Exploratory Analysis of the Malcolm Baldrige National Quality Award Model

Anyama Tettey[1]
Sampson Gholston, PhD[1]
Bryan Mesmer, PhD[1]

[1]*University of Alabama in Huntsville*
aht0005@uah.edu; gholsts@uah.edu; blm0027@uah.edu

Abstract

This paper conducts exploratory statistical analysis to assess trends in scores for the Malcolm Baldrige National Performance Excellence Award Model (MBNQA). The analysis identifies significant differences and similarities across sectors of examiner scores for the award program. The paper makes use of real consensus and site visit scores data collected over 11 years period, 2007-2017, from a States' Performance Excellence Award program to conduct the analysis. There are 2 parts to the study. In the first part, the authors use descriptive statistics and various univariate parametric procedures to observe differences/ similarities in variability and mean scores over the period of the study. The results show that the variability in the scoring approaches across sectors were not statistically different. The mean consensus scores, however, were statistically different from the mean site visit scores. The second part of the study uses multivariate analysis to assess any significant difference in means between award winners and non-award winners for the consensus and site visit scores. The results show a significant difference between award winners and non-award winners for all 7 category scores using the site visit data but a different pattern for the consensus scores. The study confirms examiner consistency and reliability, as well as the need for all applicants to be given a site visit tour during the award program.

1. Introduction

The Malcolm Baldrige Performance Excellence Model has proven to be a valuable tool for use by organizations interested in improving overall Organizational Performance. At a time when the influx of Japanese automobiles and electronics into the American market was seemingly making American companies less competitive, the MBNQA was formed in 1987 to address the ensuing problem. The goal of the MBNQA Improvement Act of 1987 is thus to enhance the competitiveness of U.S. businesses (Link & Scott, 2012).

Since its inception, its usefulness has become apparent to quality experts and other parties interested in the growth of American businesses, and a form of it has been rolled out by most states in the US and all industrialized countries including Japan (Townsend & Gebhardt, 1996). Studies have shown numerous benefits of the award program to the US economy. In 2011, Link et. al, showed that the economic and social benefit to cost ratio had increased considerably from 207-1 for an earlier study they did in 2001 to 820-1 in 2012 (Link & Scott, 2012). Their 2012 study used the ratio of social benefits to social costs for the population of all Baldrige applicants from 2007-10 to come up with the 820:1 ratio in 2012. The authors, however, acknowledged 351:1 as a more conservative ratio since it is more representative of the 45 (16.5%) applicants that responded to their survey.

The importance and usefulness of the model further resulted in the emergence of local and state

quality award programs. As of 2016, there were 30 independent Baldrige-based state and regional award programs covering nearly all 50 states (NIST, 2016b).

Despite the success stories for the MBNQA, data used for most research work relating to the Baldrige award has had some obvious limitations. Up until 2009, no data relating to the Baldrige program and scoring process had been made available to the general public due to the sensitive nature of the program. In 2009, however, NIST released the blinded Baldrige scoring data covering years from 1990 to 2006.

Within the period from 1990 to 2006 however, there were some major revisions to the Baldrige framework that makes it unreasonable to identify any observable trends within that stretch of time effectively (Link & Scott, 2012). Within this period the criteria evolved from a quality assurance focus to an overall Performance Excellence focus (Lee, Zuckweiler, & Trimi, 2006) and it was not until 1999 that the scope of the award was increased to include healthcare and education.

This research uses more recent scoring data having both consensus and site visit scores of 59 applicants from a states' MBNQA program from 2007 to 2017. The state has an extensive corps of more than 150 examiners, 40-member board of directors and a panel of 5 judges that make recommendations for award recipients to the state's governor.

The idea of the state programs is to ensure that applicants attain a high level of maturity before applying at the national level. Thus, applicants to the national award program are required to have first applied to the state program unless they apply for a waiver of the standard requirement of first achieving a top-level alliance for Performance Excellence Award.

The national Baldrige scoring procedure involves a joint review of submitted applications by examiners and the subsequent award of scores referred to as consensus scores based on the applicants' achievement in 7 areas known as the Baldrige Criteria for Performance Excellence. At the national level, this process helps to decide whether the applicant will go to the next step of receiving a site visit and subsequently a site visit score. The state award programs closely mimic the national programs with minor modifications, and this is one of such modifications with this states' program where all applicants receive a site visit.

Using the states' data this study tests the following propositions:

a. There is significant difference in scores across sectors, and between site visit and consensus scores
b. There is significant difference in variability across sectors, and between site visit and consensus scores
c. There is significant difference in mean scores to reflect the minor changes in criteria or improvement in applicant performance over the years
d. The significance of each of the seven criteria is the same for determining award winners for the 2 scoring approaches

The study concludes with findings that could lead to a useful assessment for all stakeholders involved in or interested in the Performance Excellence agenda in the US.

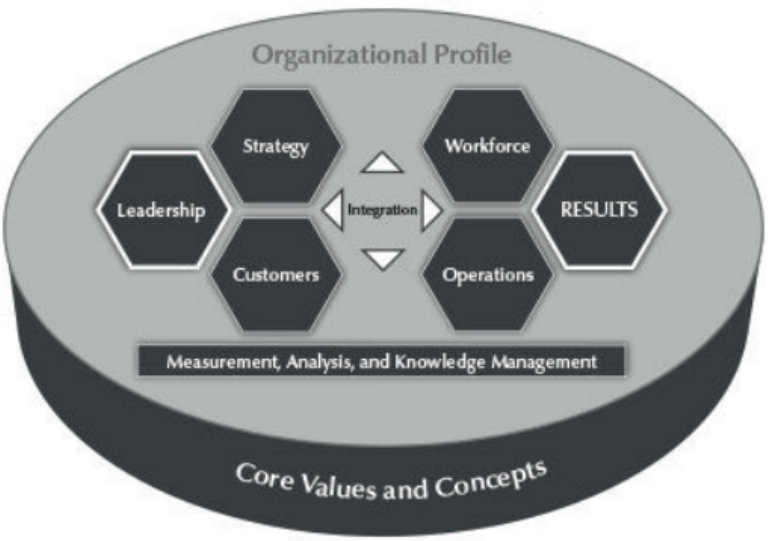
2. Overview of the Baldrige Excellence Framework and award program

The Malcolm Baldrige National Quality Award (MBNQA) was instituted in 1987 (Steeples, 1994) by the US congress to award United States organizations that have attained a high level of performance excellence based on an assessment by independent examiners.

The importance and usefulness of the award process further resulted in the emergence of local and state quality award programs that ensured that applicants attained a high level of maturity before applying at the national level. Consensus scores are given to applicants based on their answers to questions that represent 7 aspects of the organization. These 7 aspects are used as

benchmarks to award scores for applicants in the areas of Leadership, Strategy, Customers, 'Measurement Analysis and Knowledge Management' (MAKM), Workforce, Operations and Results (ASQ, 2018). At this state level program, all applicants receive site visits by examiners after their applications have been thoroughly reviewed. At the national level, however not all applicants graduate to the stage of receiving site visits, only a selected few progress to that stage. Again, at the national level, there are 6 different eligible categories for manufacturing, service, small business, nonprofit, healthcare, and education. At the state level whose data is being used for this study however, the same framework and criteria is used for all sectors. One of the motivations for this research is thus to look at any underlying differences in scores that exist between sectors and the two scoring approaches used.

Figure 1 [illegible] both the national and state a

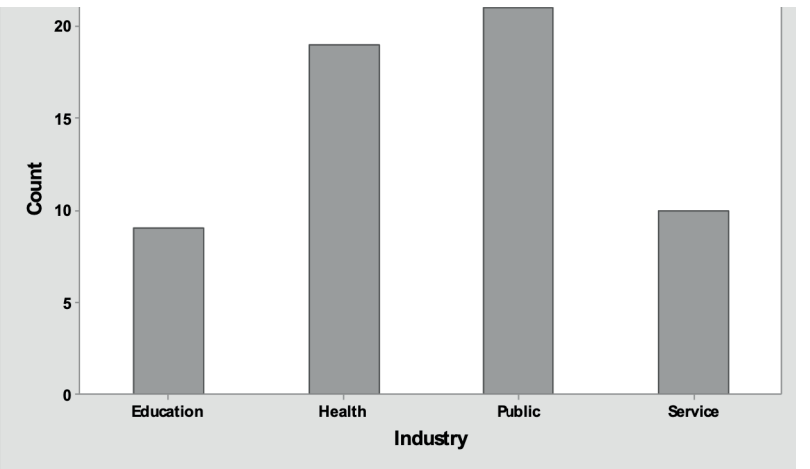


This state level award program has applicants categorized as healthcare small, healthcare large, service small, service large, education small, education large, public small and public large within this period. For the purposes of this study small and large sectors have been grouped together. The categories are weighted with maximum attainable as: 140, 100, 100, 100, 100, 100 and 360 for each of the categories Leadership, Strategy, Customers, 'Measurement Analysis and Knowledge Management' (MAKM), Workforce, Operations and Results respectively, a total maximum attainable score of 1000. After a thorough review of the applicant's submissions, the total for the 7 categories becomes the consensus score for the applicant. Each applicant then receives a site visit by assigned examiners who visit the organization's premises. Applicants are rescored after the visit and an applicant is nominated for an award based on the performance. At the national level, up to 18 awards may be given annually across the six eligibility categories. At this State level program however depending on the examiner assessments, there can be as many winners as the number of applicants. Further details on the scoring guidelines and the award national procedure are available at (NIST, 2016a).

3. Descriptive statistics

This section carries out basic descriptive statistics of the data. A brief description of the data is

shown in figure 2 below. There are a total of 9 applicants from the educational sector and 19, 21, and 10 applicants respectively from the healthcare, public, and services sectors.



Figures 3 and 4 provide a visual assessment of the trends in mean scores and the variability observed between consensus scores and site visit scores. It can be seen from Figure 3 that the mean scores for the site visits are always higher than those for the consensus scores.

Figure 3 also shows that the pattern across sectors is the same for each of the 2 scoring approaches with the educational sector scores having the lowest means whilst the healthcare sector consistently has the highest means.

Figure 4 shows a box plot of the scores for the four sectors compared over the 2 scoring approaches. The main takeaway from this figure is that there seems to be relatively more variability in the site visit scores compared to the consensus scores for the same sector.

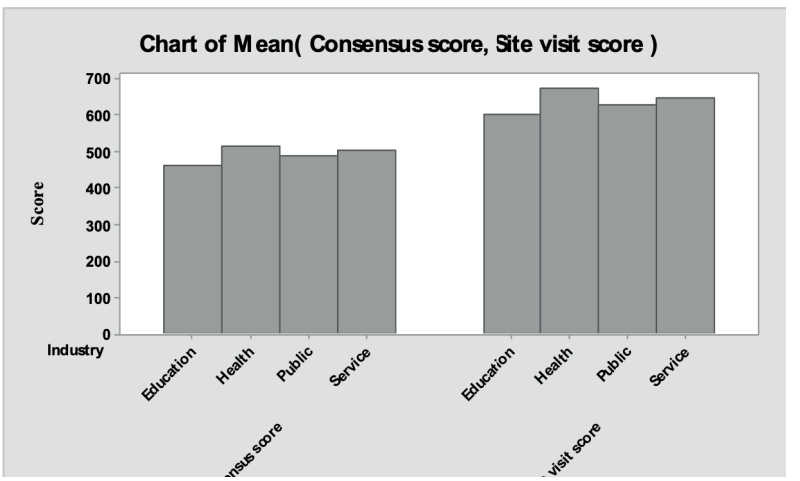


Figure 3. Comparison of mean scores across sectors for the 2 scoring approaches

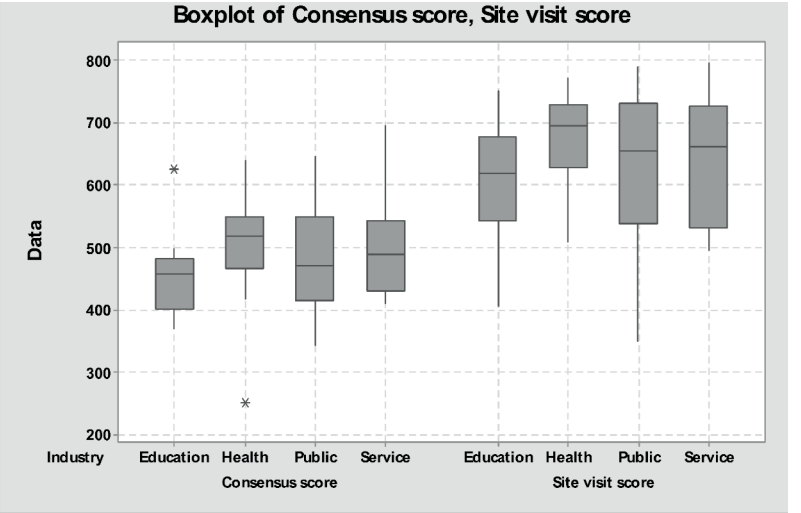


Figure 4. Variability and mean scores for the 2 scoring approaches compared

The authors went ahead to test if the observed differences in figures 3 and 4 were statistically significant using inferential statistical tests. The results are discussed in section 4 below.

4. Inferential analysis

This section uses a 2-sample t-test, one-way ANOVA and posts ANOVA techniques for the analysis. The null and alternate hypothesis of the 2-sample test has $H_0: \mu_{(\text{consensus})}=\mu_{(\text{site visit})}$ and $H_1: \mu_{(\text{consensus})} \neq \mu_{(\text{site visit})}$ respectively. With a p-value < 0.0001 (table 1), we conclude that there is a difference in means for the 2 scoring paradigms. A one-way ANOVA was performed to determine if there is a difference in mean scores across sectors. This was done separately for both scoring approaches. The p-values were 0.465 and 0.351 respectively for the consensus and site visit scores. Thus, in both cases, the conclusion is that the observed differences across sectors (seen in Figure 3) are not statistically significant.

Due to its robustness to the normality assumption, the Levene's test was used for the test of significance of the differences in variability across the 4 sector scores. The test results in Table 1 did not reject the assumption of equal variances. We thus conclude equal variances across all sectors for the consensus and site visit scores within the period 2007-2017. The difference in variability between the site visit and consensus scores was also not significant for both the Levene and Bonett tests as seen in Table 1. Thus all the observed differences in figure 4 were not statistically significant upon investigation.

Table 1. Tests of equality of mean scores and variability within scores

Test method	Result	Equality of the 2 groups (Site visit and consensus scores)	Test method P-value
95% CI for difference (Mean of consensus - Site visit score)	Interval : -183.4 to -11		
		Levene test of equality across sectors	
			Consensus 0.714

The third proposition alludes that as minor criteria changes occurred in the years 2007, 2009 and 2013 it reflected in changes in the mean scores over those periods. One-way ANOVA is used to test the equality of means over these periods after a residual analysis showed model adequacy. The null hypothesis is: H_0: $\mu_{(07)}$=$\mu_{(08-09)}$=$\mu_{(10-13)}$=$\mu_{(14-17)}$. The ANOVA results showed a significant p-value and therefore the rejection of the null hypotheses at the 0.05 significant level. Table 2 summarizes the results.

Table 2. ANOVA and Fisher pairwise comparisons across years

Year	Number of applicants	Consensus			Site visit		
		Mean scores	P value	Fisher comparison	Mean scores	P value	Fisher comparison
2007	13	437	0.042	B	573	0.048	B
08--09	14	495		AB	641		AB
10--13	13	517		A	671		A
14--17	19	514		A	670		A

The post ANOVA technique, 'Fisher pairwise comparisons' showed mean differences from 07-10 for both scores. There were no significant differences in mean scores between 2008-2017. Under the column titled 'Fisher comparison' in table 2, years with the same letter show equal means. There is a slight increasing trend from 2007 to 2008 for both the site visit and consensus scores.

5. Inferential analysis using MANOVA

This part of the analysis uses MANOVA to test the hypothesis of equal mean vector scores between award winners and non-award winners for both the consensus and site visit scores. Here we refer to an applicant's category score for all 7 categories, Leadership (x1), Strategy (x2), Customers (x3), MAKM (x4), Workforce (x5), Operations (x6), and Results (x7), (X_j' =x1; x2; x3; x4; x5; x6; x7), as a vector score.

The general case of the MANOVA model is for comparing 'g' (2 or more) population means, with each population consisting of 'p' (more than one) responses (Johnson & Wichern, 2002). We consider the simplest case of comparing 2 population mean vectors: Winners and non-winners, to statistically determine if the 'seven category scores/ responses vary across different levels of the factor. This is done separately for the consensus score and site visit scores. We test the hypotheses that: H_0: $\boldsymbol{\tau_1} = \boldsymbol{\tau_2} = \cdots = \boldsymbol{\tau_g} = \mathbf{0}$, where $\boldsymbol{\tau}_\ell$ represents the ℓ^{th} scoring difference; estimated as the ℓ^{th} population mean vector minus the overall sample mean vector.

The statistic Wilks' Lambda corresponds to the equivalent of the F-test in the univariate case. A significant result is followed up by post hoc tests to determine what accounts for the differences. As with all parametric studies, some assumptions needed to be satisfied to pave the way for proper MANOVA analysis.

The following assumptions about the structure of the data for one-way MANOVA were satisfied: The random samples from the 2 populations are independent. The test for the equality of covariance matrix for the 2 populations failed to reject the null hypotheses of equal covariance matrix at a significance level of 0.05. Finally, to assess multivariate normality, Mahalanobis distances within the 7 category scores of an observation, for both the consensus and site visit data did not raise any concerns for alarm. Normality assumption was thus assumed satisfied to carry out a one-way

MANOVA.

The analysis in SPSS gave a significant value of Wilks' Lambda; <0.0001 and 0.031 for the site visit and consensus scores respectively. The result of Fisher's post-hoc 'test of between-subjects effects' is shown in Table 3 below.

Table 3. Results of one-way MANOVA for consensus and site visit score data

Dependent Variable	Consensus score data		Site visit score data	
	F	Sig.	F	Sig.
Leadership	10.5	0.002	25.2	<0.0001
Strategy	8.2	0.006	26.1	<0.0001
Customers	5.7	0.021	16.5	<0.0001
MAKM	4.3	0.044	26.7	<0.0001
Workforce	2.5	0.119	15.1	<0.0001
Operations	2.0	0.160	36.3	<0.0001
Results	14	0.001	72.4	<0.0001

The results in table 3 show a significant difference between award winners and non-award winners for all 7 category scores using the site visit data but not different for all with the consensus scores. Specifically, there was no significant difference between award winners and non-award winners for the workforce and operations category scores.

6. Conclusions

This study explored trends and differences in consensus and site visit scores from a States' Performance Excellence Award Program that have been recorded over 11 years period from 2007-2017, with a total of 59 applicants.

The descriptive statistics showed the educational sector scores having the lowest means whilst the healthcare sector always had the highest means for both the consensus and site visit scores. Scoring trends over the years and across sectors increased slightly in 2008 but not too significantly afterward.

We observe examiner consistency in scoring trends by comparing the site visit data against the consensus data. Criteria changes have not been major within the period of study, and this paves the way for data consistency and validity when performing statistical data analysis with this Baldrige data.

There was no significant difference between award winners and non-award winners for the workforce and operations categories with the consensus scores. There were however significant differences between the 2 sets of scores using the site visit data. This result agrees with what was observed in the study to assess scoring differences between award winners and non-award winners for the MBNQA (Tettey, Gholston, & Mesmer, 2018). These are interesting outcomes that seem to suggest that all applicants deserve to be given a site visit tour.

Further work will investigate the adequacy of theoretical linkages underlying the Baldrige Performance Excellence Model.

7. References

ASQ. (2018). Malcolm Baldrige National Quality Award (MBNQA). Retrieved from http://asq.org/learn-about-quality/malcolm-baldrige-award/overview/overview.html

Johnson, R. A., & Wichern, D. W. (2002). Applied multivariate statistical analysis (Vol. 5): Prentice hall Upper Saddle River, NJ.

Lee, S. M., Zuckweiler, K. M., & Trimi, S. (2006). Modernization of the Malcolm Baldrige National Quality Award. International Journal of Production Research, 44(23), 5089-5106. doi:10.1080/00207540500161043

Link, A. N., & Scott, J. T. (2012). On the social value of quality: An economic evaluation of the Baldrige Performance Excellence Program. Science and Public Policy, 39(5), 680-689. doi:10.1093/scipol/scs052

NIST. (2016a). 2017-2018 Baldrige Excellence Framework and Criteria. Retrieved from https://www.nist.gov/news-events/news/2016/12/2017-2018-baldrige-excellence-framework

NIST. (2016b). Four U.S. organizations receive nations highest honor for Performance Excellence. Retrieved from https://www.nist.gov/news-events/news/2016/11/four-us-organizations-receive-nations-highest-honor-performance-excellence

NIST/BPEP. (2018). How Baldrige Works Retrieved from https://www.nist.gov/baldrige/how-baldrige-works

Steeples, M. M. (1994). The Baldrige Award and ISO 9000 in the quality management processes. IEEE Communications Magazine, 32(10), 52-56. doi:10.1109/35.329025

Tettey, A., Gholston, S., & Mesmer, B. (2018). ASSESSING SCORING DIFFERENCES BETWEEN AWARD WINNERS AND NON-AWARD WINNERS FOR THE MALCOLM BALRIGE NATIONAL QUALITY AWARD.

Townsend, P., & Gebhardt, J. (1996). The importance of the Baldrige to US economy. The Journal for Quality and Participation, 19(4), 6.