of research related to the evolution of the Baldrige framework and to causal analysis of the Baldrige framework. The research literature on causal analysis fell into one of three main areas—(1) research done utilizing data from manufacturing, service, small business, nonprofit, and Government organizations (collective referred to as Business/Non-profit) for which the criteria were originally intended, (2) research done utilizing data from health care organizations and the health care criteria developed in 1995, and (3) research conducted utilizing data from various organizations outside of the United States.

2.2 Background on Model Evaluations and Criteria Revisions

The Baldrige model is one of the most prominent business or performance excellence model in use today, accounting for approximately 15% of all business or performance excellence models worldwide (Ghafoor et al., 2022). Analysis of the Baldrige Performance Excellence Criteria began at its original inception in 1988 and continues to this day.

Most recently, Rangsungnoen (2022) tested causal relationships within grouped categories. Ghosh et al. (2003) compared the relationships between strategic and operational quality constructs within the model. The researchers’ analysis found support
the strategic quality planning process as the driver of operational quality planning. Curkovic et al. (2000) examined the extent to which the Baldrige model captured the major dimensions of Total Quality Management. Tummala and Tang (1996) examined the core concepts of Strategic Quality Management and compared them to the Baldrige and European Quality Award Criteria as well as ISO 9001 certification requirements.

Asfa et al. (2021) compared the Baldrige model to the European Foundation for Quality Management model and the Deming Prize model. They found the Baldrige model to be the most well structured model of the three. Gupta and Vrat (2020) compared the National Quality Award models in terms of the criteria employed and their relative weights for the models from USA (Baldrige), Canada, Europe, Australia, Japan, and India. They found high variability in the parameter weights, a lack of holistic organizational performance assessment within some of the models, and variation in how often (if at all) the criteria weights are changed. Alanazi (2021) developed a measurement model to compare the Baldrige model, European Foundation of Quality Management (EFQM), and the King Abdul Aziz Quality Award, but did not perform empirical tests. Ghafoor et al. (2021) researched the strengths and opportunities of the Baldrige model’s promotion, facilitation, and award related activities. He identified a key opportunity of increasing award awareness and facilitating organizational use of the Baldrige approach. Ghobadian and Woo (1996) compared the characteristics, benefits, and shortcomings of the Baldrige Award, Deming Application Prize, the European Quality Award, and the Australian Quality Award. Bohoris (1995) compared the Japanese, European, and US (Baldrige) Quality awards against the attainment of Total Quality Management Objectives.
According to Lee et al. (2006), the original Baldrige framework was released in 1988 with seven categories that were worth a total of 1000 points. Though major revisions to the framework occurred in 1992 and 1997 with a minor revision in 2003, the criteria have continued to consist of seven categories with 1000 total points allocated to those categories. The original criteria were prescriptive in nature. Table 2.1 beginning on the next page summarizes the Baldrige Categories, Subcategories, and Points for the revisions in 1992, 1997, and 2003.

The 1992 revision to the criteria changed the criteria from being prescriptive to less rigid, more descriptive areas to address. While the number of subcategories and number of areas to address changed, the overall criteria revision for 1992 was roughly analogous to the 1988 criteria.

The 1997 criteria revision was a major one. The results section was changed from a focus on primarily customer satisfaction to now also emphasize financial results, productivity, safety, and employee morale as critical outcomes of the applicants’ processes. Additionally, the Baldrige categories were reordered.
Table 2.1 Baldrige Categories, Subcategories, and Points by Major Revision Year (Lee et al., 2006).

| Year | Leadership, 150 points | Senior corporate leadership, 50 points | Policy, 30 points | Management system and quality improvement processes, 30 points | Resource allocation and utilization, 20 points | Public responsibility, 20 points | Information and analysis, 25 points | Data and analysis, 20 points | Supplier quality and data analysis, 10 points | Distributor and/or dealer quality and data analysis, 10 points | Employee related data and analysis, 5 points | Unique and innovative leadership techniques, 10 points | Leadership, 100 points | Senior executive leadership, 45 points | Management for quality, 25 points | Public responsibility and citizenship, 20 points | Leadership system, 80 points | Company responsibility and citizenship, 80 points | Organizational leadership, 80 points | Social responsibility, 40 points |
|------|------------------------|----------------------------------------|-------------------|------------------------------------------------|-----------------------------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------------|-----------------------------------|---------------------------------|-----------------------------|------------------------|-------------------|-------------------|------------------------|-------------------------|------------------|----------------|-----------------|------------------|
| 1988 | 1. Leadership, 150 points | 1. Senior corporate leadership, 50 points | 1. Policy, 30 points | 1.3. Management system and quality improvement processes, 30 points | 1. Resource allocation and utilization, 20 points | 1. Public responsibility, 20 points | 2. Information and analysis, 25 points | 2. Use of analytical techniques or systems, 15 points | 2.4. Customer data and analysis, 20 points | 2.5. Distributor and/or dealer quality and data analysis, 10 points | 2.6. Employee related data and analysis, 5 points | 1. Leadership, 150 points | 1. Leadership system, 80 points | 1.2. Management for quality, 25 points | 1.3. Public responsibility, 20 points | 1.1. Leadership system, 80 points | 1.2. Company responsibility and citizenship, 20 points | 1. Leadership, 120 points | 1.2. Social responsibility, 40 points |
| 1992 | 1. Leadership, 100 points | 1. Senior executive leadership, 45 points | 1. Policy, 30 points | 1.3. Management system and quality improvement processes, 30 points | 1. Resource allocation and utilization, 20 points | 1. Public responsibility, 20 points | 2. Information and analysis, 25 points | 2. Use of analytical techniques or systems, 15 points | 2.4. Customer data and analysis, 20 points | 2.5. Distributor and/or dealer quality and data analysis, 10 points | 2.6. Employee related data and analysis, 5 points | 1. Leadership, 150 points | 1. Leadership system, 80 points | 1.2. Management for quality, 25 points | 1.3. Public responsibility, 20 points | 1.1. Leadership system, 80 points | 1.2. Company responsibility and citizenship, 20 points | 1. Leadership, 120 points | 1.2. Social responsibility, 40 points |
| 1997 | 1. Leadership, 100 points | 1. Senior executive leadership, 45 points | 1. Policy, 30 points | 1.3. Management system and quality improvement processes, 30 points | 1. Resource allocation and utilization, 20 points | 1. Public responsibility, 20 points | 2. Information and analysis, 25 points | 2. Use of analytical techniques or systems, 15 points | 2.4. Customer data and analysis, 20 points | 2.5. Distributor and/or dealer quality and data analysis, 10 points | 2.6. Employee related data and analysis, 5 points | 1. Leadership, 150 points | 1. Leadership system, 80 points | 1.2. Management for quality, 25 points | 1.3. Public responsibility, 20 points | 1.1. Leadership system, 80 points | 1.2. Company responsibility and citizenship, 20 points | 1. Leadership, 120 points | 1.2. Social responsibility, 40 points |
| 2003 | 1. Leadership, 120 points | 1. Senior executive leadership, 45 points | 1. Policy, 30 points | 1.3. Management system and quality improvement processes, 30 points | 1. Resource allocation and utilization, 20 points | 1. Public responsibility, 20 points | 2. Strategic planning, 40 points | 2.1. Strategy development process, 40 points | 2.4. Customer data and analysis, 20 points | 2.5. Distributor and/or dealer quality and data analysis, 10 points | 2.6. Employee related data and analysis, 5 points | 1. Leadership, 150 points | 1. Leadership system, 80 points | 1.2. Management for quality, 25 points | 1.3. Public responsibility, 20 points | 1.1. Leadership system, 80 points | 1.2. Company responsibility and citizenship, 20 points | 1. Leadership, 120 points | 1.2. Social responsibility, 40 points |
|------|------|------|------|
| 3. Strategic quality planning, 75 points | 3. Strategic quality planning, 60 points | 3. Customer and market focus, 80 points | 3. Customer and market focus, 85 points |
| 3.2. Planning function, 20 points | 3.2. Quality and performance plans, 25 points | 3.2. Customer satisfaction and relationship enhancement, 40 points | 3.2. Customer relationships and satisfaction, 45 points |
| 3.3. Planning for quality improvement, 30 points | | | |
| 3.4. Unique and innovative planning, 5 points | | | |
| 4.2. Employee quality awareness and involvement, 50 points | 4.2. Employee involvement, 40 points | 4.2. Selection and use of comparative information and data, 40 points | 4.2. Information and knowledge management, 45 points |
| 4.3. Quality training and education, 30 points | 4.3. Employee education and training, 40 points | 4.3. Analysis and review of company performance, 40 points | |
| 4.4. Evaluation, incentive and recognition systems, 30 points | 4.4. Employee performance and recognition, 25 points | | |
| 4.5. Unique and innovative approaches, 10 points | 4.5. Employee well-being and morale, 25 points | | |
| 5. Quality assurance of products and services, 150 points | 5. Management of process quality, 140 points | 5. Human resource development and management, 100 points | 5. Human resource focus, 85 points |
| 5.1. Customer input to products and services, 20 points | 5.1. Design and introduction of quality products and services, 40 points | 5.1. Work systems, 40 points | 5.1. Work systems, 35 points |
| 5.2. Planning for new or improved products or services, 20 points | 5.2. Process management, product and service production and delivery processes, 35 points | 5.2. Employee education, training and development, 30 points | 5.2. Employee learning and motivation, 25 points |
| 5.3. Design of new or improved products and services, 30 points | 5.3. Process management, business processes and support services, 30 points | 5.3. Employee well-being and satisfaction, 30 points | 5.3. Employee well-being and satisfaction, 25 points |
| 5.4. Measurements, standards and data system, 10 points | 5.4. Supplier quality, 20 points | | |
| 5.5. Technology, 15 points | 5.5. Quality assessment, 15 points | | |
| 5.6. Audit, 15 points | | | |
| 5.7. Documentation, 10 points | | | |
| 5.8. Safety, health and environment, 10 points | | | |
| 5.9. Assurance/validation, 15 points | | | |
| 5.10. Unique and innovative approaches, 10 points | | | |
| 6. Results from quality assurance of products and services, 100 points | 6. Quality and operational results, 150 points | 6. Process management, 100 points | 6. Process management, 85 points |
| 6.2. Reductions in scrap, rework, and rejected products or services, 20 points | 6.2. Company operational results, 45 points | 6.2. Management of support processes, 20 points | 6.2. Support processes, 35 points |
The 2003 criteria revised categories are nearly identical to the ones in 1997 with only a slight change in nomenclature for Category 4 and 5. The most notable changes were the removal of subcategories specifically related to supplier and partnering processes and results to reflect a more enterprise approach.

Revisions to the criteria are now completed in 2-year cycles beginning in 2007. The 2007 criteria contained significant revision to address four areas of growing importance: (1) strategic advantages and core competencies; (2) innovation; (3) work systems for producing business results; and (4) workforce engagement, workforce capability, and workforce capacity. Additionally, the award expanded to include Government and non-profit categories in 2007. Evolutions of the criteria concepts include the addition of succession planning in 2005; strategic challenges, advantages, and core competencies in 2007; customer engagement in 2009; a focus on the alignment of core competencies with strategy and work processes in 2011; a focus on the role of innovation in strategic

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6.4. Reductions in warranty or field support work, 20 points</td>
<td>6.4. Supplier quality results, 35 points</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.5. Unique or innovative indicators of quality improvements or economic gains, 10 points</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.2. Competitive comparison of products or services, 50 points</td>
<td>7.2. Commitment to customers, 15 points</td>
<td>7.2. Financial and market results, 130 points</td>
<td>7.2. Product and service results, 75 points</td>
</tr>
<tr>
<td>7.3. Customer service and complaint handling, 75 points</td>
<td>7.3. Customer satisfaction determination, 35 points</td>
<td>7.3. Human resource results, 35 points</td>
<td>7.3. Financial and market results, 75 points</td>
</tr>
<tr>
<td>7.4. Customer views of guarantees/warranties, 50 points</td>
<td>7.4. Customer satisfaction results, 75 points</td>
<td>7.4. Supplier and partner results, 25 points</td>
<td>7.4. Human resource results, 75 points</td>
</tr>
<tr>
<td>7.5. Unique or innovative approaches to assessing customer satisfaction, 25 points</td>
<td>7.5. Customer satisfaction comparison, 75 points</td>
<td>7.5. Company-specific results, 130 points</td>
<td>7.5. Organizational effectiveness results, 75 points</td>
</tr>
<tr>
<td>7.6. Future requirements and expectations of customers, 35 points</td>
<td></td>
<td>7.6. Governance and social responsibility results, 75 points</td>
<td></td>
</tr>
<tr>
<td>Total points: 1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>
opportunities and intelligent risk taking as well as the role of social media in 2013. The 2015 criteria featured a renewed focus on managing and leading all the components of the organization as a unified whole, managing change, and dealing with data analytics, data integrity, and cybersecurity. These revisions were made to ensure that the criteria remained on the leading edge of validated leadership and performance practice.

The Baldrige criteria have evolved since their original inception. They have expanded to include applicant categories in education, health care, non-profit, small business, and Government organizations with their own tailored criteria. Initially based on a single set of business award criteria, the program today boasts a suite of performance improvement tools tailorable to all market sectors. Performance excellence has been redefined in the Baldrige context to reflect a more complex approach. While initially a part of each individual category, results became its own category with a substantial emphasis from a points perspective (450 out of 1000 total possible points) to emphasize that a process could not be truly effective without achieving desired results. The criteria have evolved over the years in an effort of continuous improvement and evolution to match the ever-changing organizational environment and landscape. While the names have changed each of the seven categories represented in the criteria today have been present from the beginning.

2.3 Research in Business/Non-Profit

Wilson and Collier (2000) were the first to test the causal performance linkages implied by the Baldrige framework utilizing structural equation modeling. They utilized a comprehensive set of 101 questions that were directly tied to the 1995 Baldrige criteria. The theory behind the Baldrige framework states, “leadership drives the system that
creates results.” For purposes of their analysis the “system” included the Process Management, Human Resource Development & Management, Strategic Planning, and Information & Analysis categories. The Baldrige theory at the time suggested a recursive causal model containing no reciprocal causation or feedback loops, yet when the Baldrige quality experts defined this system in a diagram (Figure 2.1) they included both reciprocal causation and feedback loops. Essentially, they had admitted they were not sure of the performance relationships and directions of causation and hence defaulted to the idea that everything is related to everything else (Wilson & Collier, 2000).

Figure 2.1 Malcolm Baldrige National Quality Award Original Causal Model (Wilson & Collier, 2000).

Wilson and Collier (2000) hypothesized that the relationship was actually a recursive causal model as originally stated and that the sign of each path coefficient is positive
(“leadership drives the system that creates results”). Their research objectives were to (1) develop a measurement model and associated constructs and scales that are accurate in capturing the content and intent of the Baldrige criteria, (2) test the general theory that “leadership drives the system that creates results,” and (3) provide better insights into the specific directions of causation among the seven Baldrige categories than the original two headed arrow scheme shown previously in Figure 2.1. They proposed to evaluate the model through structural equation modeling.

The survey developed by Wilson and Collier (2000) was evaluated for content validity and comprehensive coverage of the seven categories (referred to as constructs) and sub-categories (referred to as dimensions). The population for the main sample consisted of manufacturing companies in the United States automobile industry. The survey sample size was 800 manufacturing firms with a response rate of 31.4%. The analysis of the results by structural equation modeling was a two-stage process. First, the measurement model was estimated using factor analysis to obtain a factor score for each Baldrige category. Second, the measurement model was fixed when the structural model was estimated.

The initial analyses were divided into two models to reduce the number of estimated parameters. The only difference in the two models was the performance variable (Financial Results or Customer Satisfaction). The results are shown in Figures 2.2 and 2.3 below. Only the statistically significant paths are shown including the within system paths, the path coefficients, and the hypotheses that were tested.
Figure 2.2 Final Causal Model with Financial Results as the Performance Variable (Wilson & Collier, 2000).

Figure 2.3 Final Causal Model with Customer Satisfaction as the Performance Variable (Wilson & Collier, 2000).
The research results provided strong evidence that the Baldrige underlying theory of “leadership drives the system that causes results” is correct. The research results also provided a proven model of the causal relationships between the seven categories or constructs. Wilson and Collier (2000) drew many conclusions about the relationships and drivers of the various categories. Furthermore, they suggested additional research should be done to see if these relationships held true in different industries, countries, and cultures.

Pannirselvam and Ferguson (2001) also sought to test the validity and strength of the proposed relationships between the categories in the Baldrige framework. Unlike Wilson and Collier (2000), Pannirselvam and Ferguson (2001) utilized actual Baldrige scoring data from the 1993 Arizona Governor’s Quality Award process. The data consisted of 69 organizations who had applied for the award. Additionally, they tested an alternative Quality Performance model.

The alternative model along with the results of the analysis is shown in Figure 2.4 below. The results only partially validated the notion that leadership is the driving force that influences all other elements of the quality system. Pannirselvam and Ferguson (2001) hypothesized that since their research utilized trained examiner scoring data it was more objective than self-reported scale item data such as that used by Wilson and Collier (2000). They suggested further research should concentrate on more varied data sources to confirm their results. Additionally, they pointed out that their research was based on criteria and data from 1993 and there had already been a major revision to the criteria in 1997.
Evans and Jack (2003) utilized survey data and canonical correlation to validate the relationships of the criteria categories. Their results also validated all category linkages in the overall structural model, although they noted that structural equation modeling would produce more statistically robust results that would account for measurement error and concurrent validation of all the linkages within the model. Validation is defined in the context of this and future research as a statistically significant path coefficient representing a causal relationship between two categories of the Baldrige model.

Evans (2010) and Evans and Mai (2014) utilized blinded applicant total scoring data with descriptive statistical analysis and some basic statistical inference tests to develop insights regarding examiner performance. He observed significant differences between large organizations and the small business sector, a decline in scoring-related
performance relative to the criteria within the for-profit sectors, steady increases within health care and education, and an improvement in examiner performance as measured by scoring variability. Evans noted the limitations of only having total scoring data rather than the category scoring data and the recent release of such data from NIST that would allow for a much richer analysis. In 2012, Evans et al. would publish later research focused on common opportunities for improvement provided to award recipients that could be used to accelerate the process of improving performance.

Karimi et al. (2014) identified critical factors that predict quality management program success using the Baldrige criteria as a measurement proxy. He performed longitudinal and cross-sectional data mining analyses using ANOVA, multiple regression, discriminant analysis, and decision tree analysis to compare applicant data released from NIST. This research found the Leadership and Measurement, Analysis, and Knowledge Management factors to be the most important.

Flynn and Saladin (2001) sought to statistically validate utilizing structural equation modeling that the major revisions (1988, 1992, and 1997) in the Baldrige criteria were in fact improvements. Their data set was pulled from survey data of the World Class Manufacturing Database, Round II which included both domestic and foreign plants. Their results indicated that each of the three frameworks included robust relationships and that the subsequent revisions in 1992 and 1997 did in fact improve on the previous model by utilizing path analysis to confirm that the model fit improved with each subsequent iteration of the model. Lee et al. (2006) attributed these revisions to shifts in the global economy and technology that changed the role of quality from an order winner to an order qualifier.
Mai et al. (2018) built off the work of Flynn and Saladin (2001) by taking their model (Figure 2.5), individual review blinded applicant scoring data from 1999 to 2006 and utilized partial least squares structural equation modeling to analyze the data. The model was found to have considerable predictive and explanatory power. The individual pathways shown in Figure 2.5 were all found to be statistically significant. Additionally, multiple group analyses were conducted to examine whether the model was consistent for various industries—manufacturing, service, small business, education, and health care. This group comparison was conducted at both the measurement and structural level. The measurement model was confirmed as adequate and similar for all groups; however, the structural coefficients were found to be different among the various industries. Pathway coefficients were found to be not statistically significant in each industry specific model, ranging from 1 in 15 non-significant in the health care industry to 5 in 15 not statistically significant in the service sector.

Mai et al. (2018) found that although Measurement, Information, and Analysis is important in the overall model, its predictive power was less within the Service category having no direct impact on the Customer Focus and Results categories. They attributed this to the role of human factors in determining quality within the Service category. Additionally, Mai et al. (2018) found that the link between the Strategy and Customer focus is more important in the Healthcare and Service industry categories. In all other market categories this link was weaker and not statistically significant. Mai et al. (2018) suggested that this relationship was different in the other categories due to a focus on product rather than services. Lee et al. (2003) similarly found that the principles of
supplier relationships, leadership commitment, and customer orientation were inconsistent between manufacturing and service organizations.

Figure 2.5 Flynn and Saladin’s (2001) Path Model Based on 1997 Baldrige Framework.

2.4 Research in Health Care

In 1995, separate Baldrige Criteria was designed specifically for health care and educational organizations. The first applications for award were accepted in 1999, but no health care or educational organization winners were awarded in 1999 or 2000. Health care researchers suggested that evidence needed to be presented that proved the Baldrige criteria were an effective measure of quality in their industry (Meyer and Collier, 2001).

Meyer and Collier (2001) set out to prove that the Baldrige Health Care criteria were in fact an effective set of criteria. Like Wilson and Collier before them, they utilized a survey to collect data for use in structural equation modeling. The survey data came from 220 US hospitals. Their results are shown in Figure 2.6 below. The research validated the causal model suggested by the Health Criteria and the authors made several
observations from the research about where hospitals should invest their resources to most significantly influence outcomes.

Figure 2.6 Results of Testing Causal Relationships in Baldrige Health Care Criteria (Meyer and Collier, 2001).
2.5 Research in Other Countries


In China, Sun (2011) attempted to validate the framework using a survey of 200 manufacturing companies. Interestingly, his research indicated that several of the causal paths previously proven out in studies of American organizations did not hold true for the Chinese. It should be noted that Hongyi himself noted that his research was based on a rather small sample size and as such was fallible. He also commented on the relative absence of research literature on the subject after 2003. Also in China, He et al. (2011) validated the causal model utilizing a survey of 2,302 manufacturing organizations. Interestingly, their research indicated process management to be the most important construct (not leadership as had been found in all other previous studies). They attributed this to differences in Quality Management practices in China versus the United States.

Also in China, Goldstein et al. (2002) developed a measurement model tied directly to the Baldrige criteria at the construct and dimension levels that validated the theoretical model underlying the Baldrige framework utilizing exploratory factor analysis and confirmatory factor analysis on data collected from manufacturing and service firms in China.
Jayamaha et al. (2008) empirically assessed the criteria validity for New Zealand organizations. His research found 11 of the 13 implied causal relationships to be statistically significant and found Leadership and Measurement, Analysis, and Knowledge Management to be the two most important overall categories. His follow on research (Jayamaha et al., 2011) did similar work for applicants in New Zealand. Masood et al. (2006) tested the causal relationships of the model using university and college data from the United Arab Emirates.

2.6 Summary

An extensive literature search was conducted to examine the body of research related to the evolution of the Baldrige framework and to causal analysis of the Baldrige model. The research literature on causal analysis fell into one of three main areas—(1) research done utilizing data from manufacturing, service, small business, nonprofit, and Government organizations (collective referred to as Business/Non-profit) for which the criteria were originally intended, (2) research done utilizing data from health care organizations and the health care criteria developed in 1995, and (3) research conducted utilizing data from various organizations outside of the United States.

According to Lee et al. (2006), the original Baldrige Criteria were released in 1988 with seven categories that were worth a total of 1000 points. Though major revisions to the framework occurred in 1992 and 1997 with a minor revision in 2003, the criteria have continued to consist of seven categories with 1000 total points allocated to those categories.

Wilson and Collier (2000) were the first to test the causal performance linkages implied by the Baldrige model utilizing structural equation modeling. They utilized a
comprehensive set of 101 questions that were directly tied to the 1995 Baldrige framework and structural equation modeling to evaluate the model. The results provided strong evidence that the Baldrige underlying theory of “leadership drives the system that causes results” is correct.

Pannirselvam and Ferguson (2001) also sought to test the validity and strength of the proposed relationships between the categories in the Baldrige framework. Unlike Wilson and Collier, Pannirselvam and Ferguson utilized actual Baldrige scoring data from the 1993 Arizona Governor’s Quality Award process.

Evans (2010) utilized blinded applicant total scoring data with descriptive statistical analysis and some basic statistical inference tests to develop insights regarding examiner performance.

Flynn and Saladin (2001) sought to statistically validate utilizing structural equation modeling that the major revisions (1988, 1992, and 1997) in the Baldrige criteria were in fact improvements. Their results indicated that each of the three frameworks included robust relationships and that the subsequent revisions in 1992 and 1997 did in fact improve on the previous framework based on a statistical analysis that measured goodness of fit of the model showing improvement in each subsequent iteration.

Mai et al. (2018) built off the work of Flynn and Saladin (2001) by taking their model (Figure 2.5), individual review blinded applicant scoring data from 1999 to 2006 and utilized partial least squares structural equation modeling to analyze the data. The model was found to have considerable predictive and explanatory power. The
measurement model was confirmed as adequate and similar for all groups; however, the structural coefficients were found to be different among the various industries.

2.7 Conclusions and Observations

The research available on causal analysis of the Baldrige framework is far from all encompassing. Several gaps exist within the literature. The initial research literature was focused on the criteria for manufacturing, service, small business, nonprofit, and government organizations collectively before shifting to focus on the Health Care and Education criteria that emerged in 1995. The most recent literature has been spurred by the release of applicant scoring data in recent years.

Further gaps identified were with regard to the timeframe of the research and the data used for the structural equation modeling. While several research literature items are present in the 2011 to current year period utilizing data from companies in Asia, the literature review indicates that there has been little research utilizing data from US companies in the past decade. There have been major revisions to the criteria within that timeframe. A significant portion of the causal model research uncovered in the literature review utilized a survey method to collect data. As was suggested by Pannirselvam and Ferguson (2001), utilization of actual examiner scoring data would lead to much more statistically valid results than self-assessment survey data. Mai et al. (2018) was the first to use a large population of examiner scoring data for model evaluation, but the data used from 1999 to 2006 is antiquated given subsequent changes to the Award Criteria since 2006. While the research literature largely supports all of the causal linkages identified in the Flynn and Saladin (2001) model while using composite data, the linkages did not hold up when examining data that had been segmented by market sector. Most notably,
Mai et al. (2018) found all of the path coefficients as significant when looking at composite data, but found distinct pathway differences when segmenting the data by market sector.
Chapter 3. Research Questions and Hypotheses

The purposes of this section are to define the research questions for this research effort and present a specific set of hypotheses to address the research question.

3.1 Research Question

The Baldrige framework has evolved since its inception in 1988. While this evolution has been well documented and researched, a suitable validation of the framework has remained elusive due to the lack of available data. Although NIST provided historical scoring data in 2011, the scoring data provided was already two years old at that point and was not appropriately analyzed and published until 2018. While the results from Mai et al. (2018) research successfully validated the existing framework, continual changes in the criteria since 2009 represent a research gap in the model validation.

The research available on causal analysis of the Baldrige framework is far from all encompassing. Several gaps exist within the literature. The initial research literature lacked examiner data and was focused on the criteria for manufacturing, service, small business, nonprofit, and government organizations collectively before shifting to focus on the Health Care and Education criteria that emerged in 1995. The most recent literature has been spurred by the release of applicant scoring data in recent years.

Further gaps identified were with regard to the timeframe of the research and the data used for the structural equation modeling. While several research literature items are
present in the 2011 to current year period utilizing data from companies in Asia, the
literature review indicates that there has been little research utilizing data from US
companies in the past decade. There have been major revisions to the criteria within that
timeframe. A significant portion of the causal model research uncovered in the literature
review utilized a survey method to collect data. As was suggested by Pannirselvam and
Ferguson (2001), utilization of actual examiner scoring data would lead to much more
statistically valid results than self-assessment survey data. Mai et al. (2018) was the first
to use a large population of examiner scoring data for framework evaluation, but the data
used from 1999 to 2006 is antiquated given subsequent changes to the Award Criteria
since 2006.

The purpose of this study was to empirically assess the existing Baldrige
framework to determine its validity in context of its continuing evolution (Research
Objective 1). Specifically, this research attempted to empirically validate the causal
relationships suggested by the framework (Research Objective 2) and determine which
criteria items are more important (Research Objective 3). The Baldrige framework was
developed based on knowledge and collaborative input from various Subject Matter
Experts. This research sought to empirically validate these determinations. The most
recent empirical validations in the research literature utilized data from 2009 and later.
This research utilized more recent data than any previous research.

This research made several contributions. First, this research using analytical
techniques that have not been practical for use in most previous studies with the
exception of Mai et al. (2018). Most of the analysis found in research literature utilized
covariance based structural equation modeling. Partial Least Squares (PLS) based
structural equation modeling has more recently began to be used to model complex cause-effect relationships. Second, this research used a data set of applicant consensus scoring data from 2008 to 2017. This data is more recent than anything that has been published in the existing research literature. Third, this research highlighted several observations about the relationships expressed in the model that are practical and relevant to managers seeking to utilize the Baldrige framework to improve their organization as well as researchers seeking to better understand the Baldrige framework and its application.

3.2 Research Model

The next phase of the research effort was to determine which model will be evaluated. Rather than adding to the proliferation of models studied in past research efforts, a review of the existing research was conducted and Flynn and Saladin’s (2001) model shown previously in Figure 2.5 was selected. The rationale for this selection is both obvious and numerous. First, Flynn and Saladin’s “third generation” model coincides with the framework still suggested by and in use today for Baldrige as depicted in the 2015 visual update previously shown in Figure 1.1. Second, this model was conceptually linked and developed based on the evolution of previous Baldrige models. Additionally, this model has been empirically validated (Mai et al., 2018) for antiquated data sets.

3.3 Hypotheses

Flynn and Saldin’s (2001) model (Figure 2.5) posits Leadership as the driver. Leadership is proposed to directly influence Strategic Planning; Customer and Market
Focus; Measurement, Analysis, and Knowledge Management; and Results.

Measurement, Analysis, and Knowledge Management is proposed to directly influence Strategic Planning, Customer and Market Focus, Human Resources (HR) Development and Management, Process Management, and Results. Strategic Planning is proposed to directly influence Customer and Market Focus and HR Development and Management. HR Development and Management are proposed to directly influence Process Management and Results. Customer and Market Focus is proposed to directly influence Process Management. Process Management is proposed to directly affect Results. Stated and designated as hypotheses to be tested:

H1a. Leadership has a direct, positive influence on Strategic Planning
H1b. Leadership has a direct, positive influence on Customer and Market Focus
H1c. Leadership has a direct, positive influence on Measurement, Analysis, and Knowledge Management
H1d. Leadership has a direct, positive influence on Results

H2a. Measurement, Analysis, and Knowledge Management has a direct, positive influence on Strategic Planning
H2b. Measurement, Analysis, and Knowledge Management has a direct, positive influence on Customer and Market Focus
H2c. Measurement, Analysis, and Knowledge Management has a direct, positive influence on HR Development and Management
H2d. Measurement, Analysis, and Knowledge Management has a direct, positive influence on Process Management
H2e. Measurement, Analysis, and Knowledge Management has a direct, positive influence on Results

H3a. Strategic Planning has a direct, positive influence on Customer and Market Focus

H3b. Strategic Planning has a direct, positive influence on HR Development and Management

H4a. HR Development and Management has a direct, positive influence on Process Management

H4b. HR Development and Management has a direct, positive influence on Results

H5. Customer and Market Focus has a direct, positive influence on Process Management

H6a. Process Management has a direct, positive influence on Results
3.4 Summary

The Baldrige framework has evolved since its inception in 1988. While this evolution has been well documented and researched, a suitable validation of the framework has remained elusive due to the lack of available data.

The research available on causal analysis of the Baldrige framework is far from all encompassing. Several gaps exist within the literature. The initial research literature was focused on the criteria for manufacturing, service, small business, nonprofit, and government organizations collectively before shifting to focus on the Health Care and Education criteria that emerged in 1995. Further gaps identified were with regard to the timeframe of the research and the data used for the structural equation modeling. There
has been little research utilizing data from US companies in the past decade. There have been major revisions to the criteria within that timeframe. A significant portion of the causal model research uncovered in the literature review utilized a survey method to collect data. As was suggested by Pannirselvam and Ferguson (2001), utilization of actual examiner scoring data would lead to much more statistically valid results than self-assessment survey data. Mai et al. (2018) was the first to use a large population of examiner scoring data for model evaluation, but the data used from 1999 to 2006 is antiquated given subsequent changes to the Award Criteria since 2006.

Rather than adding to the proliferation of models studied in past research efforts, a review of the existing research was conducted and Flynn and Saladin’s (2001) model shown previously in Figure 2.5 was selected. The rationale for this selection is both obvious and numerous. First, Flynn and Saladin’s “third generation” model coincides with the framework still suggested by and in use today for Baldrige as depicted in the 2015 visual update previously shown in Figure 1.1. Second, this model was conceptually linked and developed based on the evolution of previous Baldrige models. Additionally, this model has been empirically validated (Mai et al., 2018) for antiquated data sets.

Flynn and Saladin’s (2001) model posits Leadership as the driver. Leadership is proposed to directly influence Strategic Planning; Customer and Market Focus; Measurement, Analysis, and Knowledge Management; and Results. Measurement, Analysis, and Knowledge Management is proposed to directly influence Strategic Planning, Customer and Market Focus, Human Resources (HR) Development and Management, Process Management, and Results. Strategic Planning is proposed to directly influence Customer and Market Focus and HR Development and Management.
HR Development and Management are proposed to directly influence Process Management and Results. Customer and Market Focus is proposed to directly influence Process Management. Process Management is proposed to directly affect Results.
Chapter 4. Research Methodology

4.1 Introduction

The purpose of this study was to empirically assess the existing Baldrige framework to determine its validity in context of its continuing evolution. The research method consisted of three distinct phases—inital data manipulation, data analysis, and analysis validation. These phases will be followed by recommendations and conclusion adding to the body of knowledge. This chapter addresses each of these phases while also providing a description of the test data, analysis method selection, and a background and general understanding of the analysis method.

4.2 Description of Test Data

The data utilized for this research consists of blinded site visit scoring data for applicants to a state awards program for the years 2008 to 2017. For the years 2008 to 2013, a percentage score was utilized for each subcategory level. Table 4.1 below shows an example of an individual applicant’s scoring data from 2008 to 2013. Table 4.2 below shows an example of an individual applicant’s scoring data for years 2014-2017 in which the scoring data transitioned to a range scoring system.
Table 4.1 Example Applicant Scoring Data from 2008 to 2013.

<table>
<thead>
<tr>
<th>Applicant Item</th>
<th>Consensus Score</th>
<th>Site Visit Score</th>
<th>Winner?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL01 1.1</td>
<td>60%</td>
<td>75%</td>
<td>No</td>
</tr>
<tr>
<td>PL01 1.2</td>
<td>60%</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>PL01 2.1</td>
<td>55%</td>
<td>70%</td>
<td></td>
</tr>
<tr>
<td>PL01 2.2</td>
<td>55%</td>
<td>65%</td>
<td></td>
</tr>
<tr>
<td>PL01 3.1</td>
<td>60%</td>
<td>75%</td>
<td></td>
</tr>
<tr>
<td>PL01 3.2</td>
<td>65%</td>
<td>75%</td>
<td></td>
</tr>
<tr>
<td>PL01 4.1</td>
<td>45%</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>PL01 4.2</td>
<td>65%</td>
<td>70%</td>
<td></td>
</tr>
<tr>
<td>PL01 5.1</td>
<td>60%</td>
<td>70%</td>
<td></td>
</tr>
<tr>
<td>PL01 5.2</td>
<td>60%</td>
<td>75%</td>
<td></td>
</tr>
<tr>
<td>PL01 6.1</td>
<td>55%</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>PL01 6.2</td>
<td>55%</td>
<td>65%</td>
<td></td>
</tr>
<tr>
<td>PL01 7.1</td>
<td>50%</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>PL01 7.2</td>
<td>60%</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>PL01 7.3</td>
<td>45%</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>PL01 7.4</td>
<td>50%</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>PL01 7.5</td>
<td>50%</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>556</td>
<td>637</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.2 Example Applicant Scoring Data from 2014 to 2018.

<table>
<thead>
<tr>
<th>Applicant</th>
<th>Item</th>
<th>Consensus Score</th>
<th>Site Visit Score</th>
<th>Winner?</th>
</tr>
</thead>
<tbody>
<tr>
<td>EL01</td>
<td>1.1</td>
<td>50%-65%</td>
<td>70%-85%</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>1.2</td>
<td>30%-45%</td>
<td>70%-85%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.1</td>
<td>50%-65%</td>
<td>50%-65%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.2</td>
<td>50%-65%</td>
<td>70%-85%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.1</td>
<td>30%-45%</td>
<td>50%-65%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.2</td>
<td>30%-45%</td>
<td>50%-65%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.1</td>
<td>50%-65%</td>
<td>50%-65%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.2</td>
<td>50%-65%</td>
<td>50%-65%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.1</td>
<td>50%-65%</td>
<td>50%-65%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.2</td>
<td>50%-65%</td>
<td>70%-85%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.1</td>
<td>30%-45%</td>
<td>50%-65%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.2</td>
<td>30%-45%</td>
<td>50%-65%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.1</td>
<td>30%-45%</td>
<td>70%-85%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.2</td>
<td>50%-65%</td>
<td>50%-65%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.3</td>
<td>30%-45%</td>
<td>30%-45%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.4</td>
<td>30%-45%</td>
<td>50%-65%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.5</td>
<td>30%-45%</td>
<td>50%-65%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>30%-45%</td>
<td>50%-65%</td>
<td></td>
</tr>
</tbody>
</table>

The applicant designator refers to the market sector of the applicant and is not taken into consideration in the analysis because of sample size considerations. Some previous research as noted in Chapter 2 allowed for a market comparison, but a sufficient, recent, and suitable set of data was not found for this undertaking. Table 4.3 below shows the distribution of applicant data by year. The Item Numbers correlate to the Leadership; Strategic Planning; Customer and Market Focus; Measurement, Analysis, and Knowledge Management; HR Development and Management; Process Management; and Results Categories.
Table 4.3 Number of Applicants Per Year.

<table>
<thead>
<tr>
<th>Year</th>
<th># of Applicants</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>9</td>
</tr>
<tr>
<td>2009</td>
<td>5</td>
</tr>
<tr>
<td>2010</td>
<td>6</td>
</tr>
<tr>
<td>2011</td>
<td>3</td>
</tr>
<tr>
<td>2012</td>
<td>2</td>
</tr>
<tr>
<td>2013</td>
<td>2</td>
</tr>
<tr>
<td>2014</td>
<td>4</td>
</tr>
<tr>
<td>2015</td>
<td>9</td>
</tr>
<tr>
<td>2016</td>
<td>5</td>
</tr>
<tr>
<td>2017</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>46</strong></td>
</tr>
</tbody>
</table>

As discussed in Chapter 1, applicants are scored by a board of examiners trained in assessing the implementation of the criteria. Individual scores produced by examiners are elevated for discussion and debated at the team level to produce a consensus score for the applicant. High performing applicant organizations at the national level are selected for a site visit by a team of examiners and given a further refined consensus score based on the findings of that visit by the team of examiners. Both the consensus score and site visit score are shown in Tables 4.1 and 4.2 above. The site visit score was used for this analysis.

The scoring data was manipulated to provide a single point score for each category. This was accomplished by taking the combined average of the sub-category percentage and multiplying by the points available at the category level for each year. For the 2014 to 2018 data, an average of the sub-category scoring band was calculated and used as the sub-category percentage for purposes of analysis. For instance, a scoring range of 70%–85% at the sub-category was calculated to be a score of 77.5% for that particular sub-category.
Descriptive statistics of the data are shown in Table 4.4 below. Additionally, the data was analyzed for normality. The p-value results of the Anderson-Darling test are shown in Table 4.5 below. The test compares the empirical cumulative distribution function of the data with the distribution expected if the data were normal. The null hypothesis of population normality is rejected if the observed difference is sufficiently large enough. The descriptive statistics show that winners on average had a score at least one standard deviation higher in every category evaluated. The Anderson-Darling test for Normality confirms the scores are not approximately normally distributed.

Table 4.4 Descriptive Statistics of Examiner Data.

<table>
<thead>
<tr>
<th>Category</th>
<th>All</th>
<th></th>
<th>Non-Winners</th>
<th></th>
<th>Winners</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Mean</td>
<td>Std. Dev.</td>
</tr>
<tr>
<td>Leadership</td>
<td>90.52</td>
<td>11.84</td>
<td>81.56</td>
<td>11.71</td>
<td>95.30</td>
<td>9.14</td>
</tr>
<tr>
<td>Strategy</td>
<td>57.86</td>
<td>10.02</td>
<td>49.94</td>
<td>10.06</td>
<td>62.09</td>
<td>7.34</td>
</tr>
<tr>
<td>Customers</td>
<td>57.10</td>
<td>8.93</td>
<td>50.87</td>
<td>9.61</td>
<td>60.42</td>
<td>6.78</td>
</tr>
<tr>
<td>MAKM</td>
<td>59.28</td>
<td>8.30</td>
<td>52.45</td>
<td>7.74</td>
<td>62.93</td>
<td>6.24</td>
</tr>
<tr>
<td>Workforce</td>
<td>58.60</td>
<td>9.86</td>
<td>52.46</td>
<td>9.37</td>
<td>61.87</td>
<td>8.77</td>
</tr>
<tr>
<td>Operations</td>
<td>58.48</td>
<td>8.90</td>
<td>51.00</td>
<td>9.15</td>
<td>62.48</td>
<td>5.92</td>
</tr>
<tr>
<td>Results</td>
<td>258.41</td>
<td>61.97</td>
<td>191.81</td>
<td>52.51</td>
<td>293.93</td>
<td>30.24</td>
</tr>
</tbody>
</table>

Table 4.5 Anderson-Darling Test for Normality Results.

<table>
<thead>
<tr>
<th>Category</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leadership</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>Strategy</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>Customers</td>
<td>0.032</td>
</tr>
<tr>
<td>MAKM</td>
<td>0.057</td>
</tr>
<tr>
<td>Workforce</td>
<td>0.188</td>
</tr>
<tr>
<td>Operations</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>Results</td>
<td>0.011</td>
</tr>
</tbody>
</table>

4.3 Background on Structural Equation Modeling

Swedish econometrician Herman O.A. Wold "vigorously pursued the creation and construction of models and methods for the social sciences, where 'soft models and soft
data’ were the rule rather than the exception, and where approaches strongly oriented at prediction would be of great value” (Dijkstra, 2009). Wold’s work (Wold, 1982; Wold, 1985) was the basis of Partial Least Squares Structural Equation Modeling. The flexibility and application of structural equation modeling in theory testing and empirical model building has long been recognized in academia (Fornell & Larcker, 1981). Partial least squares path modeling was a result of Wold’s work and later evolved into Partial Least Squares Structural Equation Modeling (PLS-SEM).

According to Hair et al. (2021), PLS-SEM estimates the parameters of a set of equations in a structural equation model by combining principal components analysis with regression-based path analysis. PLS-SEM was developed as an alternative to factor based or covariance based SEM. Covariance based SEM has numerous and restrictive assumptions for establishing a SEM such as data distribution and sample size limitations.

Disadvantages of PLS-SEM include greater difficulty in interpreting the loadings of the independent latent variables which are not based as in common factor analysis or covariances among the manifest independents and the lack of ability to assess significance of without utilization of bootstrapping.

PLS-SEM allows researchers to estimate very complex models with many constructs and indicator variables. Generally, PLS-SEM allows more flexibility in terms of requirements and specifications of relationships between constructs and indicator variables. More specifically, PLS-SEM is ideal for the purposes of this research because the analysis is concerned with the testing of a theoretical framework from a prediction perspective; the structural model is complex including many constraints, indicators, and model relationships; the research uses ratio (scoring) data; and relatively small sample
size (Hair et al., 2021). This flexibility outweighs the disadvantages highlighted above. Furthermore, the model tested in this research only includes one independent variable. Disadvantages to bootstrapping are covered in the "Evaluation of PLS-SEM Results" section of this chapter. In addition to the advantages discussed above, the recent research literature discussed in the literature review almost exclusively uses PLS-SEM. Due to the advantages discussed in this section and to allow for comparison to previous research, PLS-SEM was chosen as the analysis method for this research. Cassel et al. (1999) used simulated data to show the robustness of PLS-SEM against data skewness, multicollinearity, and model misspecification.

There are a wide range of software packages that offer PLS-SEM analysis. SmartPLS was chosen for utilization for analysis in this research effort (Ringle et al., 2015) due to its ease of use, state of the art methods, and abundance of supplemental material available to facilitate utilization.

### 4.4 Principles of Structural Equation Modeling

A path model is a diagram that displays the hypotheses and variable relationships to be estimated in an SEM analysis. Constructs, also referred to as latent variables, are elements in statistical models that represent conceptual variables that researchers define in their theoretical models. Constructs are visually represented as circles or ovals in path models and linked via single-headed arrows that represent the predictive relationships. Indicators are directly measured or observed variables that represent raw data, are represented as rectangles, and are linked to their corresponding constructs through arrows (Hair et al., 2021).
Path models consist of two main elements—the structural model (often referred to as the inner model) and the measurement model (often referred to as the outer model). The structural model represents the structural path(s) between the constructs. The measurement model represents the relationships between each construct and its associated indicators (Hair et al., 2021).

Structural theory indicates the latent variables are to be considered in the analysis of a certain phenomenon and their relationships. Path models are typically laid out from left to right. The latent variables on the left side of the model are independent variables and latent variables on the right side are dependent variables. However, latent variables may in some cases be both independent and dependent. When a latent variable is only an independent variable it is referred to as an exogenous latent variable. When a latent variable is a dependent variable or both a dependent and independent variable, it is referred to as an endogenous latent variable. Endogenous latent variables always have error terms associated with them. The error terms for an endogenous variable represent the sources of variance not captured by any preceding construct in the model. For an exogenous latent variable, the error term is constrained to zero in PLS-SEM. The strength of the relationships between latent variables is represented by path coefficients, which are the results of regressions of endogenous latent variables on their direct predecessor constructs (Hair et al., 2021).
Measurement theory specifies how to measure latent variables. Measurement models consist of two primary types: reflective measurement models and formative measurement models. Reflective measurement models have direct relationships from the construct indicators. They treat indicators as error-prone manifestations of the underlying construct. For reflective indicators, the items should be representative samples of all items of the construct’s conceptual domain and should be highly correlated (Hair et al., 2021).

In formative measurement models, a linear combination of a set of indicators forms the construct. That is to say that the variation in the indicators precedes variation in the latent variable. Strong correlation is not a requirement for Indicators of formatively measured constructs, although correlation can occur and is not necessarily an indication that the measurement model is reflective in nature (Hair et al., 2021).
4.5 Path Model Estimation with PLS-SEM

Unlike factor-based SEM, PLS-SEM explicitly calculates case values for the latent variable as part of the algorithm. The unobservable variables are estimated as exact linear combinations of their empirical indicators such that the resulting composites capture most of the variance of the exogenous constructs’ indicators that is useful for predicting the endogenous constructs’ indicators. PLS-SEM can process both reflectively and formatively specific measurement models without identification issues. Thus, PLS-SEM requires only that each construct is linked with a significant path to the net of constructs (Hair et al., 2021).

PLS-SEM draws on a three-stage approach as shown in Figure 4.2 below. In step #1 of Stage 1, the PLS-SEM algorithm iteratively determines the inner weights and latent variable scores by means of a four-step procedure. In Step #2, proxies are computed for all latent variables by using the weighted sums of its adjacent latent variables. In step #3, new outer weights indicating the strength of the relationship between each latent variable and its corresponding indicator are calculated. These new outer weights and indicators are linearly combined to update the latent variable scores in step #4. After step #4, a new iteration is started which terminates only when the weights obtained from step #3 change only marginally in a given iteration or a maximum number of iterations are achieved (Hair et al., 2021).
Stages 2 and 3 utilize as input the final latent variable scores from Stage 1 for a series of ordinary least squares regressions which produce the final outer loadings, outer weights, and path coefficients. Additionally related coefficients calculated include indirect and total effects, $R^2$ values for the endogenous latent variables, and the indicator and latent variable correlations (Hair et al., 2021).

**Figure 4.2** Basic PLS-SEM Algorithm (Hair et al., 2021).

<table>
<thead>
<tr>
<th>Stage 1:</th>
<th>Iterative estimation of weights and latent variable scores</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step #1</strong></td>
<td>Inner weights (here obtained by using the factor weighting scheme)</td>
</tr>
<tr>
<td>$v_{ji} = \text{cov}(Y_j; Y_i)$ if adjacent, 0 if otherwise</td>
<td></td>
</tr>
<tr>
<td><strong>Step #2</strong></td>
<td>Inside approximation</td>
</tr>
<tr>
<td>$\tilde{Y}<em>j := \sum_i b</em>{ji} Y_i$</td>
<td></td>
</tr>
<tr>
<td><strong>Step #3</strong></td>
<td>Outer weights; solve for</td>
</tr>
<tr>
<td>$\tilde{Y}<em>{jn} = \sum_k \tilde{w}</em>{kj} x_{kn} + d_{jn}$ in a Mode A block</td>
<td></td>
</tr>
<tr>
<td>$X_{kn} = \tilde{w}<em>{kj} \tilde{Y}</em>{jn} + e_{kn}$ in a Mode B block</td>
<td></td>
</tr>
<tr>
<td><strong>Step #4</strong></td>
<td>Outside approximation</td>
</tr>
<tr>
<td>$Y_{jn} := \sum_k \tilde{w}<em>{kj} x</em>{kn}$</td>
<td></td>
</tr>
</tbody>
</table>

Stage 2: Estimation of outer weights, outer loadings, and path coefficients

Stage 3: Estimation of location parameters

4.6 Additional Considerations for PLS-SEM

Research literature contains considerable debate about the restrictions and considerations of using PLS-SEM. This section deals with the distribution assumptions, statistical power, sample size, and goodness of fit considerations when utilizing PLS-SEM.

In 2011, the Journal of Marketing Theory and Practice published “PLS-SEM: Indeed a Silver Bullet.” This article (Hair et al., 2011) was widely considered to have
established the standard approach for estimating models with latent variables and had considerable impact on the dissemination of the PLS-SEM method as evidenced by its massive citation count. Rönkkö and Evermann (2013) and others have aggressively critiqued the article as an advocacy paper. Numerous other scholars have countered the critics on the grounds of analytical evidence (Cook & Forzani, 2020; Rigdon, 2012), simulation findings (Hair et al., 2017; Henseler et al., 2014; Sharma et al., 2023), and conceptual arguments (e.g., Rigdon et al., 2017; Sarstedt et al., 2016). Russo and Stol (2023) observe an increasingly polarized debated among scholars as to the efficacy of PLS-SEM. Petter (2018) summarizes such critiques as “haters gonna hate,” and concludes that scholars should “feel the love for PLS.” Review studies in this research area are numerous and include the categories of Accounting (Lee et al., 2011; Nitzl, 2016), Management (Hair et al., 2012a; Nicole et al., 2016), Knowledge Management (Cepeda-Carrion et al., 2019), and Marketing (Hair et al., 2012b; Sarstedt et al., 2022).

As previously mentioned, factor-based SEM approach has more stringent requirements for distributional assumptions. However, maximum likelihood estimation in factor-based SEM is robust against violations of normality and can avoid normality violations for smaller sample sizes utilizing procedures for parameter estimation. Consequently, data distribution assumptions should not be used as the sole basis for utilization of PLS-SEM (Hair et al., 2021). Data distribution for this research was previously discussed in Chapter 4, “Description of Test Data” section.

According to Hair et al. (2021) PLS-SEM works efficiently with complex models having a small sample size and reviews of SEM in research literature indicates that the average number of constructs per model and the number of indicators per construct tend
to be higher in PLS-SEM versus factor-based SEM. Additionally, PLS-SEM provides solutions when other methods do not converge or develop inadmissible solutions. PLS-SEM can be applied with smaller samples in many instances where other methods fail. Hair et al. (2021) cites the rule of thumb for minimum sample size in PLS SEM as being equal to the larger of 10 times the larger of the largest number of formative indicators used to measure a single construct or 10 times the largest number of structural paths directed at a particular construct in the structural model (Hair et al., 2021). Sample size considerations for this research are discussed in Chapter 5, “Structural Model” section.

PLS-SEM does not have a standard established goodness-of-fit measure, but does have several suggested goodness-of-fit measures: Standardized Root Mean-square Residual (SRMR), the Root Mean Square Residual Covariance (RMS\_{theta}), the Normed Fit Index (NFI), the Non-Normed Fit Index (NNFI), and the Exact Model Test. Hubona et al. (2021) call for the routine use of SRMR. Some discretion is needed in using goodness-of-fit measures in PLS-SEM. Literature casts doubt on whether measured fit is a relevant concept for PLS-SEM, specifically in instances where the model has not proven to have high predictive accuracy (Hair et al., 2021). Hair et al. (2021) suggests that an SRMR of <0.1 is desirable. Goodness of Fit evaluation for this research is discussed in Chapter 5, “PLS Analysis” section.

4.7 Evaluation of PLS-SEM Results

For purposes of this research, the Baldrige model is considered to exclusively contain formative constructs (see previous discussion of formative versus reflective constructs). Evaluation of formative constructs consists of an examination of: (1) the convergent validity, (2) indicator collinearity, and (3) statistically significant and
relevance of the indicator weights. The convergent validity of a formatively measured construct is determined on the basis of the extent to which the construct correlated with a reflectively measured construct capturing the same concept (Hair et al., 2021).

Collinearity is assessed by computing each item's Variance Inflation Factor (VIF) through running a multiple regression of each indicator in the measurement model of the formatively measured construct on all other items of the same construct. The higher the VIF, the greater the collinearity with a rule of thumb being a VIF above 5 being indicative of collinearity (Hair et al., 2021).

Statistical significance and relevance of indicator weights are done via bootstrapping in which a large number of samples from the original data sets is drawn with replacement for model estimation. The standard errors of the bootstrapping samples are then calculated which allows for determining the statistical significance of the original indicator weights. These results can be utilized for computation of t-values and their corresponding p-values as well as confidence intervals of these values. Bootstrapping is typically used with small and/or non-normal data sets where direct computation of path coefficients is not feasible due to the sample size and non-normality. While large, normally distributed samples would be preferred there are currently no data sets available that would meet this criteria. Bootstrapping sometimes produces broad confidence intervals. Bias corrected and accelerated bootstrapping was utilized for this research as it is considered to be the most stable of bootstrap methods available for the analysis. Indicators that have both a not statistically significant weight and a low loading are candidates for deletion from the model (Hair et al., 2021).
The final structural model assessment focuses on learning about the predictive capabilities of the model as indicated by the coefficient of determination ($R^2$), cross-validated redundancy ($Q^2$), and the path coefficients. The coefficient of determination indicates the variance explained in each of the endogenous constructs with a range of 0 to 1 corresponding to no predictive accuracy to complete predictive accuracy (Hair et al., 2021). Chin (1998) proposed tiers of $>0.67$, $>0.33$, and $>0.19$ as Substantial, Moderate, and Weak designations for the coefficient of determination.

Similarly, the cross-validated redundancy ($Q^2$) value is also a measure of the model’s predictive accuracy. This value omits single points in the data matrix, imputes the omitted elements, and estimates the model parameters. Using these estimated inputs, the omitted data points are predicted. This process is then repeated until every data point has been omitted and the model re-estimated. The $Q^2$ value varies inversely with the difference between the predicted and the original values, therefore a higher $Q^2$ value is indicative of more accurate and relevant model. Blindfolding is a sample re-use technique, which systematically deletes data points and provides a prognosis of their original values. The procedure requires an omission distance selection of between 5 and 12 is recommended in the literature (Hair et al., 2021).

Finally, the strength and significance of the path coefficients is evaluated regarding the relationships (structural paths) hypothesized between the constructs. Similar to the formative indicator weights, the significance assessment of path coefficients utilizes bootstrapping to determine standard errors that are utilized to calculate corresponding t-values, p-values, and confidence intervals (Hair et al., 2021). The path coefficients are representative of the resulting standard deviation change in a
dependent construct given a standard deviation change in the independent construct. For the model tested in this research, a higher path coefficient represents a statistically stronger relationship between the Categories and is hence more desirable. Statistical significance is assessed against a Type I error rate of 5% which is typical in all previous similar research evaluated by the researcher.

Multiplicity, the inflation of Type I error rate as a result of multiple hypotheses tests, is a concern with any research that utilizes multiple hypotheses tests. The researcher notes that in testing 15 hypotheses the cumulative probability of at least one type one error (rejecting the null hypothesis of not having category correlation in this research when in fact the null hypothesis is true) utilizing an alpha value of 0.05 calculates to approximately 54%. To account for this issue the researcher utilized the Holm-Bonferroni method (Holm, 1979) to counteract the problem of multiplicity. The Holm-Bonferroni method is the most prevalent correction utilized for multiplicity within the research literature at large. The researcher also notes that none of the previous research noted in the Literature Review for this document mentioned or accounted for multiplicity.

The Holm-Bonferroni method consists of taking the m p-values for the statistically significant hypotheses and comparing them to significance level (alpha) divided by the number of statistically significant hypotheses. The p-values are ordered from least to greatest and the initial p-value is compared to the significance level alpha divided by the number of statistically significant hypotheses. If the first p-value is less than alpha divided by the number of statistically significant hypotheses the result is accepted. The second p-value is compared to alpha divided by the number of statistically
significant hypotheses minus one and again accepted if the p-value is the lesser of the two values. This comparison is completed for all p-values with the number of statistically significant being reduced by an additional one each time as shown in the equation below:

$$p_k < \frac{\alpha}{m + 1 - k}.$$ 

This ensures that the family wise error is at most alpha.

### Table 4.6 Evaluation Parameters and Desired Results.

<table>
<thead>
<tr>
<th>Evaluation Parameter</th>
<th>Desired Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collinearity via Variance Inflation Factor (VIF)</td>
<td>VIF&lt;5 indicative of no collinearity</td>
</tr>
<tr>
<td>Indicator Weight and Statistical Significance</td>
<td>Open to interpretation, but both a non-significant weight and a low loading are needed to justify deletion from the model</td>
</tr>
<tr>
<td>Coefficient of Determination ($R^2$)</td>
<td>&gt; 0.67 Substantial</td>
</tr>
<tr>
<td></td>
<td>&gt; 0.33 Moderate</td>
</tr>
<tr>
<td></td>
<td>&gt;0.19 Weak</td>
</tr>
<tr>
<td>Cross-Validated Redundancy ($Q^2$)</td>
<td>&gt;0.5 Large</td>
</tr>
<tr>
<td></td>
<td>&gt;0.25 Medium</td>
</tr>
<tr>
<td></td>
<td>&gt;0 Small</td>
</tr>
<tr>
<td>Path Coefficients</td>
<td>Higher Absolute Value is Better</td>
</tr>
<tr>
<td>Goodness of Fit via SRMR</td>
<td>&lt;0.1</td>
</tr>
</tbody>
</table>
4.8 Mann-Whitney Test

The Mann-Whitney Test determines whether two independent samples represent two populations with different median values. Test statistical significance indicates that there is a difference between the two sample medians. As a result, one can draw the conclusion that there is a high likelihood that the samples represent populations with different median values. The Mann-Whitney test is ideal for comparing two sample groups that are pulled from a population that is non-normal. As previously discussed in Chapter 4, “Description of Test Data” section, the data utilized for this research does not follow a normal distribution. A t-test for independent samples is more common for this type of analysis but requires the data to be approximately normally distributed (Sheskin, 2020).

4.9 Summary

The purpose of this study is to empirically assess the existing Baldrige framework to determine its validity in context of its continuing evolution. The research method consists of three distinct phases—initial data manipulation, data analysis, and analysis validation. These phases will be followed by recommendations and conclusion adding to the body of knowledge. This chapter will address each of these phases while also providing a description of the test data, analysis method selection, and a background and general understanding of the analysis method.

The data utilized for this research consists of blinded site visit scoring data for applicants to a state Baldrige Award for the years 2008 to 2017.

PLS-SEM allows researchers to estimate very complex models with many constructs and indicator variables. Generally, PLS-SEM allows more flexibility in terms
of requirements and specifications of relationships between constructs and indicator variables. More specifically, PLS-SEM is ideal for purposes of this research because the analysis is concerned with the testing of a theoretical framework from a prediction perspective; the structural model is complex including many constraints, indicators, and model relationships; the research uses ratio (scoring) data; and relatively small sample size. There are a wide range of software packages that offer PLS-SEM analysis.

For purposes of this research, the Baldrige model is considered to exclusively contain formative constructs (see previous discussion of formative versus reflective constructs). Evaluation of formative constructs consists of an examination of: (1) the convergent validity, (2) indicator collinearity, and (3) statistically significant and relevance of the indicator weights. The convergent validity of a formatively measured construct is determined on the basis of the extent to which the construct correlated with a reflectively measured construct capturing the same concept.
Chapter 5. Analysis and Results

5.1 Introduction

As previously discussed in chapter 4, the examiner scoring data was manipulated utilizing Microsoft Excel. This data was input into Smart PLS along with Flynn and Saladin’s 3rd generation model. SmartPLS was manipulated as described in the proceeding sections of Chapter 4 to produce the results provided below. After input of the Structural Model and raw data, the researcher ran the original PLS algorithm for the path model seen below along with Bootstrapping. An additional analysis utilizing Blindfolding was conducted in SmartPLS using the settings described below.

5.2 Structural Model

As discussed in Chapter 4, Flynn and Saladin’s (2001) 3rd generation model was used for our analysis. The majority of previous research that utilizes Structural Equation Modeling performed the analysis in covariance-based SEM. Based on factors outlined in Chapter 4, PLS-SEM was chosen for this analysis. The structural path model is shown in Figure 5.1 below. The rectangular boxes represent the indicators, which in this case are the raw input data. The circles represent the constructs. As previously stated, Hair et al. (2021) cites the rule of thumb for minimum sample size in PLS SEM as being equal to the larger of 10 times the larger of the largest number of formative indicators used to measure a single construct or 10 times the largest number of structural paths directed at a
particular construct in the structural model. Using this rule of thumb we conclude that the minimum sample size for this model utilizing PLS-SEM would be 40 due to there being four structural paths directed at the Results construct. Using Hair et al. (2021) minimum sample size needed based on statistical power analysis with a significance level of 5%, a minimum $R^2$ value of 0.25, and maximum number of arrows pointing at a particular construct of 4 suggests a similar minimum sample size of 41. As discussed in Chapter 4, this research utilizes a data set with a sample size of 46.

![Figure 5.1 Structural Path Model.](image)
5.3 PLS Analysis

The initial PLS-SEM analysis was conducted using a path-loading scheme with a maximum iteration of 300 and a stop criterion of $10^{-7}$. SmartPLS will stop when the outer weights (path weights connecting the indicator variables to the latent variables) do not change more than $10^{-7}$. Path coefficients and $R^2$ values for the individual constructs are shown in Figure 5.2 below. $R^2$ values range from 0.249 for Measurement, Analysis, and Knowledge Management (MAKM) to 0.722 for Strategy. The range of $R^2$ values are consistent with those found in previous research literature and all but one (MAKM) falls in the moderate range as previously defined in Chapter 4. Several of the path coefficients are weak. The path coefficients are representative of the resulting standard deviation change in a dependent construct given a standard deviation change in the independent construct. For instance, the path coefficient from Leadership to Results is 0.250. This would suggest that for a 1 standard deviation increase in the Leadership Score, the Results score would increase by 0.250 standard deviations. Notable weak path coefficients in the model include Strategy to Customers, Workforce to Operations, and Workforce to Results.
Total effects were calculated as part of the analysis and can be seen in Table 5.1 below. Total effects represent the summation of the direct effects (path coefficient) as well as the indirect effects. Indirect effects are the totality of an independent variable on a dependent variable through antecedent dependent variables. For instance, MAKM has a direct effect on Results, but it also has an indirect effect on Results through its effect on Workforce. Total effects highlighted in green were statistically significant at the p=0.05 level.
Convergent validity is the extent to which a measure correlates positively with alternative measures of the same construct. The convergent validity of this model could not be measured due to the fact that only one indicator was used per construct. Single indicator structural equation models are statistically sound and perhaps even more desirable in certain situations (Hayduk & Littvay, 2012). The indicator collinearity was evaluated utilizing the VIF as previously discussed in Chapter 4. All VIFs were found to be below 5 which is indicative of a lack of multicollinearity. Goodness of fit was evaluated by calculation of the Standardized Root Mean Square Residual and found to be 0.099 which is just within the suggested range of less than 0.1.

**Table 5.1** Total Effects by Category.

<table>
<thead>
<tr>
<th></th>
<th>Customers</th>
<th>Leadership</th>
<th>MAKM</th>
<th>Operations</th>
<th>Results</th>
<th>Strategy</th>
<th>Workforce</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.286</td>
</tr>
<tr>
<td>Leadership</td>
<td>0.456</td>
<td>0.499</td>
<td>0.250</td>
<td>0.682</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAKM</td>
<td>0.413</td>
<td></td>
<td>0.541</td>
<td>0.438</td>
<td>0.271</td>
<td>0.168</td>
<td></td>
</tr>
<tr>
<td>Operations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.175</td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategy</td>
<td>-0.125</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.457</td>
</tr>
<tr>
<td>Workforce</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.039</td>
<td>0.092</td>
<td></td>
</tr>
</tbody>
</table>

**Table 5.2** VIF Scores by Category.

<table>
<thead>
<tr>
<th></th>
<th>Customers</th>
<th>Leadership</th>
<th>MAKM</th>
<th>Operations</th>
<th>Results</th>
<th>Strategy</th>
<th>Workforce</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.019</td>
</tr>
<tr>
<td>Leadership</td>
<td>3.002</td>
<td>1</td>
<td></td>
<td>2.989</td>
<td>1.331</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAKM</td>
<td>1.595</td>
<td>1.496</td>
<td>2.168</td>
<td>1.331</td>
<td>1.595</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operations</td>
<td></td>
<td></td>
<td></td>
<td>2.251</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategy</td>
<td>3.597</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.595</td>
</tr>
<tr>
<td>Workforce</td>
<td></td>
<td></td>
<td></td>
<td>1.719</td>
<td>2.73</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.4 Bootstrapping

A bootstrapping procedure was used in order to analyze the statistical significance of the path coefficients. As discussed in Chapter 4, bootstrapping determines the statistical significance and relevance of indicator weights by producing a large number of samples from the original data set, drawn with the replacement for model estimation. These results can be utilized for computation of t-values and their corresponding p-values as well as confidence intervals. The number of subsamples utilized for bootstrapping was set at 5,000 as recommended in the PLS-SEM guide for final analysis for stability of estimates. A complete bootstrapping analysis was conducted. The results are shown in Figure 5.3 below. Non-statistically significant path coefficients at the p=0.05 level are highlighted in red. Recall that most previous research had validated all causal linkages as being statistically significant. The exception to this was research conducted by Mai et al. (2018) where he found inconsistencies when conducting analysis but only when segmenting the data by market sector.
Finally, the cross validated redundancy ($Q^2$) was calculated using a Blindfolding technique. An omission distance of seven was selected (every seventh data point of a construct’s indicator would be removed). This selection was made in order to ensure that the sample size (46) was not a multiple of the omission distance, which would result in full observations being deleted as part of the procedure. The values are shown in Table 5.3 below and are all non-zero positive numbers, which is an indication of suitable and suggestive of retaining all constructs in the model.

Figure 5.3 Statistical Significance of Path Coefficients via Bootstrapping.

5.5 Blindfolding
Table 5.3 Blindfolding Results.

<table>
<thead>
<tr>
<th></th>
<th>( Q^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customers</td>
<td>0.319</td>
</tr>
<tr>
<td>MAKM</td>
<td>0.202</td>
</tr>
<tr>
<td>Operations</td>
<td>0.504</td>
</tr>
<tr>
<td>Results</td>
<td>0.56</td>
</tr>
<tr>
<td>Strategy</td>
<td>0.684</td>
</tr>
<tr>
<td>Workforce</td>
<td>0.254</td>
</tr>
</tbody>
</table>

5.6 Discussion of Hypotheses

H1a. Leadership has a direct, positive influence on Strategic Planning

H1b. Leadership has a direct, positive influence on Customer and Market Focus

H1c. Leadership has a direct, positive influence on Measurement, Analysis, and Knowledge Management

H1d. Leadership has a direct, positive influence on Results

The postulated relationships are partially supported by the data. All path coefficients were statistically significant (see p-values shown in Table 5.4 below) except for the Leadership to Results path which was 0.117.

H2a. Measurement, Analysis, and Knowledge Management has a direct, positive influence on Strategic Planning

H2b. Measurement, Analysis, and Knowledge Management has a direct, positive influence on Customer and Market Focus
H2c. *Measurement, Analysis, and Knowledge Management has a direct, positive influence on HR Development and Management*

H2d. *Measurement, Analysis, and Knowledge Management has a direct, positive influence on Process Management*

H2e. *Measurement, Analysis, and Knowledge Management has a direct, positive influence on Results*

The postulated relationships are partially supported by the data. All the path coefficients were statistically significant with acceptable p-values with the exception of the linkage to Workforce. The path coefficient from Measurement, Analysis, and Knowledge Management to HR Development and Management is still relatively strong at 0.168 but the p-value of 0.254 is high.

H3a. *Strategic Planning has a direct, positive influence on Customer and Market Focus*

H3b. *Strategic Planning has a direct, positive influence on HR Development and Management*

The postulated relationships are partially supported. The path coefficient between Strategic Planning and Customer and Market Focus was negative and somewhat weak (-0.125) with a statistically significantly high p-value (0.644) indicating that the relationship was not statistically significant. The path coefficient between Strategic Planning and HR Development and Management (0.457) was strong and the p-value (0.008) was statistically significant.

H4a. *HR Development and Management has a direct, positive influence on Process Management*

H4b. *HR Development and Management has a direct, positive influence on Results*
The postulated relationships were not supported. The path coefficients were weak and the p-values were not statistically significant.

**H5. Customer and Market Focus has a direct, positive influence on Process Management**

The postulated relationship was not supported. The path coefficient (0.286) showed a moderated relationship with a p-value of 0.146 indicating that it was not statistically significant.

**H6a. Process Management has a direct, positive influence on Results**

The postulated relationship was not supported. The path coefficient (0.175) was not as strong as other path values and the p-value (0.258) was not statistically significant.

<table>
<thead>
<tr>
<th></th>
<th>Customers</th>
<th>Leadership</th>
<th>MAKM</th>
<th>Operations</th>
<th>Results</th>
<th>Strategy</th>
<th>Workforce</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.146</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership</td>
<td>0.031</td>
<td>0.000</td>
<td></td>
<td></td>
<td>0.117</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>MAKM</td>
<td>0.003</td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td>0.004</td>
<td>0.254</td>
</tr>
<tr>
<td>Operations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategy</td>
<td>0.644</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.008</td>
</tr>
<tr>
<td>Workforce</td>
<td></td>
<td></td>
<td></td>
<td>0.811</td>
<td></td>
<td></td>
<td>0.559</td>
</tr>
</tbody>
</table>

The Holm-Bonferroni method was utilized to account for multiplicity and all p-values were found to be acceptable as they are less than the adjusted alpha value suggested by the Holm-Bonferroni method. Results can be seen in Table 5.5 below.
Table 5.5 Holm-Bonferroni Correction Comparisons.

<table>
<thead>
<tr>
<th>k</th>
<th>P-Value</th>
<th>Adjusted Alpha Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.0063</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.0071</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.0083</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>5</td>
<td>0.003</td>
<td>0.0125</td>
</tr>
<tr>
<td>6</td>
<td>0.004</td>
<td>0.0167</td>
</tr>
<tr>
<td>7</td>
<td>0.008</td>
<td>0.025</td>
</tr>
<tr>
<td>8</td>
<td>0.031</td>
<td>0.05</td>
</tr>
</tbody>
</table>

5.7 Summary

As discussed in Chapter 4, Flynn and Saladin’s (2001) 3rd generation model was used for our analysis. SmartPLS was used as the analysis tool due to its ease of use, state of the art methods, and abundance of supplemental material available to facilitate discussion. The majority of previous research that utilizing Structural Equation Modeling performed the analysis in covariance-based SEM. Based on factors outlined in Chapter 4, PLS-SEM was chosen for this analysis.

The initial PLS-SEM analysis was conducted using a path loading scheme. Path coefficients and R2 values for the individual constructs were calculated. R² values range from 0.249 for Measurement, Analysis, and Knowledge Management (MAKM) to 0.722 for Strategy. The range of R² values are consistent with those found in previous research literature. Several of the path coefficients are weak. The path coefficients are representative of the resulting standard deviation change in a dependent construct given a standard deviation change in the independent construct. Notable weak path coefficients
in the model include Strategy to Customers, Workforce to Operations, and Workforce to Results.

Total effects were calculated as part of the analysis and can be seen in Table 5.1. The convergent validity of this model could not be measured due to the fact that only one indicator was used per construct. The indicator collinearity was evaluated utilizing the VIF as previously discussed in Chapter 4. All VIFs were found to be below 5 which is indicative of a lack of multicollinearity. A bootstrapping procedure was used in order to analyze the statistical significance of the path coefficients. Finally, the cross-validated redundancy ($Q^2$) was calculated using a Blindfolding technique. The values are shown in Table 5.3 and are all non-zero positive numbers, which is an indication of suitable and suggestive of retaining all constructs in the model.

The postulated relationships for Hypothesis 1, 2, and 3 were partially supported. The postulated relationships for 4, 5, and 6 were not supported by the data.
Chapter 6. Conclusions and Recommendations

6.1 Introduction

Over the 30+ years of its existence, the Baldrige Performance Excellence Program (originally known as the Malcolm Baldrige National Quality Award Program) has sought to define the basis for high performing organizations. The Presidential award, though originally created for business, has expanded to include applicant categories in education, health care, non-profit, small business, and government organizations.

There have been four major iterations of the Baldrige framework—the original framework developed in 1988, a slight change to the categories in 1992, a significant change to the framework and categories in 1997, and a visual change in representation made in 2015.

The Criteria for Performance Excellence were originally in 1988 a set of statements that evolved into a set of questions in 1999 and subsequent years. The framework was developed anecdotally by industry leaders and has evolved over the years in an effort of continuous improvement and evolution to match the ever-changing organizational environment and landscape.

The Baldrige framework evolution since its inception in 1988 has been well documented and researched; a final validation of the framework is ever elusive due to the due to the continuous evolution of the framework in an effort to account for societal and organizational performance advancement. While the 2011 release by NIST of blinded...
applicant scoring data from 1988 to 2009 has fueled recent research, changes in the criteria since 2009 represent a significant gap in the research literature.

The purpose of this study was to empirically assess the existing Baldrige framework to determine its validity in context of its continuing evolution (Research Objective 1). Specifically, this research attempted to empirically validate the causal relationships suggested by the framework (Research Objective 2) and determine which criteria items are more important (Research Objective 3). The Baldrige framework was developed based on anecdotal evidence and collaborative input of various Subject Matter Experts. This research seeks to empirically validate these determinations.

This research makes several contributions. First, this research utilizes analytical techniques that have not been practical for use in most previous studies. Second, this research uses a data set of applicant consensus scoring data from 2008 to 2017 from a state Baldrige Award (modeled after the Baldrige Performance Excellence Award). This data is significantly more recent than anything that has been published in the existing research literature. Third, this research will highlight several observations about the relationships expressed in the model that are practical and relevant to managers seeking to utilize the Baldrige framework to improve their organization as well as researchers seeking to better understand the Baldrige framework and its application.

An extensive literature search was conducted to examine the body of research related to the evolution of the Baldrige framework and to causal analysis of the Baldrige criteria. The research literature on causal analysis fell into one of three main areas—(1) research done utilizing data from manufacturing, service, small business, nonprofit, and Government organizations (collective referred to as Business/Non-profit) for which the
criteria were originally intended, (2) research done utilizing data from health care organizations and the health care criteria developed in 1995, and (3) research conducted utilizing data from various organizations outside of the United States.

According to Lee et al. (2006), the original Baldrige framework was released in 1988 with seven categories that were worth a total of 1000 points. Though major revisions to the framework occurred in 1992 and 1997 with a minor revision in 2003, the criteria have continued to consist of seven categories with 1000 total points allocated to those categories. Revisions to the criteria are now completed in 2-year cycles beginning in 2007.

Wilson and Collier (2000) were the first to test the causal performance linkages implied by the Baldrige framework utilizing structural equation modeling. They utilized a comprehensive set of 101 questions that were directly tied to the 1995 Baldrige framework and structural equation modeling to evaluate the model. The results provided strong evidence that the Baldrige underlying theory of “leadership drives the system that causes results” is correct.

Pannirselvam and Ferguson (2001) also sought to test the validity and strength of the proposed relationships between the categories in the Baldrige framework. Unlike Wilson and Collier, Pannirselvam and Ferguson utilized actual Baldrige scoring data from the 1993 Arizona Governor’s Quality Award process. The data consisted of 69 organizations who had applied for the award. Additionally, they tested an alternative Quality Performance model.
Evans (2010) utilized blinded applicant total scoring data with descriptive statistical analysis and some basic statistical inference tests to develop insights regarding examiner performance.

Flynn and Saladin (2001) sought to statistically validate utilizing structural equation modeling that the major revisions (1988, 1992, and 1997) in the Baldrige criteria were in fact improvements. Their results indicated that each of the three frameworks included robust relationships and that the subsequent revisions in 1992 and 1997 did in fact improve on the previous framework.

Mai et al. (2018) built off the work of Flynn and Saladin (2001) by taking their model (Figure 2.5), individual review blinded applicant scoring data from 1999 to 2006, and utilized partial least squares structural equation modeling to analyze the data. The model was found to have considerable predictive and explanatory power. The measurement model was confirmed as adequate and similar for all groups; however, the structural coefficients were found to be different among the various industries.

In 1995, separate Baldrige Criteria was designed specifically for health care and educational organizations. Meyer and Collier (2001) set out to prove that the Baldrige Health criteria were in fact an effective set of criteria. The research validated the causal model suggested by the Health Criteria and the authors made several observations from the research about where hospitals should invest their resources to most significantly influence outcomes.

attempted to validate the framework using a survey of 200 manufacturing companies. Interestingly, his research indicated that several of the causal paths previously proven out in studies of American organizations did not hold true for the Chinese.

The research available on causal analysis of the Baldrige framework is far from all encompassing (see summary in Table 6.1 below). Several gaps exist within the literature. The initial research literature was focused on the criteria for manufacturing, service, small business, nonprofit, and government organizations collectively before shifting to focus on the Health Care and Education criteria that emerged in 1995. The most recent literature has been spurred by the release of applicant scoring data in recent years.

<table>
<thead>
<tr>
<th>Study</th>
<th>Year Published</th>
<th>Data Set</th>
<th>Analysis Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilson and Collier</td>
<td>2000</td>
<td>Survey Data</td>
<td>SEM</td>
</tr>
<tr>
<td>Flynn and Saladin</td>
<td>2001</td>
<td>Survey Data</td>
<td>SEM</td>
</tr>
<tr>
<td>Pannirselvam and Ferguson</td>
<td>2001</td>
<td>Applicant Data</td>
<td>SEM</td>
</tr>
<tr>
<td>Evans and Jack</td>
<td>2003</td>
<td>Survey Data</td>
<td>Canonical Correlation</td>
</tr>
<tr>
<td>Evans</td>
<td>2010</td>
<td>Applicant Data (1990 to 2006)</td>
<td>Descriptive Statistics and Basic Statistical Inference Tests</td>
</tr>
</tbody>
</table>

Further gaps identified were with regard to the timeframe of the research and the data used for the structural equation modeling. While several research literature items are present in the 2011 to current year period utilizing data from companies in Asia, the literature review indicates that there has been little research utilizing data from US companies in the past decade. There have been major revisions to the criteria within that
timeframe. A significant portion of the causal model research uncovered in the literature review utilized a survey method to collect data. As was suggested by Pannirselvam and Ferguson (2001), utilization of actual examiner scoring data would lead to much more statistically valid results than self-assessment survey data. Mai et al. (2018) was the first to use a large population of examiner scoring data for model evaluation, but the data used from 1999 to 2006 is antiquated given subsequent changes to the Award Criteria since 2006.

The Baldrige framework has involved since its inception in 1988. While this evolution has been well documented and researched, a suitable validation of the framework has remained elusive due to the lack of available data.

Rather than adding to the proliferation of models studied in past research efforts, a review of the existing research was conducted and Flynn and Saladin’s (2001) model shown previously in Figure 2.5 was selected. The rationale for this selection is both obvious and numerous. First, Flynn and Saladin’s “third generation” model coincides with the framework still suggested by and in use today for Baldrige as depicted in the 2015 visual update previously shown in Figure 1.1. Second, this model was conceptually linked and developed based on the evolution of previous Baldrige models. Additionally, this model has been empirically validated (Mai et al., 2018) for antiquated data sets.

Flynn and Saladin’s (2001) model posits Leadership as the driver. Leadership is proposed to directly influence Strategic Planning; Customer and Market Focus; Measurement, Analysis, and Knowledge Management; and Results. Measurement, Analysis, and Knowledge Management is proposed to directly influence Strategic Planning, Customer and Market Focus, Human Resources (HR) Development and
Management, Process Management, and Results. Strategic Planning is proposed to
directly influence Customer and Market Focus and HR Development and Management.
HR Development and Management are proposed to directly influence Process
Management and Results. Customer and Market Focus is proposed to directly influence
Process Management. Process Management is proposed to directly affect Results.

The purpose of this study was to empirically assess the existing Baldrige
framework to determine its validity in context of its continuing evolution (Research
Objective 1). Specifically, this research attempted to empirically validate the causal
relationships suggested by the framework (Research Objective 2) and determine which
criteria items are more important (Research Objective 3). The research method consists
of three distinct phases—initial data manipulation, data analysis, and analysis validation.
The data utilized for this research consists of blinded site visit scoring data for applicants
to a state Baldrige Award for the years 2008 to 2017.

PLS-SEM allows researchers to estimate very complex models with many
constructs and indicator variables. Generally, PLS-SEM allows more flexibility in terms
of requirements and specifications of relationships between constructs and indicator
variables. More specifically, PLS-SEM is ideal for purposes of this research because the
analysis is concerned with the testing of a theoretical framework from a prediction
perspective; the structural model is complex including many constraints, indicators, and
model relationships; the research uses ratio (scoring) data; and relatively small sample
size. There are a wide range of software packages that offer PLS-SEM analysis.

For purposes of this research, the Baldrige model is considered to exclusively
contain formative constructs (see previous discussion of formative versus reflective
constructs). Evaluation of formative constructs consists of an examination of: (1) the convergent validity, (2) indicator collinearity, and (3) statistically significant and relevance of the indicator weights. The convergent validity of a formatively measured construct is determined based on the extent to which the construct correlated with a reflectively measured construct capturing the same concept.

The initial PLS-SEM analysis was conducted using a path-loading scheme. Path coefficients and $R^2$ values for the individual constructs were calculated. The path coefficients are representative of the resulting standard deviation change in a dependent construct given a standard deviation change in the independent construct. Total effects were calculated as part of the analysis. The convergent validity of this model could not be measured due to the fact that only one indicator was used per construct. The indicator collinearity was evaluated utilizing the VIF as previously discussed in Chapter 4. A bootstrapping procedure was used in order to analyze the statistical significance of the path coefficients. Finally, the cross-validated redundancy ($Q^2$) was calculated using a Blindfolding technique.

6.2 Conclusions

Conclusion 1. The model is still valid.

As shown in Table 6.2 below, all evaluation parameter results were within desired results range based on thresholds established within the research literature.
Table 6.2 Evaluation Parameters, Desired Results, and Achieved Results.

<table>
<thead>
<tr>
<th>Evaluation Parameter</th>
<th>Desired Result</th>
<th>Achieved Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collinearity via Variance Inflation Factor (VIF)</td>
<td>VIF&lt;5 indicative of no collinearity</td>
<td>All VIF&lt;5</td>
</tr>
<tr>
<td>Coefficient of Determination (R²)</td>
<td>&gt; 0.67 Substantial</td>
<td>All are &gt; 0.33 except MAKM (0.249)</td>
</tr>
<tr>
<td></td>
<td>&gt; 0.33 Moderate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;0.19 Weak</td>
<td></td>
</tr>
<tr>
<td>Cross-Validated Redundancy (Q²)</td>
<td>&gt;0.5 Large</td>
<td>3 Large, 2 Medium, 1 Small (MAKM is 0.202)</td>
</tr>
<tr>
<td></td>
<td>&gt;0.25 Medium</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;0.19 Small</td>
<td></td>
</tr>
<tr>
<td>Path Coefficients</td>
<td>Higher Absolute Value is Better</td>
<td>See Framework with Path Coefficients</td>
</tr>
<tr>
<td>Goodness-of-Fit via Standardized Root Mean Square Residual (SRMR)</td>
<td>&lt;0.1</td>
<td>0.099</td>
</tr>
</tbody>
</table>

Statistical analysis of organizations who won the award versus those who did not win the award was conducted utilizing a Mann-Whitney test due to the non-normality of the data. The Mann-Whitney test compares differences between independent groups to determine if the median values differ significantly. This test was conducted on each set of the category scoring data for both winning and non-winning organizations. The results showed that the 95% Confidence Interval of the median difference did not contain zero suggesting that the winning organizations scored statistically significantly higher in every category. This analysis further confirms the criteria and model’s ability to successfully identify and differentiate high performing organizations.

The criteria are frequently used to assess an organization’s performance by identifying its strengths and opportunities for improvement. Many associations model their performance excellence programs on Baldrige. Organizations that utilize the criteria have been empirically shown to improve organizational performance. Hospitals and health care organizations represent a growing area of Baldrige utilization. The results of this research confirm the continued usefulness of the Baldrige framework and criteria to drive organizational performance.
This research confirms the Baldrige approach to organizational performance improvement. Organizations should concentrate on measures of effectiveness and efficiency that are indicative of a suitable approach to the criteria items and then deploying those approaches throughout the organization as appropriate. Once an approach has been developed and deployed institutional learning through evaluation and improvement allow the further improvement of the approach before integration the approach by aligning to the organization’s needs to support organizational goals.

**Conclusion 2: Relationships between categories are evolving and are not adequately captured by the model.**

Although the analysis conducted found that the Baldrige model components are all still relevant; the path coefficients and statistical significance suggest that how the components related to each other may have evolved. Seven of the path coefficients were found to be not statistically significant. This raises questions about whether the framework is appropriately represented in the model tested. Alternatively, previous research has suggested that while the model is appropriate at an enterprise level, it may not be appropriate for all industry types (Mai et al., 2018). While the data set utilized is more recent, its’ size did not lend itself to an analysis by industry type. The continuing evolution of the criteria will continue to drive the need for further validation of the model and criteria.

The banded scoring method focuses on the development of the approach at the lower levels while addressing the basic criteria questions before advancing to more systematic approaches to the detailed criteria that are deployed with evidence of institutional learning and integration with the overall organization strategy. The results
scoring bands follow a similar pattern focused on showing responses to the basic criteria questions before expanding to more detailed criteria items that address customer/stakeholder, market, process and action plan requirements while charting projections of future competitor performance.

Of the seven pathways identified as being not statistically significant, three still have relatively low p-values: the path between Measurement, Analysis, and Knowledge Management to Workforce (0.254); the path between Customers and Operations (0.146); and the path between Operations and Results (0.258). While a p-value of 0.05 or less is desirable the pathways significance for each of these paths still lies at a greater than 15 in 20 chance of being statistically significant.

The remaining three pathways with p-values greater than 0.05 cannot be construed in any other way that to state that the data clearly shows these pathways are not statistically significant in the model. The relationship between Strategy and Customers has a p-value of 0.644 and in addition the path coefficient was found to be negative by this analysis. The remaining two pathways with large p-values are from Workforce to Operations (0.811) and Workforce to Results (0.559). These two paths represent the only two arrows originating out of the Workforce category.

Flynn and Saladin (2001) were the first to produce analysis that questioned the linkage from Strategy to Customers. In addition to having a statistically not significant path coefficient, the path coefficient was also found to be negative. The strategic planning process is intended to analyze current and future customer requirements as well as the competitive market environment that will facilitate filling these requirements. The way organizations obtain customer data has evolved significantly in the last decade with
the advent of indirect tracking of data electronically. The results of this study would suggest that the model has not evolved sufficiently to account for the differences in the relationship between organizational strategy and the focus on customer requirements, needs, expectations, and preferences.

The way organizations manage, engage, and develops their workforce has changed drastically. The results of this research suggest that the Baldrige model has not adequately captured this evolution as both of the linkages coming from the Workforce Category (to Operations and to Results) are not statistically significant (p-values of 0.811 and 0.559 respectively). Chief among these evolution drivers is the widespread use of telework started due to the Covid-19 pandemic and still maintained in many organizations in the present day. Additional considerations that may be driving this evolution are a much more transient workforce with individuals changing jobs and companies on a much more frequent basis, a renewed focus on work-life balance by employees, and the advent of Artificial Intelligence and Machine Learning within the HR field.

**Conclusion 3: Leadership and Measurement, Analysis, and Knowledge Management (MAKM) are the cornerstone of the framework.**

This research does confirm previous research findings the continued importance of both Leadership and Measurement, Analysis, and Knowledge Management as the cornerstone of the Baldrige framework. Their total effects scores (0.589 and 0.580) tower above the total effects scores of other categories. This research confirms their importance as being the two highest non-result category point allocations in the Baldrige scoring system.
Leadership has a positive, direct effect on the categories of Customer Service; Measurement, Analysis, and Knowledge Management; Strategy; and Results. Leadership has the strongest total effect on 5 of the 6 other Categories highlighting the continued emphasis and importance leadership has in the performance of organizations. Flynn et al. (1995) concluded that leadership has a significant effect on the workforce. Adam et al. (1997) show through their research that leadership has a significant impact on training and product quality. Winn and Cameron's (1998) model, based on exploratory factor analysis, also revealed that the main effect of leadership was on the system dimensions, not on the outcome dimensions. Deming (Walton, 1988) was one of the first to point to the tie between leadership and results as well as the interdependency of various structures and functions of the organization in Deming’s 14 points:

1. Create constancy of purpose toward improvement of products and service, with the aim to become competitive and to stay in business, and to provide jobs.

9. Break down barriers between departments. People in research, design, sales, and production must work as a team, to foresee problems of production and in use that may be encountered with the product and service.

Collins (2001) attributed “Level 5 Leadership” as a key aspect in the transformation of good companies to great companies noting that the data in support of this conclusion was overwhelming and convincing. Peters and Waterman (1982) noted the impact of transformational leadership in excellent companies they studied. This research further confirms the importance of leadership in a high performing organization in an ever-evolving society and organizational performance construct.
Measurement, Analysis, and Knowledge Management is an important construct overall. It significantly effects the other non-Leadership categories and provides nearly as significant of a total effect to the Results of the organization. Walton (1986) noted the criticality of decision making that was based on accurate, timely data. Dr. Deming noted the need for data measurement and analysis in all areas of a high performing organization, not just those in technical positions. Mai et al. (2018) also noted the importance of this construct highlighting the empirical support for the criteria evolution in 1997, which posited this category as the foundation for effective performance management. This research further confirms these conclusions with more recent data than previous studies.

6.3 Assumptions and Limitations

Several assumptions and limitations of this research are of value to be mentioned. First, the applicant data utilized is blinded in terms of the information available about the applicant. A detailed analysis of the organizations concerning market sector, organizational structure, quality program/system type, etc. would serve to help draw greater insights into how these factors may/may not have affected their scores.

Second, the applicant data is isolated to organizations in a specific geographical area. No accounting of sample statistics for examining for uniformity and diversity in the previously mentioned factors could be accounted for. Geographically homogeneity is assumed to have had no negative effect on the statistical analyses employed.

Third, as to be discussed in the Future Research section below a larger sample size and from a population of national award level applicants would have been preferable for this analysis.
Fourth, as previously discussed in Chapter 4 PLS-SEM is not without its disadvantages: greater difficulty in interpreting the loadings of the independent latent variables which are not based as in common factor analysis or covariances among the manifest independents and the lack of ability to assess significance of without utilization of bootstrapping.

6.4 Future Research

Future research might consider utilization of statistical analysis techniques to account for reverse causality. While almost all previous research has considered the Baldrige constructs to be connected in a series of recursive, unidirectional path arrows; the third generation Baldrige model included two headed arrows, which allows for the possibility of non-recursive relationships. Non-recursive relationships are challenging to model from a computational perspective, but represent a target of opportunity for future research in this arena.

Non-recursive structural equation models generally take the form of feedback loops. These require the use of instrument variable to achieve model identification. When instrument variables are utilized with categorical indicators, existing research is lacking on how this interaction works (Finch and French, 2015). Research in these analysis methods and techniques will need to mature in order for future research to examine non-recursive relationships within the Baldrige model.

While this research focuses on an analysis at the Enterprise level of the Baldrige model, an analysis of the sector-by-sector differences of the Baldrige model also represents a fruitful future research opportunity. While this sort of analysis has been conducted before (Mai et al., 2018) it utilized an obsolete data set. Should an updated
While Baldrige was originally developed and intended for use in for profit, manufacturing organizations it has spread to be utilized throughout all types of organizations. Manufacturing organizations have evolved significantly in the United States in what has been called the fourth industrial revolution and characterized by emergent communication, information, and intelligence technologies such as additive manufacturing, artificial intelligence, big data analytics, blockchain, cloud computing, and simulation (Bai et al., 2020). These significant changes within the market sector and the attempt of the Baldrige criteria to capture them represent a significant opportunity for future research.

Finally, with the suspension of the award in 2022 for a comprehensive, independent review to take place suggests there are likely to be significant changes forthcoming. These changes which are as of yet unknown will no doubt be fertile ground for future research opportunities. While this research found inconsistencies in the current model it did not seek to propose an alternative model. This too represents an opportunity for future research.

6.5 Summary

The primary goal of this Dissertation was to validate that the Baldrige framework is still valid and relevant today in context of its continuing evolution. The use of recent, actual scoring data addressed a hole in the existing research literature. The model chosen was firmly established in research literature and provides a basis of comparison with
previous research. This research is a major contribution to the research literature with several theoretical implications.

Managers have utilized the Baldrige framework as a basis for organization assessment and improvement since its inception. Past research has validated the Baldrige framework for this utilization at a conceptual level with strong profitability empirically proven. This research provides a validation of the framework utilizing recent scoring data and offers confidence about the validity of the framework construct while producing insights about how those constructs operated in today’s environment.

This Dissertation validated that the Baldrige framework is still valid and relevant today in context of its continuing evolution. Although the analysis conducted found that the Baldrige model components are all still relevant, the path coefficients suggest that how the components related to each other may have evolved. Six of the path coefficients were found to be not statistically significant. The results of this research confirm the continued usefulness of the Baldrige framework to drive organizational performance. This research also confirms the continued importance of both Leadership and Measurement, Analysis, and Knowledge Management as the cornerstone of the Baldrige framework.
References


89


_History : Who we are : The foundation for the Malcolm Baldrige National Quality Award._ (n.d.). https://baldrigefoundation.org/who-we-are/history.html


Russo, D., & Stol, K-J. (2023). Don’t Throw the Baby Out With the Bathwater: Comments on “Recent Developments in PLS”. Communications of the Association for Information Systems, 52 https://doi.org/10.17705/1CAIS.05231


