Insights from the Baldrige Award Item-Level Blinded Applicant Scoring Data

Abstract

In this study, we examine the item-level applicant scoring data of the Malcolm Baldrige Award from 1991 to 2006, which were released by the National Institute of Standards and Technology (NIST) in 2011. By reviewing the evolution and changes in the Baldrige Award Criteria, we aggregate some of the Criteria items so that they are consistent over the entire time period and amenable to analytical evaluation. We investigate trends and inter-industry differences in the Criteria concepts and find support for declining performance of manufacturing industry and steady improvement of non-profit sectors in Baldrige evaluation scores. Moreover, using these national data, we provide further evidence for the validity of the theoretical Baldrige framework using confirmatory factor analysis.

1. Introduction

The Malcolm Baldrige Award (formerly known as the Malcolm Baldrige National Quality Award) has been one of the most powerful catalysts of quality and organizational performance excellence in the United States, and indeed, throughout the world. A recent economic evaluation study of the Baldrige program by the U.S. Department of Commerce [Link and Scott (2011)] concluded that the ratio of net social benefits to the costs of operating the Baldrige program since 2006 was 820:1. The Baldrige criteria have been used extensively by organizations across the globe seeking to improve their management performance and business results. The Baldrige Award is given in eligibility categories of manufacturing, service, small business, education, health care, and not-for-profit.

A limited amount of scholarly research has been performed using the Baldrige criteria. For example, Pannirselvam, Siferd, & Ruch (1998); Wilson & Collier (2000); Meyer & Collier (2001); and Flynn and Saladin (2001) have conducted empirical studies that validated the theoretical models underlying the Baldrige framework as a whole using surveys and other data sources. Other studies [Evans (1997), Ford and Evans (2000), Jack and Evans (2003), Evans (2004), and Stephens, et al. (2005)] investigated conceptual linkages among elements of the criteria and empirical relationships among results items, the conceptual validity of the propositions implied within the strategic planning category, the maturity of measurement practices and their correlation with business results, and Baldrige perceptions and practices in small businesses.

To date, however, very little research has been performed using Baldrige applicant data. This is a result of the Baldrige program's policy to preserve the confidentiality of applicant information and maintain high standards of integrity of the program and Award process. The Baldrige program carefully guards all information related to applicants as proprietary and has never released any individual data. However, using applicant-provided feedback reports, Evans (2010)

and Evans, et al. (2012), have conducted qualitative text-based analyses of themes related to opportunities for improvement to provide insight into performance excellence practices.

In 2009, the National Institute of Standards and Technology (NIST) and the Baldrige program prepared and released a set of blinded applicant scoring data covering 17 years from 1990 to 2006. The data were made public in response to many requests and to facilitate further analysis by interested researchers. The 2009 data set includes the application year, sector, applicant number, Independent Review median score by category, and Consensus Review score by category if applicable. Using a descriptive analysis of the data, supplemented by basic statistical inference tests, Evans (2010) conducted an exploratory analysis of these data to investigate how the performance of applicants had evolved as reflected in examiner category scores, and developed some insights regarding examiner performance. He observed significant differences between large organizations and the Small Business sector, which has consistently lagged the other sectors, a decline in scoring-related performance relative to the Baldrige Criteria within the for-profit sectors, steady increases within Health Care and Education, and improvement in examiner performance as measured by scoring variability.

One of the significant shortcomings of the 2009 dataset, however, was that only category scores were provided. As each Baldrige category consists of multiple items that reflect different concepts and types of management practices, it was difficult to analyze the impacts of specific practices or relationships among them. For example, the Customer Focus category incorporates two items that reflect different concepts: customer knowledge acquisition and customer relationship management. Because each item is scored independently by examiners, category scores alone do not reflect differences in the item scores that may exist.

In 2011, Baldrige released a second version of the blinded data, which included more detailed item scores. The data sets are publicly available and can be obtained from the Baldrige

website, www.nist.gov/baldrige. As with the earlier data release, only median scores are provided. A FAQ document on the Baldrige website notes: "Before 1996, the average Independent Review scores were reported and used during the Award process. Thereafter, median scores were used. …Since each applicant is reviewed by approximately 8 Examiners, the median score describes the applicant more suitably than does the average. For consistency within this dataset, the median Independent Review scores were reported for all applicants. Beginning in 2007, all applicants receive a Consensus Review, which was not the case in prior years. Consequently, data were only released up through 2006 because of this change in the Award process."

This study offers several contributions to the quality management literature. First, we provide a unique review and analysis of the Baldrige Criteria as they have changed over the course of its history corresponding to the time frame of the data that results in consistent constructs that may be analyzed longitudinally. Second, this study provides both longitudinal and cross-sectional view of how award applicants' scores have evolved over time and how they differ across industry sectors. We are able to gain insights on the progress in quality improvement and performance excellence made in these sectors. Third, to our knowledge, this is the first study to comprehensively analyze the national item-level applicant scoring data released by NIST. Finally, we provide validation of the current Baldrige framework using these national Baldrige data. In prior work, the Baldrige model has been evaluated empirically using data from surveys and state quality awards [Pannirselvam, Siferd, & Ruch (1998); Wilson & Collier (2000); Meyer & Collier (2001); Flynn and Saladin (2001)]. The Baldrige evaluation process is designed to be rigorous and objective, and examiners are carefully chosen and more experienced than state-level examiners, providing data having higher integrity and reliability. Compared with state-level data and self-reported surveys, the use of objective and reliable applicant data allows us to make more generalizable conclusions.

The remainder of the paper is organized as follows: Section 2 reviews the Baldrige evaluation process on which the scoring is based, and the evolution of the Baldrige criteria, leading to our research hypotheses. Section 3 describes how we reorganized the items so they can be consistently analyzed across time. Section 4 describes the methodology used in our research. Section 5 presents the main results from our empirical study. Finally, we summarize the conclusions and implications of our research in Section 6.

2. Baldrige Evaluation Process and Criteria Evolution

The Baldrige evaluation process is rigorous, and designed to be objective and tamper-proof against political pressures. In the first stage, each application is thoroughly reviewed by 6 to 10 volunteer examiners chosen from among leading quality professionals in business, education, healthcare, and nonprofits. Examiners evaluate the applicant's response to each examination item, listing major "strengths" and "opportunities for improvement" relative to the Criteria. Strengths demonstrate an effective and positive response to the Criteria. Opportunities for improvement do not prescribe specific practices or examiners' opinions on what the company should be doing, but rather deficiencies in responding to the criteria. Based on these comments, a percentage score from 0 to 100 in increments of 5% is given to each item. This is called the Independent Review (IR) score.

Each examination item is evaluated on approach/deployment or results. *Approach* refers to the methods the company uses to achieve the requirements addressed in each category. The factors used to evaluate approaches include:

- The appropriateness of the methods to the requirements
- The effectiveness of methods, namely, the degree to which the approach is repeatable, integrated, and consistently applied; the degree to which the approach embodies evaluation/improvement/learning cycles; and is based on reliable information and data

- Alignment with organizational needs
- Evidence of innovation

Deployment refers to the extent to which the approaches are applied to all requirements of the item. The factors used to evaluate deployment include:

- Use of the approach in addressing item requirements relevant to the organization
- Use of the approach by all appropriate work units

Results refers to outcomes in achieving the purposes given in the item. The factors used to evaluate results include:

- Current performance levels
- Performance levels relative to appropriate comparisons and benchmarks
- Rate, breadth and importance of performance improvements
- Linkage of results to key customer, market, process, and action plan performance requirements identified in the approach/deployment items and other important factors to the organization

The scoring guidelines by which an application is evaluated can be found in the Criteria document. The scoring guidelines are constructed so that 50 percent represents a very solid and effective approach; few organizations – including Award recipients – receive item scores exceeding 70 or 80. Numerical scores for each examination item are then computed by multiplying the examiner's score by the maximum point value that can be earned. For example, the Senior Leadership item is worth 70 points. Thus, a score of 60 percent would result in 42 points toward the maximum possible point total of 1000.

The examiner team then conducts a consensus process in which they discuss variations in individual scores and arrive at Consensus Review (CR) scores and comments for each item. Next, the Panel of Judges reviews the scores and selects the highest scoring applicants that they believe

have the potential to be recipients for site visits. At this point, an examiner team visits the organization for several days to verify information contained in the written application and resolve issues that are unclear. The site visit report is used by the Panel of Judges to ultimately recommend the recipients.

2.1 Criteria Evolution

The Baldrige Criteria are constantly evolving to reflect, as one former judge stated, "the leading edge of validated management practice." The years 1995 and 1999 were two pivotal years in the evolution of the Criteria. Prior to 1995, the Criteria had a "little q" orientation that was focused on the provincial notion of product quality as reflected in item and category titles such as Item 1.2: Management for Quality, Item 2.1: Scope and Management of Quality and Performance Data and Information, Category 3: Strategic Quality Planning, Category 5: Management of Process Quality, and Category 6: Quality and Operational Results. The 1995 version took a distinct "Big Q" orientation that was directed toward a comprehensive and systems-based business performance model that introduced the notion of a leadership system, broad-based strategic planning and information management, a systems view of process management, and results that included financial metrics. In 1997, the criteria strengthened the concept of alignment among the components of a performance management system by significantly changing the dynamic relationships among categories by relegating Customer Focus as a key element of the "leadership triad" rather than a results-oriented goal, and positioned Information and Analysis as the foundation for effective performance management. Research conducted by Wilson & Collier (2000) and Meyer & Collier (2001) validated the older framework, while Flynn and Saladin (2001) demonstrated the validity of the new framework, which has remained to this day. In 1999, all items were rewritten as questions – a key change that focused the Criteria around a process view. This

year also introduced Criteria for Health Care and Education sectors. Thus, our analysis will segment the time periods from 1990-1994, 1995-1998, and 1999-2006.

Other changes that may have bearing on trends and scoring patterns were the emergence of state and local quality award programs during the 1990s. These programs provided an opportunity for applicants to "test the waters" and only apply to the national level after gaining maturity and strengthening their applications. Thus, we would expect stronger applicants over time.

2.2 Research Focus and Propositions

Of particular concern is the apparent decline in interest in quality and performance excellence among the traditional sectors of manufacturing, service, and small business as reflected by recent changes in the mix of applicants in these sectors for the Award. The first years of eligibility for the newer sectors of health care, education, and not-for-profit showed clear differences in average total scores, indicating early adoption and immaturity in performance excellence practices. Therefore, a key question to investigate is differences among these sectors. Secondly, as we have noted, we are interested in the change of performance over time. Investigating these issues provides insight into the maturity of each sector and their effective implementation of the Baldrige Criteria for performance improvement. This leads to the formal statements of the following hypotheses:

H1: Differences exist among applicant scores for different industry sectors.

H2: Applicants from the manufacturing industry achieved better scores during the earlier year.

H3: Differences between industry sectors have decreased over time (reflecting growing maturity and application of quality management)

The second part of our study is concerned with the reliability and validity of the seven categories in the Baldrige Criteria. As we noted earlier, several empirical studies have validated theoretical models of the Baldrige framework; however, none of them used applicant scoring data at

the national level. Our study can provide a comprehensive evaluation of the Baldrige measurement model using data that has a broader geographic and industry representation. Drawing upon the theory and previous studies of the validity of the Baldrige framework, we propose the following hypothesis:

> *H4: The items in each of the seven categories of Baldrige Criteria measure the theoretical constructs that they were purported to capture.*

3. Reconciling Criteria Items Over Time

One of the challenges in working with Item-level data is the fact that the Criteria have changed significantly over the years. Thus, before attempting any analysis, it is necessary to review the Item-level Criteria changes over the time frame represented by the data and to describe how we dealt with the changes. Specifically, the Item numbers used in the data set often reflect different criteria concepts in different years. For example, "Public responsibility and citizenship" was Item 1.4 in 1990 and 1991, then Item 1.3 through 1996, and finally Item 1.2 from 1997. In addition, some of the embedded elements in the Criteria were not included in the Criteria in all years, or were aggregated differently in different years. For example, many of the detailed elements of the customer focus category (such as customer relationship management, service standards and contact requirements, and satisfaction determination) were independent items in the early 1990s; beginning in 1997, these were aggregated together into one item. Thus, the item scores after 1997 reflected multiple elements that were individually scored in prior years. In order to be able to effectively analyze scoring data across multiple years having different criteria, some aggregation of the data were required, which resulted in the elimination of some criteria elements. In addition, assumptions were made in weighting the data. To understand this, we will review each category, trace its

history from 1990 through 2006, and explain how we defined the Criteria elements used in this research and aggregated the data to reflect these. Table 1 provides a summary of the criteria concepts and items for each year, along with the aggregation of items to identify the constructs used in this research.

When aggregating items, we took the ratio of the sum of the point totals assigned to the items to the sum of the maximum points available of the items for our analyses. For instance, in 1990, the measurement item *Leadership Approach and Deployment* consists of item 1.1, 1.2, and 1.3. An applicant scored 24 out of 30 in item 1.1, 14 out of 20 in item 1.2, and 21 out of 30 in item 1.3, therefore its score *Leadership Approach and Deployment* is calculated by $\frac{24+14+21}{30+20+30} = 0.7375$. Using this method we can adjust the measurement scores using relative consistent scoring criteria across years and industry sectors. The descriptive statistics of the aggregated item scores are presented in Table 2. Table 3 presents the correlation matrix. Not all applicants received Consensus Review scores before 2006, therefore there are fewer observations under Consensus Review.

[Insert Table 1 and Table 2 about here]

[Insert Table 3 about here]

Category 1: Leadership

In 1990 and 1991, the Leadership Category consisted of four items:

1.1 Senior Executive Leadership – focusing on leadership, personal involvement, and visibility in developing and maintaining an environment for quality excellence

1.2 Quality Values – quality values, and how they are communicated and deployed, assessed, and reinforced

1.3 Management for Quality – how quality values are integrated into day-to-day management

1.4 Public Responsibility – leadership in the community and integration of public responsibilities and ethical practices in to policies and activities

In 1992, the basic content of the Quality Values item was integrated into items 1.1 and 1.3, resulting in three items: 1.1 Senior Executive Leadership, 1.2 Management for Quality, and 1.3 Public Responsibility. Some minor changes and enhancements to the Public Responsibility item resulted in a name change to Public Responsibility and Corporate Citizenship in 1993. In 1995, the criteria were streamlined and Item 1.2 was renamed as Leadership System and Organization, however, the general theme of integrating quality values and reviewing performance was maintained. The criteria requirements in the category were generally stable and consistent through 1996.

In 1997, the category was further reduced to two items: 1.1 Leadership System, and 1.2 Company Responsibility and Citizenship. Item 1.1 was renamed Organizational Leadership in 1999, and later as Senior Leadership in 2005; and Item 1.2 was also renamed in 1999 as Public Responsibility and Citizenship. In 2003, it was renamed Social Responsibility, and Organizational Governance was introduced as an Area to Address in Item 1.1. In 2005, two key changes were made in the category. Organizational Governance was moved to Item 1.2, which was renamed Governance and Social Responsibilities, and criteria requirements related to leadership review and assessment of organizational performance was moved to category 4.

Because of the fact that since 1997, all leadership approaches were integrated into one item (1.1), we need to aggregate the data for items 1.1-1.3 for 1990 and 1991, and items 1.1-1.2 for 1992 through 1996 in our analysis. Public Responsibility stands alone as a separate item (1.4 in 1990-91; 1.3 from 1992-96, and 1.2 from 1997 through 2006) and can be analyzed independently. Thus, we defined two constructs that reflect the following elements:

1. Leadership approach and deployment (*Leadership*):

- Senior leadership, personal involvement, visibility in maintaining a performance excellence environment
- Quality value adoption, communication, assessment, and reinforcement
- Integration of quality values into day to day management
- Leadership deployment, review, and assessment,

2. Societal Responsibility (*Societal*):

- leadership in the community
- integration of public responsibilities and ethical practices in to policies and activities

Category 2: Strategic Planning

The Strategic Planning Category was formerly called Strategic Quality Planning in 1990 and consisted of three items: 3.1 Strategic Quality Planning Process, 3.2 Quality Leadership Indicators in Planning, and 3.3 Quality Priorities. Item 3.2 focused on benchmarking and competitive comparisons. Item 3.3 was focused on summarizing key short-term and longer-term plans (priorities), resource commitments, and projections of major changes in competitive position. In 1991, Item 3.2 was moved to the Information and Analysis category and the former 3.3 Quality Priorities was renamed as 3.2 Quality Goals and Plans with basically similar requirements.

In 1992, Item 3.1 was renamed as Strategic Quality and Company Performance Planning Process, and included the strategy development process, deployment of strategic plans, and evaluation and improvement. Item 3.2, Quality and Performance Plans, included previous requirements and projections along with some basic information about key competitive factors (which was subsequently relegated to the Organizational Profile when it was introduced). The 1993 and 1994 criteria were essentially identical.

In 1995, Item 3.1 was renamed Strategy Development, but now addressed how strategies and plans were translated into action plans as a basis for deployment. Deployment was removed

from Item 3.1 and included into Item 3.2, now named Strategy Deployment. This stayed the same in 1996.

In 1997, the entire category was renumbered as Category 2. Item 2.1, Strategy Development Process, included both strategy development and strategy deployment, although as separate areas to address. Item 2.2, Company Strategy, more clearly articulated the requirements of summarizing strategy and action plans and human resource plans, along with performance projections. In 1998, the area to address 2.1b, Strategy Deployment, was dropped from the criteria, but was not included in 2.2 either. Finally, the year 1999 solidified the separation of strategy development (Item 2.1) from strategy deployment (Item 2.2) and as remained constant through 2006.

While the strategy development process has been the focal point of item 3.1 or 2.1 throughout the time frame, many of the requirements that eventually settled in the deployment item (2.2) were included as part of the strategy development item. These changes make it somewhat difficult to clearly define independent constructs during the full time frame and to separate the scoring for individual items between 1990 and 1997. However, these differences are relatively minor and involve such issues as resource allocation, performance measures, and alignment that were not the fundamental focus of the items. The principal constructs boil down to the following:

1. Strategy development (*StrategyDev*)

- Strategy development process

2. Strategy deployment (*StrategyDep*)

- Development of action plans and related human resource plans
- Resource allocation for action plans
- Performance measurement
- Alignment of short- and longer-term action plans with strategic objectives
- Performance projections

Category 3: Customer Focus

In the 1990 and 1991 Baldrige Criteria, customer focus concepts were intermingled with customer satisfaction results. Category 7, Customer Satisfaction, consisted of eight items, two of which were focused on results. The process concepts were clearly separated into different items: Knowledge of Customer Requirements and Expectations, Customer Relationship Management, Customer Service Standards, Commitment to Customers, Complaint Resolution for Quality Improvement, and Customer Satisfaction Determination. In 1992, the Criteria were streamlined somewhat into four process items: Customer Relationship Management, Commitment to Customers, Customer Satisfaction Determination, and Future Requirements and Expectations of Customers. In 1995, further aggregation resulted in three process items: Customer and Market Knowledge, Customer Relationship Management, and Customer Satisfaction Determination. A key change this year was dropping explicit reference to customer commitment (which was a small portion of the category point total).

In 1997, the new framework created an independent results category, and Customer and Market Focus became a separate process category consisting of two items: Customer and Market Knowledge and Customer Satisfaction and Relationship Enhancement. The latter item subsumed most of the detailed items in previous years.

Thus, we identified two principal constructs:

1. Customer and market knowledge (*CustomerKnowledge*)

- Identifying customer groups
- Listening to the voice of the customer
- Using customer information

2. Customer relationship building and satisfaction determination (*CustomerRelation*)

- Relationship building and access mechanisms

- Complaint management
- Satisfaction determination and use

Category 4: Measurement, Information, and Analysis

This category has focused on the selection, analysis, and use of data and information throughout its history. The 1990 Criteria had two items: Scope and Management of Quality Data and Information, and Analysis of Quality Data and Information. In 1991, Competitive Comparisons and Benchmarks was added as a new item. Throughout the 1990s, the scope of data and information became increasingly broader as the Criteria evolved from a pure quality orientation to a more comprehensive organizational framework. In 1999, selection and use of data and information and comparisons and benchmarks were combined into a single item, Measurement of Organizational Performance. In 2001, measurement and analysis were combined into one item, and information management was added to the category. In 2003, knowledge management was added to the information management item, expanding its scope. One concept – review of data and information – was introduced to the category in 1997, but then moved to the Leadership category in 1999 and back again in 2005. This confounds the scoring a bit, but it cannot be cleanly separated.

Two constructs emerged:

1. Measurement and analysis of data and information (*Measurement*):

- Selection, analysis, and use of data
- Comparative data and information
- Analysis and deployment of results

2. Information management (*InformationMgt*) – 2001 and later:

- Data and information availability and reliability
- Hardware and software management
- Management of organizational knowledge

Category 5: Workforce Focus

The role of human resources has been one of the core principles in quality management throughout history. Among all the categories in the Baldrige Criteria, this one has been the most consistent over time. In 1990, the category consisted of five items: Human Resource Management, Employee Involvement, Quality Education and Training, Employee Recognition and Performance Measurement, and Employee Well-Being and Morale. These did not change substantially until 1995, when the category was updated to four items: Human Resource Planning and Evaluation, High Performance Work Systems, Employee Education, Training, and Development, and Employee Well-Being and Satisfaction. In 1997, the criteria were reduced to three items: Work Systems, Employee Education, Training, and Development, and Employee Well-Being and Satisfaction. In 2003 the Employee Education, Training, and Development item was changed to Employee Learning and Motivation.

Three constructs are identified:

1. Work Systems (*WorkSys*):

- Workforce planning, organization, and management
- Performance management, recognition and reward
- Hiring and career progression

2. Education, Training, and Development (*Education*):

- Education design and delivery
- Motivation and career development

3. Employee well-being and satisfaction (*EmployeeWellbeing*):

- Work environment
- Employee support
- Measurement, assessment, and improvement

Category 6: Process Management

In contrast to Category 5, this category has undergone more changes over the years than any other. During the early years of the Baldrige program, the Criteria assumed a strong manufacturing orientation. Thus, in 1990, this category (called Quality Assurance of Products and Services) consisted of seven detailed items: Design and Introduction of Quality Products and Services, Process and Quality Control, Continuous Improvement of Processes, Products and Services, Quality Assessment, Documentation, Quality Assurance, Quality Assessment and Quality Improvement of Support Services and Business Processes, and Quality Assurance, Quality Assessment and Quality Improvement of Suppliers. In 1992, the Criteria were streamlined to five items: Design and Introduction of Quality Products and Services; Process Management – Product and Service Production and Delivery Processes; Process Management – Business Processes and Support Services; Supplier Quality; and Quality Assessment. In 1995, Quality Assessment, which involved evaluation and improvement, was removed, leaving four items: Design and Introduction of Products and Services; Process Management: Product and Service Production and Delivery; Process Management: Support Services; and Management of Supplier Performance. In 1997, the category was reduced to three items: Management of Product and Service Processes, Management of Support Processes, and Management of Supplier and Partnering Processes. In 2001, supplier processes were removed as a distinct area to address, and Business Processes was added. In 2003, the Criteria were reduced to two items: Value Creation Processes and Support Processes. In 2005, Operational Planning was added to the Support Processes item.

Despite all these changes, in one way or another, the category focused on managing what we now call value creation processes and support processes. Most of the items in the early years can be folded into one of these two areas, resulting in two constructs:

1. Value Creation Processes (*ValueCreation*):

- Design
- Measurement
- Management

2. Support Processes (*Support*)

- Non-value creation processes
- Business processes
- Supplier and supply chain processes as appropriate

Category 7: Results

All results in the Baldrige Criteria have not historically been a separate category. In 1990, the Quality Results category included Quality of Products and Services, Comparison of Quality Results, Business Process, Operation and Support Service Quality Improvement, and Supplier Quality Improvement. However, Customer Satisfaction Results and Customer Satisfaction Comparison were incorporated in the Customer Satisfaction category. The category was renamed Quality and Operational Results in 1992 and Business Results in 1995 when financial performance was also included. In 1996, Human Resource Results was added to the category. In 1997, with the new framework, the Business Results category was reorganized to include Customer Satisfaction Results, Financial and Market Results, Human Resource Results, Supplier and Partner Results, and Company-Specific Results (which was renamed Organizational Effectiveness Results in 1999). In 2001, Supplier and Partner Results was dropped as a separate item and integrated into Organizational Effectiveness Results. In 2003, Governance and Social Responsibility Results was added as a new item.

Because of these differences, the constructs that make sense across the time frame could only be defined as:

1. Customer and Product Results (*ExternalResult*)

- Customer-focused outcomes
- Product and service outcomes

2. Process and Operational Results (*InternalResult*)

- Operational performance
- Supplier/partner
- Leadership and social responsibility

3. Financial and Market Results (*FinancialResult*) – 1997 and later

- Financial performance
- Market performance

4. Human Resource Results (*HRResult*) – 1996 and later

- Work systems
- Employee well-being and development

4. Methodology

Since 1995 and 1999 are two pivotal years in the evolution of the Criteria, we divided the data into the three groups of years: 1990-1994, 1995-1998, and 1999-2006. The distributions of industry sectors and time segment are provided in Table 4. Figure 1 displays the smoothed mean plots for the items. The Shapiro-Wilks statistics rejects the normality assumption of the data. Nevertheless the statistical procedures in our analysis are considered relatively robust to less severe non-normality (the kurtoses are within ± 3 and the skewness are within ± 1).

[Insert Table 4 about here]

[Insert Figure 1 about here]

To test hypothesis H1-H3, we conducted two-way ANOVA to study differences in applicant's industry sectors and application year range. If a two-way interaction effect is not

significant for an item, we use one-way ANOVA to test differences on two main effects individually. If the interaction effect is present, we conducted ANOVA on each slice of the data – testing the significance of one effect at each level of another effect. To provide further insights, Tukey's HSD test was used for all pairwise comparisons (summarized by the homogeneous subsets) when F-tests showed significant differences.

We used confirmatory factor analysis (CFA) to test H4 by examine the reliability and validity of the 7 categories in Baldrige framework quantitatively. CFA allows us to assess overall measurement model adequacy as well as individual items. On the theoretic basis of the Baldrige framework, we examined whether the applicants' scores empirically support the proposed seven-factor structure in Baldrige framework: (1) Leadership; (2) Strategic Planning; (3) Customer Focus; (4) Measurement, Information, and Analysis; (5) Workforce Focus; (6) Process Management; (7) Results.

5. Results

5.1 Differences in applicant scores over time and across industries

Table 5 presents results for the ANOVA analysis (.05 significance). The results for the items that do not have interaction effects, reported in Panel A, provided limited evidence for H1. Panel A reveals that sector differences are not significant among the defined constructs within two categories: Leadership and Workforce Focus. Although we also found time period difference in these items, namely the significantly higher scores for some items in 1990-1994, at least part of the reason for these differences is due to the over represented number of applicants from manufacturing sectors, which had higher scores during the early years of Baldrige award.

Panel B and C of Table 5 summarizes the results of multiple comparison of sector differences and time differences for those items with two-factor interaction effects. The results provide support

for H2. Between 1990 and 1994, applicants from the manufacturing sector received significantly higher scores in the Baldrige evaluation process when compared with applicants from service sector and small business sector. The gap is particularly evident between manufacturing applicants and small business applicants, as the analysis finds significant differences in all seven items.

Our analysis reveals that service and small business applicants are catching up in most of the categories. Compared with the results of H2, small business applicants showed the most obvious improvement. For applicant scores in 1999-2006, the statistically significant difference between small business and manufacturing in earlier years has disappeared for all items. Therefore, H3 is supported in our study.

In addition to the surge of small business, we noticed the declining trend of manufacturing applicants' scores. When comparing the rank with other industry sectors, manufacturing sector had a significant lead during 1990-1994, but the lead diminished as time progressed. Panel C also offers support for the claim, as in all of items the mean scores of manufacturing applicants are significantly higher in 1990-1994 than those in 1999-2006.

In line with Evans (2010)'s finding, health care, education, and non-profit applicants started participating in Baldrige award later than the traditional for-profit sectors and lagged behind. Panel B shows that health care applicants have significant lower mean scores across all items during 1995-1998, and education along with non-profit applicants are in the group with lowest mean scores for 6 out of 9 items during 1999-2006. Due to the structure of our ANOVA, we were not able to test whether education and non-profit applicants are improving, but we found significant increases in mean scores of health care applicants in 1999-2006 compared with 1995-1998.

[Insert Table 5 about here]

5.2 Validation of Baldrige Award Criteria Measurement Model

Overall, our analysis provides strong support for validity of the Baldrige measurement model (H4). As shown in Table 6, the overall fit of the CFA model is satisfactory. Although the $\chi 2$ is significant at 0.01 level, it is known to be inherently biased when the sample size is large (Shah & Goldstein, 2006). We reported several other commonly used fit indices. Both absolute and incremental fit indices suggest a high goodness of fit between the Baldrige constructs and the aggregated items.

[Insert Table 6 about here]

We used Cronbach's alpha (Cronbach, 1951) to assess reliability of the constructs. The alpha values are above the recommended value of 0.70, supporting reliability high internal consistency of the constructs (O'Leary-Kelly, 1998). Convergent validity implies that all variation in items of each Baldrige category can be accounted for by the theoretical concepts that the items are intended to measure plus random error (Bagozzi, Yi, & Phillips, 1991). Convergent validity has been achieved for our model as all factor loadings are close to 1 and significant at $p < 0.01$, indicating the items adequately reflect the same corresponding concept. In addition, we examined the average variance extracted (ρ_{vc}) as a more conservative measure for discriminant validity (Fornell & Larcker, 1981). All ρ_{vc} are greater than 0.5, as Table 7 shown. Therefore we conclude that the convergent validity of the Baldrige constructs is adequate.

We performed pair-wise chi-square difference test to assess discriminant validity (Bagozzi & Phillips, 1982). The null hypothesis is that there is no discriminant validity among constructs. For each pair of the constructs we constrain inter-factor correlation to be 1 and compare the chi-square values to the original model. A significant increase in the chi-square value indicates that there is sufficient evidence to reject the null hypothesis. The chi-square difference ranged from 12.82 to

453.97 ($P < 0.01$), which demonstrates that discriminant validity exists between pairs of constructs. Two alternative criterions for discriminant validity: average variance extracted being greater than squared correlations between any two constructs, and Cronbach's alpha being greater than correlations between any two constructs, are also satisfied (Kim, Kumar, & Kumar, 2012).

[Insert Table 7 about here]

6. Discussion and Conclusions

In this paper we present the first comprehensive analysis of the item-level applicant data released by NIST for the Baldrige Award. Our results showed differences among different industry sectors in the defined constructs in all categories except leadership and workforce focus. This is perhaps not very surprising, as leadership and workforce (human resource) issues have been traditional components of management practice in all organizations for over a century, while the other categories (customer focus, strategic planning, process management, and information/analysis) developed and matured much more recently as total quality management concepts evolved and became popular. The differences in culture and management practices among different sectors help explain these differences.

Over time, we observed a declining trend of manufacturing applicants' scores. The results provide support for the hypothesis that applicants from the manufacturing industry achieved better scores during earlier years, suggesting that a focus on performance excellence practices is diminishing. The recent decline in the number of Baldrige applicants appears to suggest the same thing. The criticality of manufacturing and these trends have recently led to the development of a Baldrige-sponsored panel to investigate this issue.

The maturity and application of quality management appears to have leveled off among sectors as evidenced by support of hypothesis 3. Our analysis reveals that service and small business

applicants have caught up in most of the categories. Sharing of best practices – one of the principal objectives of the Baldrige program – has apparently been achieved, particularly across sectors.

Finally, using confirmatory factor analysis, we found strong evidence of reliability, convergent and discriminant validity in the Baldrige framework, consistent with other research that used survey or state-level data.

One of the limitations of this study is the subjective definition of the constructs, necessitated by the Criteria changes over the years. The assumptions used in weighting the scoring data may not have accurately reflected the specific nature of the Criteria on which the scores were based. Nevertheless, the constructs defined are consistent and logical, and provide a means to reasonably develop key insights from the data.

References

Bagozzi, R. P., & Phillips, L. W. (1982). Representing and Testing Organizational Theories: A Holistic Construal. *Administrative Science Quarterly*, *27*, 459-489.

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly*, *36*(3).

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Curcovic, S., Melnyk, S., Calantone, R., & Hardfield, R. (2000). Validating the Malcolm Baldrige National Quality Award framework through structural equation modeling. *International Journal of Production Research*, *38*(4), 765–791.

Evans, J.R (1997). "Critical Linkages in the Baldrige Criteria: Research Models and Educational Challenges," *Quality Management Journal*, 5(1), 13-30.

Evans, J.R. (2004) "An Exploratory Study of Performance Measurement Systems and Relationships with Performance Results," *Journal of Operations Management* Vol. 22, Issue 3, June 2004, pp. 219-232.

Evans, James R. (2010) "Organizational Learning for Performance Excellence: A Case Study of Branch-Smith Printing Division," *Total Quality Management and Business Excellence*, Vol. 21, Numbers 3-4, 225-243.

Evans, James R. (2010), "An Exploratory Analysis of Preliminary Blinded Applicant Scoring Data

from the Baldrige National Quality Program," *Quality Management Journal*, Vol. 17, Issue 3, 35-50.

Evans, James R., M.W. Ford, S. S. Masterson, and H.S. Hertz (2012) "Beyond Performance Excellence: Research Insights form Baldrige Recipient Feedback," *Total Quality Management and Business Excellence* (accepted)

Flynn, B. B. and B. Saladin (2001), "Further Evidence on the Validity of the Theoretical Models Underlying the Baldrige Criteria," *Journal of Operations Management*, 19, 617-652.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling*, *6*(1), 1-55.

Jack, E. and J.R. Evans (2003), "Validating Key Linkages in the Baldrige Performance Excellence Model" *Quality Management Journal*, Vol. 10, Issue 2, pp. 7-24.

Kim, D.-Y., Kumar, V., & Kumar, U. (2012). Relationship between Quality Management Practices and Innovation. *Journal of Operations Management* (accepted).

Link, A.N. and J.T. Scott (2011), "Economic Evaluation of the Baldrige Performance Excellence Program," Planning Report 11-2, U.S. Department of Commerce, National Institute of Standards and Technology.

Meyer, S.M. and D. A. Collier (2001). "An Empirical Test of the Causal Relationships in the Baldrige Health Care Pilot Criteria," *Journal of Operations Management*, 19 (4), July, 403-426.

O'Leary-Kelly, S. W. (1998). The empirical assessment of construct validity. *Journal of Operations Management*, *16*(4), 387-405.

Pannirselvam, G.P., S.P. Siferd, and W.A. Ruch, (1998). "Validation of the Arizona Governor's Quality Award criteria: A test of the Baldrige criteria," *Journal of Operations Management*, 16, 529-550.

Shah, R., & Goldstein, S. M. (2006). Use of structural equation modeling in operations management research: Looking back and forward. *Journal of Operations Management*, *24*(2), 148-169.

Stephens, P.R., J.R. Evans, and C.H. Matthews (2005), "Importance and Implementation of Baldrige Practices for Small Businesses," *Quality Management Journal*, Vol. 12, Issue 3, pp. 21-38.

Wilson, D.D. and D. A. Collier (2000). "An Empirical Investigation of the Malcolm Baldrige National Quality Award Causal Model," *Decision Sciences*, 31 (2), Spring, 361-390.